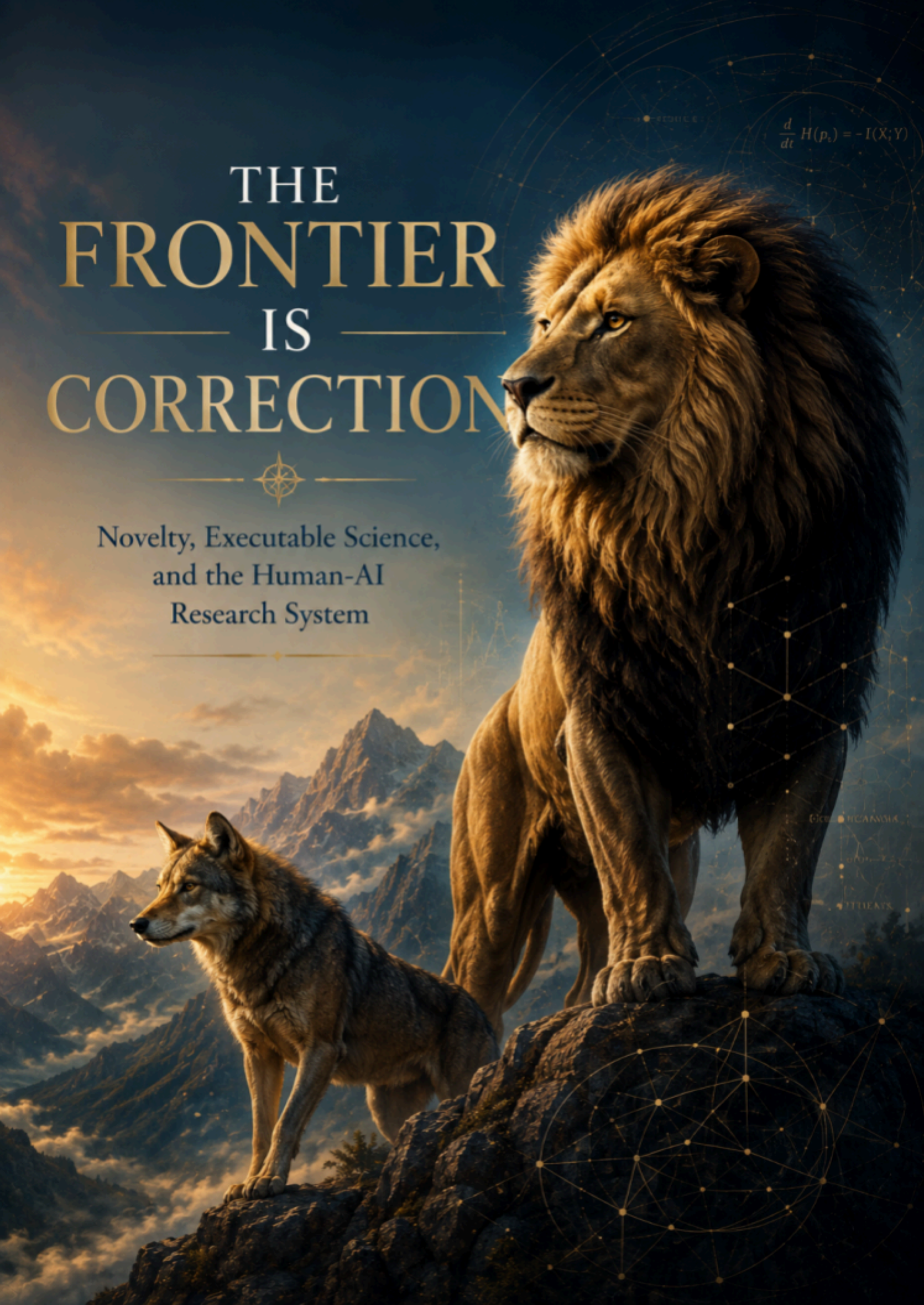
THE FRONTIER IS CORRECTION

Novelty, Executable Science, and the Human-AI Research System



THE FRONTIER IS CORRECTION

Novelty, Executable Science, and the Human-AI Research System

Sînică Alboaie

Axiologic Research SRL · ACHILLES research programme

Science should not become a factory for plausible papers. It should become an infrastructure in which claims can be inspected, challenged, revised, and responsibly released.

Copyright © [2026] Axiologic Research
All rights reserved.

KDP publishing rights held by Outfinity SRL (Iași, Romania).
Available for free on Axiologic website (www.axiologic.net).

Author's Note

This work was created under the umbrella of Axiologic Research as part of two interconnected research programs, one dedicated to Artificial Intelligence and the other to Outfinitism, both led by Sînică Alboaie. To explore the details of how these two programs intersect and build upon each other, we invite readers to visit the Outfinity Venture Studio page at www.outfinity.ch.

While Artificial Intelligence language models were used to assist with text generation, the foundational philosophy, thematic structure, and analytical approach derive exclusively from the author's decades of dedicated study of social phenomena, social technologies, and genuine hard technologies resulting from various European and self funded research projects. Recently, within the Achilles research project, we have been deeply studying AI generated literature and its automated verification using neuro symbolic approaches; all the free books made available to the public are part of the suite of works we use to test our latest technologies.

Abstract

Artificial intelligence is not merely accelerating science. It is changing the scarcity structure on which modern scientific institutions were built. Literature syntheses, hypotheses, code, equations, experiment plans, figures, reviews, and manuscripts can increasingly be generated faster than researchers and institutions can determine what is genuinely new, what caused an apparent result, which criticism is independent, and who is prepared to take responsibility for release. Candidate science is becoming abundant before justification has become equally scalable.

This book argues that the decisive frontier is therefore correction. The relevant measure of progress is not the volume of generated research but the quality-adjusted rate at which claims acquire independent support, are narrowed by better explanations, or are rejected before they consume scarce attention and physical resources. This shift creates a validation economy in which decisive criticism, causal attribution, independent reproduction, specialist judgment, and accountable authority become more valuable than fluent production.

The argument requires a different theory of novelty. Newness is not an intrinsic property of a sentence, model, molecule, or method. It is a relation to a dated and bounded state of knowledge, expressed through a representation and an equivalence rule. Two contributions can be linguistically different but mechanistically equivalent, functionally identical but historically independent, or mostly inherited while removing one assumption that changes what becomes possible. To reason over these distinctions, the book proposes an atlas of operational ideas: an evidence-grounded and revisable layer between documents and scientific judgment.

The same transformation changes the scientific record. A paper remains a signed public argument, but it can no longer serve as the complete memory of an inquiry whose branches, agents, models, instruments, and evaluations change at machine speed. The durable unit becomes an evidence-bearing research object with semantic versioning for specifications, claims, assumptions, evidence, negative results, criticism, and decisions. A paper is then a view over a larger object rather than the object itself.

The architectural proposal is specification-driven inquiry. A living research specification records purpose, admissible claims, evidence thresholds, budgets, stopping conditions, risks, and decision rights. It expands into a governed portfolio of rival hypotheses, controls, replications, and attempts at refutation. Deterministic pipelines handle stable transformations; bounded agents explore where the next useful action cannot be enumerated; independent critics return typed evidence; and humans or institutions retain authority over consequential promotion and release. The resulting system has a deterministic exterior and an adaptive interior, local autonomy and global evidence.

The future described here is not a sovereign artificial scientist but a human-AI republic of knowledge with divided powers. It could make negative results, replication, calibration, and correction visible contributions, or concentrate scientific sovereignty in the owners of models, corpora, compute, and validation services. The outcome will depend less on model scale than on whether science can preserve reasons, expose dependence, retain disagreement, and reverse error before plausibility hardens into authority.

Preface. The Scientific Machine That Must Learn to Correct Itself

The conventional promise of artificial intelligence for science is acceleration. Models will read more papers, generate more hypotheses, write more code, run more simulations, and help laboratories work continuously. All of this matters, but acceleration is the least distinctive and potentially the most misleading description of the transformation. Once the cost of producing plausible research falls sharply, speed ceases to be the central question. The central question becomes whether the scientific system can still distinguish a candidate from a contribution, a correlation from a mechanism, a repeated judgment from an independent test, and technical capability from legitimate authority.

This edition was reorganized around a demanding editorial test. A thesis was moved to the foreground only if it did more than restate that AI will help researchers. It had to explain a present distortion, alter the unit through which science should be understood, imply a concrete architectural or institutional consequence, remain meaningful across several scientific domains, and expose conditions under which it could fail. The ideas that survived this test form one connected argument rather than a catalogue of technologies.

The argument begins with a reversal of scarcity. Scientific institutions developed in a world where candidate work was expensive. A new hypothesis, implementation, experiment, or manuscript consumed enough labour that production itself limited the number of branches a team could pursue. That filter was conservative and frequently unjust, but it constrained volume. Generative and agentic systems weaken it. They can create more plausible branches than a group can understand, reproduce, or seriously criticize. The scarce achievement is no longer the production of a candidate. It is the construction of the independent resistance through which a candidate earns a more authoritative status.

This is the validation economy. It changes how research productivity should be measured. A system that eliminates a seductive explanation before a costly experiment may contribute more than one that produces another positive result. A critic that identifies a common-mode failure can be more valuable than five agents that agree. A maintained benchmark, a repaired dataset, a negative control, or a failed replication can carry greater epistemic value than the prose that eventually receives authorship credit. Scientific automation should be judged by the rate at which weak claims are rejected, important alternatives are preserved, and scarce human and physical resources are allocated to questions that survive competent criticism.

The reversal of scarcity immediately exposes a second problem: science lacks an adequate operational theory of novelty. A model can determine that a sequence is absent from its local memory or that a proposal is distant in an embedding space. Neither fact establishes historical or scientific novelty. A new term can hide an old mechanism. An old principle can become transformative when an assumption is removed, an implementation becomes possible, or the first decisive evidence arrives. A recent paper can resemble an earlier one without having inherited anything from it. Novelty therefore depends on identity, and scientific identity is not one universal relation.

The book's answer is to treat identity as typed and query-relative. Two contributions may align in purpose while diverging in mechanism. They may share a mechanism but differ in assumptions, evidence, or scope. They may be structurally analogous across domains while remaining historically independent. They may be propositionally equivalent and institutionally consequential in different ways. The correct question is not whether two vectors are close. It is which invariants are preserved, which transformations are admissible, which sources support the alignment, and what inquiry the comparison is intended to serve.

This leads to one of the book's central proposals: an atlas of operational ideas. Scientific infrastructure is organized around documents, but scientific reasoning operates over claims, mechanisms, transformations, methods, analogies, constraints, and failures that cut across documents. An operational idea is not a timeless atom of thought. It is a task-relative, evidence-grounded structure whose boundaries and interpretation remain open to challenge. The atlas records several possible decompositions, typed identity relations, historical uncertainty, and the provenance of every abstraction. Its purpose is not to become an oracle of originality. It is to make judgments of precedent, inheritance, analogy, and novelty inspectable enough to be criticized.

A related experimental hypothesis asks whether part of novelty can be represented as structured resistance to reconstruction. A candidate idea may be compared with earlier components under an explicit grammar of transformations. The unexplained residual can indicate a new mechanism, a missing source, a defective representation, the wrong granularity, or simple incoherence. Vector Symbolic Architectures and Hyperdimensional Computing are examined as one possible compositional memory for this task. They are not treated as a novelty detector by fiat. Their role is deliberately subordinate and falsifiable: if residuals do not remain stable under corpus expansion, alternative codebooks, strong baselines, and expert audit, the hypothesis fails while the larger conceptual programme remains intact.

The document-centred organization of science creates another mismatch. A paper is one of civilization's most successful social technologies, but it is a lossy view of an inquiry. It selects a clean line through failed branches, changed evaluators, revised assumptions, calibration events, disputed interpretations, and negative evidence. This compression was already problematic. It becomes structurally inadequate when machines can generate and modify more research states than any participant can remember.

The book therefore proposes an evidence-bearing research object as the durable unit of scientific memory. The object has a stable identity and a versioned history. It links specifications, claims, assumptions, data, code, protocols, models, environments, experiments, reviews, negative results, uncertainty, and decisions. Version control must operate semantically. A scientific “source control commit” should be able to say that a claim was narrowed because a control succeeded, that an evaluator was replaced because it rewarded the wrong property, or that a result ceased to transport to a new population. An epistemic diff shows changes in meaning, not only changes in files. A branch can preserve a rival mechanism. A conflict can remain unresolved when the disagreement is itself informative.

The paper does not disappear in this architecture. It becomes a signed public view over the research object. Its authors accept responsibility for the argument, the declared scope, and the release state. A reviewer can inspect the evidence slice that bears on a disputed claim. A replicator can receive an executable package and a divergence report. A regulator can inspect risk, provenance, and authority. A future research system can learn not only from successes but from what was attempted, why it failed, and which uncertainty remains open.

An executable research object needs an operating model. The proposal developed here is specification-driven inquiry. A prompt is an interaction; a research specification is a persistent commitment. It records the purpose of the inquiry, admissible claim types, assumptions, methods, evidence thresholds, budgets, stop rules, risks, and decision rights. Because research does not begin with a complete specification, the object remains living and versioned. The important distinction is between what must remain invariant and what may be explored.

A scientific specification should not compile into one plan. It should undergo specification explosion: a controlled expansion into rival hypotheses, alternative representations, baselines, negative controls, scaling tests, replications, and attempts at refutation. This is portfolio science. It makes premature convergence harder and preserves the rejected branches that later become scientific memory. The goal is not to maximize branching. Each branch receives a budget, a

stopping condition, and an expected evidential role. Generation can be broad; promotion must become progressively narrower and more expensive.

Execution follows a complementary principle: deterministic exterior, adaptive interior. Stable operations should not be repeatedly reinterpreted by a language model. Schema checks, data transformations, environment construction, proof checking, policy enforcement, and provenance capture should be deterministic where possible. Bounded agents are useful where interpretation and search remain open: literature exploration, hypothesis formation, representation choice, anomaly diagnosis, and counterexample design. Agents return named artefacts and typed evidence into an explicit outer graph. They do not become the implicit constitution of the project through a long conversation.

This arrangement produces local autonomy with global evidence. A coding agent may change a sandboxed repository. A laboratory controller may optimize inside a safety envelope. A theorem prover may certify a formal statement. A literature agent may search across vocabularies and fields. None owns the truth state of the inquiry. Outputs enter a common claim-evidence structure in which dependencies, scope, uncertainty, criticism, and decision rights remain visible. When an upstream artefact changes, the system selectively recomputes the affected subgraph rather than regenerating the project and introducing unrelated variation.

The next problem is criticism. Telling the same model to play several personas does not create scientific independence. Agents may share training data, retrieval sources, code, evaluators, prompts, and institutional incentives. Apparent consensus can therefore be a dependency pattern rather than evidence. The book proposes dependency fingerprints for critics and evidence: records of the model families, corpora, implementations, tools, laboratories, and human operators that contributed to a judgment. Independence becomes an evidential property to be demonstrated, not a theatrical property inferred from the number of voices in a transcript.

Claims should advance through evidential escalation. A candidate can pass coherence and syntax checks, then execution, adversarial challenge, causal ablation, independent reproduction, specialist adjudication, and release under a policy proportionate to consequence. Promotion requires new kinds of support, not repeated confidence from the same generator. Formal claims encounter proof kernels. Computational claims encounter execution and independent implementation. Empirical claims encounter instruments and other laboratories. Causal claims encounter interventions or observations that separate rival mechanisms. Normative and political claims encounter legitimate human deliberation.

This is why the future of scientific AI is a constitutional problem. Capability and authority are different. A system may be able to propose a clinical protocol without possessing the right to expose patients. It may predict a social outcome without deciding which outcome a society ought to optimize. It may discover an efficient process without authorizing the environmental or distributive costs of deployment. The human role is not adequately described by the phrase human in the loop. People require defined decisions, adequate evidence, genuine power to stop or revise, and responsibility proportionate to consequence.

The likely future is not a single artificial scientist but a republic of heterogeneous scientific powers. Models propose and translate. Deterministic systems enforce contracts. Instruments resist. Critics challenge bounded properties. Archives preserve ancestry and failure. Specialists adjudicate questions for which they are competent. Institutions authorize release and remain open to appeal. Such a republic can support local machine autonomy without granting one component control over generation, validation, memory, and social consequence.

This future also contains a less discussed epistemic possibility. Machines may discover operational structures that predict and support intervention before humans can compress them into familiar concepts. The correct response is neither automatic rejection nor submission to opacity. A machine-native concept should acquire a semantic passport: the observations that identify it, the operations that define it, the interventions it predicts, the domains in which it

remains stable, the alternative models that agree, and the consequences of using it. Scientific understanding can then become staged rather than falsely complete.

The political alternative is equally important. The same architecture can become a republic of knowledge or an empire of epistemic infrastructure. Open standards, public compute, shared laboratories, portable research objects, and plural evidence systems can widen scientific participation. Closed corpora, proprietary models, vendor-dependent laboratories, and opaque credibility scores can make independent challenge impossible. Scientific sovereignty will depend on who controls not only generation but the memory, ranking, and validation layers through which claims become visible and authoritative.

This book will not merely survey agents, laboratories, novelty detection, or peer review. It connects them into a theory of what scientific systems must become once candidate research is abundant. Chapter 1 gives the theory in full through one machine-generated claim; the later parts examine its epistemology, architecture, domain limits, institutional history, and constitutional consequences.

The argument exceeds any current implementation but does not claim universal automation. It distinguishes demonstrated mechanisms, architectural inference, research hypotheses, and scenarios. Production is beginning to outrun comprehension; scientifically valuable systems must make errors traceable, independently testable, and reversible.

Editorial and Evidence Note

This edition is a self-contained research and editorial manuscript developed from four working manuscripts: Executable Science; The Atlas of Operational Ideas; The Frontier of Novelty; and When Science Builds a Partner. The integration preserves their strongest terminology and arguments while removing duplicated examples, repeated definitions, and competing narrative orders.

The book distinguishes four kinds of statement. Some passages describe demonstrated systems and published results. Some draw architectural inferences from those results. Some present proposals associated with the ACHILLES research direction, including specification-driven inquiry, operational-idea representations, SOP Lang, and the emerging MRP-VM concept. Others are explicitly speculative scenarios about future scientific institutions. The language of the text is intended to keep these categories separate.

The evidential basis remains the literature and source analysis contained in the four manuscripts. Peer-reviewed primary research receives the greatest weight for empirical claims; reviews are used to organize fields; preprints and system reports are treated as provisional when they describe frontier work; and future scenarios are never presented as observations. No claim of exhaustive coverage is made, particularly in fast-moving research areas.

The term science is used broadly but not indiscriminately. Formal proof, computational experimentation, laboratory research, observational inference, clinical investigation, social inquiry, and interpretive scholarship do not share one universal oracle. Executable science therefore means proportional operational explicitness: enough structure to make transformations, evidence obligations, and authority inspectable without pretending that every scientific judgment can be reduced to an algorithm.

Prologue. Two Rooms, One Question

On a cold night in a European town near the end of the Middle Ages, a man sets a crucible over glowing coals and watches a metal change color. His room is narrow and smoky. The jars have no stable labels. Recipes have been copied by hand, altered by memory, and protected by symbols whose ambiguity is partly philosophical and partly commercial. The operator does not have a laboratory in our sense of the word. The precision of his instruments depends less on universal units than on the trained eye and hand. He raises the heat, waits, compares, fails, and tries again. In that room, explanation, craft, spiritual discipline, patronage, and hope are not yet separate domains. Yet something recognizable as research is happening. A procedure can succeed even when the theory offered for it is wrong.

Several centuries later, in a room with almost no people, a robotic arm transfers samples between a synthesis station and a characterization instrument. An artificial-intelligence system has searched the literature, proposed the next experiment, generated analysis code, estimated uncertainty, and selected a region of the design space that seems worth exploring. Sensors stream data into an experiment registry. A second model looks for anomalies. A third checks whether the protocol has drifted from its approved version. The laboratory can work through the night. It does not become tired, forget to save a file, or lose patience after the hundredth failed run. It can also pursue a poorly chosen metric with flawless persistence. It can turn calibration drift into a discovery and produce, at the same speed, both knowledge and the appearance of knowledge.

Between these two rooms lies the history of science. It is not a straight road from superstition to truth. It is the history of societies learning, slowly and conflictually, to convert curiosity into a public practice, measurement into a shared convention, doubt into an institution, and a result into a claim for which somebody must answer. Philosophy asked what it means to know. Alchemy joined the cosmology workshop. Printing made texts comparable at scale. Learned societies created rituals of testimony. Instruments extended the senses; statistics disciplined the human appetite for patterns; universities, journals, national laboratories, grants, ethics committees, and peer review gave inquiry to a social body. Every gain in rigor, however, created a new surface on which rigor could be imitated.

Artificial intelligence enters this history as more than one additional instrument. It can occupy several roles at once: librarian, programmer, statistician, critic, experimental designer, laboratory operator, and manuscript writer. Systems such as OpenScholar and PaperQA2 search large scientific corpora and return evidence-linked syntheses. The AI co-scientist uses a multi-agent process to generate, debate, rank, and evolve hypotheses. ERA combines language models with tree search to write and improve empirical scientific software when a task can be scored. The AI Scientist and Agent Laboratory connect idea generation, coding, experimentation, and paper production. Coscientist, ChemCrow, A-Lab, ChemOS, and related self-driving laboratories connect digital reasoning to physical instruments. Benchmarks such as ScienceAgentBench, MLE-bench, PaperBench, MLR-Bench, AutoResearchBench, and ResearchClawBench ask harder questions: can an agent write valid scientific code, find the right literature, replicate a paper, or rediscover a result under conditions designed to expose fabricated success? Their answer, so far, is not a simple yes. Performance is improving quickly, but robust autonomy remains much weaker than fluent demonstrations suggest (Chen et al., 2024; Chan et al., 2024; Starace et al., 2025; Xiong et al., 2026).

The question of this book is therefore broader than whether AI will replace scientists. The more consequential question is what science will become when searching, comparing, experimenting, and arguing are distributed across people and artificial systems. What made a scientific result legitimate in different historical periods? What counts as validation when no individual can mentally follow every transformation between raw data and conclusion? Can

there be a science that is predictively and interventionally reliable but whose concepts are not intelligible to humans? What happens to peer review when the number of manuscripts, hypotheses, simulations, and experiments grows by orders of magnitude? Who will be paid and honored for the slow work of checking? How will society distinguish a genuine researcher from a charlatan when both can generate polished graphs, equations, references, and a voice of serene certainty?

The argument developed here is that the future of science depends less on the arrival of an artificial genius than on our ability to rebuild the institutions of knowledge around verifiability. Science did not become credible because scientists stopped making mistakes. It became credible because, at its best, it made error detectable, disagreement permissible, and correction possible. Artificial partners will amplify research only when important claims arrive with their paths attached: sources, data, code, versions, instruments, uncertainties, rejected alternatives, negative results, and independent attempts at refutation. Without that path, AI will not create an age of knowledge. It will create an age of industrialized plausibility.

The two rooms share another feature. In each, outsiders must decide whether to trust the operator. In the first room, the operator is a person who invokes tradition, experience, and perhaps the name of a patron. In the second, the operator is a network of models, services, databases, instruments, and organizations, each carrying a different kind of technical authority. In both cases, the local marvel may be genuine, the mistake may be sincere, and fraud may borrow the vocabulary of competence. The fundamental question is not what impresses us. It is what can be traced, repeated, and contested.

This is not a story of an inevitable future. The same capabilities can strengthen open science or deepen concentration. They can relieve researchers of repetitive labor or turn them into supervisors of systems they no longer understand. They can make negative results visible or produce an ocean of synthetic papers that hides them. They can democratize access to expertise or place scientific sovereignty in the hands of a few firms and states. Technology opens a range of futures; institutions select among them.

The two rooms are not the final contrast. The book's real subject is a third room distributed across repositories, models, simulators, laboratories, reviewers, regulators, and publics. There a claim carries its ancestry, evidence, permissions, critics, and unresolved alternatives. A machine may propose it, but cannot silently define novelty, select its own judge, erase failed branches, or grant itself social authority.

The book therefore begins with the future system. Part I gives the complete argument in compressed form; later chapters justify it through epistemology, architecture, domain differences, and institutional history. The promise comes first, the detail after.

PART I: THE WHOLE ARGUMENT IN COMPACT FORM

The opening part gives the complete thesis before any technical survey or historical detour. It follows one machine-generated claim from apparent breakthrough to scientific legitimacy and exposes the infrastructure required at every transition.

Chapter 1. The Claim That Arrived Before Its Evidence

A discovery can now be generated, executed, narrated, and reviewed before science has established what is new, what caused the result, who independently checked it, or who has authority to act.

At 08:10, a research system identifies what appears to be an unexplored relation in a scientific corpus. By 09:40 it has translated the relation into a mechanism. Before noon it has written code, selected data, and produced a promising result. In a digitally native field, or in a tightly automated laboratory, it may also execute a sequence of experiments. By evening it has generated figures, a manuscript, a novelty statement, and several reviews produced by other agents. The next morning the result is described as a discovery.

Nothing in this thought experiment requires a mythical general intelligence. Every operation already exists in partial form: scientific retrieval, hypothesis generation, code execution, theorem search, active experimental design, robotic control, manuscript production, and automated criticism. The unusual feature is their speed and composition. The result can acquire the complete surface of science before any person or institution has reconstructed the path beneath it.

What, exactly, arrived at 08:10? Was the relation new to the model, new to the indexed corpus, new to one discipline, or new to science? Did the code test the intended mechanism or a convenient proxy? Was the experiment decisive, or did the system choose a region in which its evaluator was easiest to satisfy? Were the reviews independent, or did every agent inherit the same corpus, model family, and blind spot? Which failed branches disappeared from the manuscript? Who is prepared to sign the claim, authorize its use, and accept the consequences if it is wrong?

The system has not necessarily produced a discovery. It has produced a candidate scientific package whose appearance is more complete than its epistemic status. This mismatch is the central condition of the coming era. Artificial intelligence lowers the cost of candidate science more quickly than it lowers the cost of deciding which candidates deserve belief.

Modern scientific institutions were built under a different scarcity. A hypothesis, implementation, experiment, and paper consumed enough human labour that production itself constrained volume. This filter excluded unconventional ideas and preserved many inequalities, but it also limited the number of branches that demanded attention. A generative system can remove the cost of proposing a branch without removing the cost of understanding it, designing a decisive test, reproducing it independently, or taking responsibility for release.

Science is therefore entering a validation economy. The scarce resources are no longer only ideas, code, or prose. They are criticism that can change the fate of a claim, experiments that separate live explanations, specialists who can judge a narrow unresolved question, laboratories whose errors are materially independent, and institutions willing to own a consequential decision. The scientific value of automation should be measured by the quality-adjusted rate at which claims acquire independent support or are efficiently rejected, not by the number of candidate papers a system can produce.

This reverses the usual image of progress. The most valuable system may be the one that generates fewer final claims because it discovers earlier that an effect comes from leakage, calibration drift, extra compute, a weak baseline, or an invalid operational definition. A negative control that closes a branch can be more valuable than an additional positive experiment. A critic that exposes a shared dependency can contribute more than a chorus of agreeing agents. The rate of correction becomes more informative than the rate of narration.

The danger is not only hallucination. It is the mechanical research loop: a process that remains locally competent while optimizing a proxy that has drifted from the purpose of inquiry. The literature agent finds plausible sources. The planner proposes a sensible experiment. The

coding agent produces a working repository. The evaluator reports improvement. The reviewer praises clarity. Every local step can appear successful while the system has silently changed the question or selected an evaluator that cannot distinguish the proposed mechanism from a simpler explanation.

The first requirement of future science is therefore not greater confidence but organized resistance. A scientific system must be able to state what would change its conclusion, expose where an evaluator is weak, preserve rival explanations, and stop when continued optimization no longer serves the original purpose. This is the meaning of the title: the frontier is correction because reliable intelligence is not the absence of error. It is the architecture through which error becomes visible, consequential, and reversible.

Return to the morning claim. Its first unresolved property is novelty. A language model can report that it did not retrieve an equivalent phrase. An embedding system can report that the proposal lies far from familiar documents. Neither result establishes that the underlying idea lacks a precedent. The same mechanism can migrate across fields under different terminology. A recent method can repeat an old principle while adding the first effective implementation. A contribution can be mostly inherited and still become important by removing one restrictive assumption.

Newness is not a substance attached to the candidate. It is a relation among the candidate, a memory, a representation, an equivalence rule, a time, and a community. Change the corpus and the priority claim can change. Change the level of abstraction and a method can become old in mechanism but new in evidence. Change the comparison question and two contributions can become the same for one purpose and different for another.

Identity must therefore be typed. Two works may be paraphrases without sharing historical descent. They may be functionally equivalent while using different mechanisms. They may implement the same mechanism under incompatible assumptions. They may be structurally analogous across domains but historically independent. Similarity, lineage, influence, and rediscovery are not interchangeable labels on one edge. They authorize different inferences and require different evidence.

This is why document retrieval, however powerful, is insufficient. Scientific infrastructure stores papers, authors, citations, and passages. Scientific reasoning operates over problem formulations, mechanisms, transformations, methods, constraints, evidence, failures, and analogies that cross those boundaries. The morning claim needs an intermediate epistemic layer capable of saying which components were inherited, which were transformed, which sources support the comparison, and where uncertainty remains.

The proposed layer is an atlas of operational ideas. Its entries are not timeless atoms of thought. They are task-relative structures connecting goals, mechanisms, assumptions, contexts, consequences, evidence, and provenance. More than one decomposition can coexist. A historian, an engineer, and a reviewer may preserve different invariants without one representation becoming the universal ontology. The atlas becomes useful because its abstractions remain linked to sources and because identity relations can be challenged, split, merged, or superseded.

The atlas changes the novelty claim from a theatrical adjective into an argument. Instead of saying that a proposal is new, the system shows the best available reconstruction from prior components, the transformations required, and the role-specific residuals that remain. A residual may indicate substantive novelty. It may also indicate missing literature, the wrong granularity, a defective representation, extraction error, or incoherence. The value of the analysis lies in making these alternatives explicit.

The second unresolved property of the morning claim is validity. A positive result is not yet an explanation. An improved benchmark may come from a different preprocessing path, a larger search budget, a leak, or an evaluator that rewards the wrong behaviour. A laboratory signal may come from contamination or instrument drift. A formal proof may certify a statement

that fails to capture the intended informal question. Every evaluator provides resistance only for the property it actually checks.

The generator must therefore not be its own judge. Asking one model to generate, implement, interpret, and review a claim can improve consistency while reducing epistemic independence. Several personas do not create several evidence bases. They may share training data, retrieval indexes, tool implementations, benchmark assumptions, and institutional incentives. Apparent consensus can be the visible form of one common-mode error.

A correction-centred system records dependency fingerprints. For each important judgment it identifies the model family, corpus, prompt regime, implementation, evaluator, laboratory, instrument, and human authority on which the judgment depends. Independence becomes a property to be demonstrated. A proof kernel, an independently written implementation, a new dataset, a different physical measurement, or a laboratory outside the originating organization changes the evidential status because it introduces a new failure structure, not merely another vote.

Evidence should escalate. The morning candidate may first pass syntax, type, and coherence checks. It may then be executed. It may be challenged by negative controls, matched baselines, ablations, counterexample search, and alternative causal models. It may be reproduced under materially independent conditions. Specialists may adjudicate the remaining questions. Only then may an institution release the result for a declared audience and use. Promotion requires new kinds of support, not stronger self-confidence.

This asymmetry is deliberate. Generation can be broad, inexpensive, and speculative. Promotion should be narrow and increasingly costly as the claim approaches public authority or physical consequence. A system should make it easy to abandon a weak branch and difficult to erase the fact that a serious objection remains unresolved.

The third unresolved property is memory. The manuscript created by the system is likely to present the surviving line as a coherent story. It may be accurate as an argument and still conceal the structure that determines what the result means: how many mechanisms were tried, which evaluator changed, which branch failed only after a scaling test, which control narrowed the claim, and whether the same model family generated and judged the evidence.

The scientific paper remains indispensable because science needs intelligible public arguments and named responsibility. But the paper can no longer carry the entire memory of a machine-speed inquiry. It is a lossy projection. When agents can generate more transformations than any participant can remember, the durable system of record must be larger than the manuscript.

The proposed replacement is not the abolition of the paper but an evidence-bearing research object. The object has a stable identity and versioned history. It contains the living specification, claims, assumptions, data, code, environments, protocols, instrument records, experiments, negative results, reviews, uncertainty, and decisions. The paper becomes a signed view over this object. A reviewer inspects the evidence slice relevant to a claim. A replicator receives the executable package and a record of known divergence. A regulator inspects permissions and authority. A future agent can see not only what succeeded but what was attempted and why it failed.

Versioning must be semantic. A scientific committee should record that a mechanism claim was narrowed because a control succeeded, that an evaluator was replaced because it measured a proxy, or that a result no longer transported to a new population. An epistemic diff shows changes in assumptions, scope, evidence, and authority. A branch can preserve a rival causal model. A merge should occur only when the assumptions and evidence are compatible. A conflict should remain visible when the disagreement is the most informative state of the inquiry.

This history must not become total surveillance. Research requires a protected interval in which people and systems can be wrong without every provisional thought becoming a

permanent employment or funding record. The workbench should preserve committed reasons that changed the public state of the inquiry, not every token of private exploration. Layered transparency is therefore part of scientific design: private experimentation, collaborative state, institutional audit, and public release have different evidential and privacy obligations.

The fourth unresolved property is execution. A long conversation with a universal agent is not an adequate scientific architecture. It mixes transient reasoning, project memory, policy, and authority in one context. It encourages silent semantic drift and makes selective recomputation difficult. When an upstream assumption changes, the system may regenerate everything and introduce unrelated variation rather than identify the affected dependency cone.

A research programme needs a living specification. This specification records the objective, admissible claim types, assumptions, invariants, methods, evidence thresholds, resource budgets, stopping conditions, risks, and decision rights. It is not complete at the beginning because research often changes the vocabulary of the question. Its function is to distinguish what must remain stable from what may be explored and to make consequential changes explicit.

The specification should not compile into one linear plan. It should undergo specification explosion: controlled expansion into rival hypotheses, representations, baselines, negative controls, scaling tests, replications, and attempts at refutation. The morning claim should not inherit the entire programme merely because it was the first fluent completion. Research becomes a governed portfolio whose branches have budgets, evidential roles, and stopping conditions.

The execution architecture has a deterministic exterior and an adaptive interior. Stable transformations should be implemented as stable procedures: schema validation, data conversion, environment construction, formal checking, statistical diagnostics, provenance capture, access control, and policy enforcement. Bounded agents should explore where the next useful action cannot be enumerated: terminology, hypotheses, representations, anomaly diagnosis, and counterexample design. The agent returns named artefacts and evidence into an explicit outer graph. Its conversation is not the project memory.

This produces local autonomy with global evidence. A literature agent may search broadly. A coding agent may modify a sandboxed repository. A theorem prover may certify a formal claim. A robotic controller may optimize inside an approved safety envelope. None can promote its own output by narrative force. The shared workbench retains the claim-evidence graph, the relevant dependencies, and the authority required for every transition.

When an input changes, the workbench performs selective recomputation. Deterministic nodes rerun automatically. Agentic nodes receive the smallest relevant frame rather than the entire project transcript. Human review is requested where meaning, risk, or authority changes. This is how scientific automation becomes maintainable rather than merely fast.

The fifth unresolved property is authority. The morning system may be capable of proposing a clinical intervention, operating an instrument, or releasing a public claim. Capability does not grant permission. Scientific claims enter social worlds in which errors affect patients, environments, infrastructure, public budgets, and the distribution of opportunity. The authority to act must remain attributable and contestable.

A serious research object records who defined the question, who approved access to data, who accepted the operationalization, who reviewed the evidence, who authorized release, and under which policy version. Authority does not guarantee truth; institutions can be wrong or captured. Its function is to prevent consequence from becoming ownerless. A system must also preserve the path by which a claim can be challenged, corrected, superseded, or withdrawn.

The human role is therefore not a ceremonial click at the end of an automated pipeline. Human judgment becomes most important where the evaluator is weak, the representation is contested, or the consequence is high. A statistician may own the identification assumptions. A domain scientist may judge whether an assay measures the intended phenomenon. A patient or

community representative may expose a harm absent from the objective. A safety authority may stop a physically possible experiment. Each judgment should be scoped, informed, and connected to the claim it can legitimately change.

The system that emerges from these requirements is not one artificial scientist. It is a scientific constitution with divided powers. Sources and instruments constrain interpretation. Operational ideas organize comparison without pretending to be the terrain itself. Research objects preserve versions and reasons. Generators explore. Deterministic systems enforce contracts. Critics challenge bounded properties. Specialists adjudicate. Institutions authorize consequence. Publics retain standing to contest categories and uses.

This constitution also prepares science for machine-native concepts. A system may discover a representation that predicts and supports intervention while resisting human compression. Such a concept should not be accepted because the model is powerful or rejected because it is unfamiliar. It should acquire a semantic passport: the measurements that identify it, the operations that define it, the predictions it supports, the perturbations it survives, the alternative representations that agree, and the known boundary of use. Opacity raises the burden of triangulation; it does not erase the possibility of knowledge.

The architecture can now be summarized in six interacting layers. No layer is sufficient by itself, and each corrects a characteristic failure of the others.

Table 1. Six layers required before generated research becomes a corrigible scientific system

Layer	Function in the proposed scientific infrastructure
Documentary layer	Preserves publications, passages, data, code, protocols, instrument records, and other source artefacts without confusing them with later interpretation.
Operational-idea layer	Represents goals, mechanisms, assumptions, identity profiles, transformations, lineages, novelty, and disagreement at task-appropriate granularity.
Research-object layer	Versions specifications, claims, dependencies, evidence, negative results, reasons, and public release states.
Discovery layer	Coordinates generation, search, experiments, simulations, tools, archives, and bounded agentic work.
Adjudication layer	Organizes critics, evaluator independence, evidence escalation, reproduction, causal attribution, and decisions about claim status.
Institutional layer	Governs access, authority, safety, privacy, credit, contestability, correction, and responsibility for consequence.

The documentary layer prevents interpretation from floating free of sources. The operational-idea layer prevents documents and similarity scores from silently defining identity and novelty. The research-object layer prevents polished releases from erasing branches, reasons, and negative evidence. The discovery layer supplies generation, search, experiment, and archive. The adjudication layer introduces independent resistance and controls promotion. The institutional layer ensures that technical capability does not become unaccountable authority.

The six layers explain why larger models alone cannot solve the scientific problem. Better generation cannot repair an evaluator that rewards the wrong property. Provenance cannot make a biased measurement true. Formal verification cannot establish that a formalization captures the intended phenomenon. Human oversight cannot work when a person receives only a polished conclusion and lacks time, authority, or a precise decision. Open access to a paper cannot compensate for dependence on an inaccessible model or laboratory.

They also explain why this is not merely a proposal for a distant future. Present science already contains mechanical loops. Publication metrics reward visible volume. Benchmarks can become ends rather than instruments. Negative results disappear. Reviewers lack time to rerun code or reconstruct search paths. AI does not create these weaknesses. It makes them scalable, faster, and harder to notice because the resulting prose and artefacts can be unusually coherent.

The same technical capacity can produce two institutional futures. In one, open standards, public compute, shared laboratories, portable research objects, and plural critics create a wider republic of knowledge. In the other, a small number of organizations control the models, corpora, search rankings, laboratory interfaces, and validation services through which scientific claims become visible. The second future can be empirically productive and politically fragile. A civilization that cannot independently challenge its scientific infrastructure has outsourced part of its sovereignty.

The chapters that follow unpack this compressed argument. The next part explains the validation economy, the relational nature of novelty, the missing epistemic layer between documents and ideas, typed identity, and the transition from papers to versioned research objects. Later parts examine the structural limits of learning, the architecture of generators and evaluators, the different resistance supplied by proofs and laboratories, the history through which doubt became public, and the institutions required for a human-AI republic of knowledge.

The book's claim is demanding but concrete. Science can become more executable without becoming intellectually mechanical. Artificial systems can explore more broadly while evidence and authority become more explicit. Negative results, replication, calibration, criticism, and maintenance can become visible contributions. Ideas can become comparable without being reduced to one metric. Autonomy can expand where evaluators are strong and remain constrained where ambiguity and consequence are high.

The morning claim becomes science only when it survives this architecture. It must be located relative to prior knowledge, represented at the right granularity, tested by evaluators that do not merely echo its generator, preserved inside a versioned object, challenged through independent evidence, and released by an authority that can be held responsible. Until then, it is not worthless. It is a candidate. The future of science depends on remembering the difference.

PART II:
THE NEW SCIENTIFIC CONDITION

The condensed thesis is now unpacked: why correction becomes scarce, why novelty requires typed identity, why documents are the wrong unit of scientific memory, and why the paper becomes a signed view over a versioned research object.

Chapter 2. When Ideas Become Cheap and Doubt Becomes Expensive

Once hypotheses, experiments, and manuscripts can be generated at machine speed, the scarce achievements are decisive rejection, causal attribution, independent reproduction, and accountable release.

2.1 The Cheapening of Candidate Science

Give a research group an agent that can generate a hundred plausible mechanisms overnight and the group has not received a hundred discoveries. It has received a new allocation problem. Which branch deserves compute, an experiment, a specialist's attention, or the moral and institutional burden of public release? The first-order effect of scientific AI is therefore not simply higher productivity. It is the collapse of production cost before the corresponding collapse of validation cost.

This development continues a much older programme. The Logic Theory Machine demonstrated that proof search could be mechanized. DENDRAL showed that domain knowledge could constrain scientific hypothesis formation in chemistry. Robot Scientist systems linked hypothesis generation to laboratory experimentation and falsification in bounded biological domains (Newell and Simon, 1956; Lindsay et al., 1993; King et al., 2004, 2009). Program-search systems have since discovered faster matrix multiplication algorithms and new mathematical constructions, while autonomous chemistry and materials platforms have joined planning, execution, and measurement in closed or semi-closed loops (Fawzi et al., 2022; Romera-Paredes et al., 2024; Boiko et al., 2023; Szymanski et al., 2023). What language models add is a general translation layer. They can move between ordinary language, scientific prose, code, symbolic notation, database queries, and tool calls without requiring a handcrafted interface for every transition.

Recent systems make the changing frontier visible. The AI Scientist integrates ideation, literature search, code modification, experiment execution, analysis, manuscript production, and automated review in a machine-learning setting (Lu et al., 2026). Google’s Co-Scientist uses a multi-agent search-and-evolution process to generate and refine biomedical hypotheses under scientist-specified goals, with selected proposals subjected to laboratory validation (Gottweis et al., 2026). FutureHouse’s Robin combines literature agents and data-analysis agents in an iterative biology workflow that connects hypotheses to experimental results (Ghareeb et al., 2026). These systems differ in architecture and evidential strength, but they demonstrate that substantial portions of the research cycle can be represented as computational operations.

The same pattern appears in formal and computational discovery. FunSearch couples a language model to an evaluator and an evolutionary archive, allowing the model to propose program fragments while an executable function determines whether they are valid and how well they perform (Romera-Paredes et al., 2024). AlphaEvolve generalizes the idea toward algorithm discovery and optimization by combining generative models, program mutation, evaluators, and an evolving pool of candidates (Novikov et al., 2025). AlphaGeometry and related neuro-symbolic theorem-proving systems use language models or learned components to propose constructions while symbolic engines enforce proof obligations (Trinh et al., 2024). Their success is not evidence that free-form language models have acquired infallible mathematical

judgment. It is evidence that generation becomes powerful when embedded in a search process with a strong external evaluator.

Laboratory automation supplies another route. ChemOS, self-driving laboratories, A-Lab, automated optical platforms, and emerging multi-agent chemistry systems demonstrate that planning can be coupled to instruments through machine-readable protocols, active learning, Bayesian optimization, and recovery procedures (Häse et al., 2019; Burger et al., 2020; Szymanski et al., 2023; Panapitiya et al., 2026). Physical automation introduces a kind of resistance that purely textual loops lack. A synthesis fails, a sensor drifts, an instrument saturates, or a material behaves differently from the model. Yet this resistance is useful only when the system records calibration, uncertainty, execution context, and deviations. A robot that repeats an invalid assay at scale has not made the research more rigorous.

Information-centred systems are likely to spread fastest because their inputs and outputs are already digital. Tools such as Elicit, Semantic Scholar, Scite, PaperQA2, and STORM automate parts of search, evidence extraction, citation-aware synthesis, and document construction. Their practical value is considerable, particularly when they reduce the cost of mapping a literature, comparing claims, or identifying missing evidence. But an information workflow is not automatically a scientific workflow. Search quality, corpus access, source reliability, claim granularity, citation fidelity, and the treatment of negative results all determine whether the output supports inquiry or merely produces a fluent review.

The relevant distinction is not between tasks that are intrinsically human and tasks that are intrinsically mechanical. It is between operations that can be delegated under an explicit contract and decisions whose validity or legitimacy depends on context, external resistance, specialist judgment, or consequence. Data conversion can often be deterministic. A literature query can be delegated if scope and source policy are declared. Code can be generated if tests and environment constraints are explicit. A hypothesis can be proposed by a model. Whether a question is worth asking, whether an operationalization is faithful, whether a negative result closes a branch, whether a causal interpretation is warranted, and whether a finding should influence patients or public policy require more than syntactic competence.

As the cost of generation falls, the traditional relationship between proposal and attention reverses. In the old regime, a research group might explore a small number of hypotheses because each implementation and experiment was expensive. The scarcity of production indirectly filtered the space. That filter was crude and often excluded unconventional ideas, but it prevented unlimited branching. In the emerging regime, a system can formulate hundreds of plausible directions before a specialist has read the first ten. The constraint shifts from the supply of ideas to the capacity to discriminate among them.

This shift is already visible in benchmarks. PaperBench decomposes the replication of twenty machine-learning papers into thousands of gradable tasks and shows that even capable agents remain far from reliable end-to-end reproduction (Starace et al., 2025). MLE-bench measures the ability to complete machine-learning engineering tasks under competition-like conditions and similarly reveals a gap between local coding competence and sustained project execution (Chan et al., 2024). DiscoveryWorld and laboratory-oriented benchmarks expose failures in experimental reasoning, tool use, and state tracking (Jansen et al., 2024; Laurent et al., 2024). These

evaluations do not define science, but they demonstrate a recurring pattern: generating a plausible next action is much easier than maintaining a valid chain from objective to result.

The consequence is a new production function for research. More model capability and more test-time compute can increase the volume and sometimes the quality of candidate outputs. The AI Scientist reports improvements with deeper experimental search; Co-Scientist reports improvements as hypotheses undergo extended generation, debate, ranking, and evolution (Lu et al., 2026; Gottweis et al., 2026). Yet scaling search does not itself guarantee that the evaluator measures the right property. If the reward is workshop acceptability, the system may learn the form of a workshop paper. If the reward is benchmark score, it may find a benchmark-specific shortcut. If novelty is estimated through semantic distance, it may produce unfamiliar phrasing rather than a new mechanism. Generative abundance therefore increases the importance of evaluator design rather than reducing it.

A useful way to describe the transition is that science is moving from an economy of scarce conjecture to an economy of scarce validation. The phrase should not be taken literally in every field. Many good hypotheses have always been plentiful, while access to data, instruments, participants, and expert time has long been scarce. What changes is the scale and speed at which plausible research objects can be synthesized. The ability to produce them no longer indicates that the surrounding evidence can support them.

The first design principle of executable science follows. Research automation should be evaluated by the quality-adjusted rate at which claims acquire independent support or are efficiently rejected, not by the rate at which hypotheses, experiments, or papers are generated. This principle changes what should be optimized. A system that produces fewer claims but identifies decisive counterexamples earlier may be more valuable than one that fills a repository with polished results. A workbench that makes a hidden assumption visible may create more scientific value than a model that adds another percentage point to an internal score.

Table 2. The emerging validation economy

Functions becoming abundant	Functions that remain scarce
Literature retrieval, synthesis, and reformulation	Judging which omissions, assumptions, and contradictions are consequential
Generation of hypotheses, mechanisms, code paths, models, and prose	Causal attribution, discriminating experiments, and the design of decisive negative evidence
Execution of simulations and digitally instrumented workflows	Independent reproduction across tools, laboratories, datasets, and theoretical frames
Production of polished figures, reports, and manuscripts	Specialist attention, accountable release, and institutional willingness to own a claim
Repeated self-critique by the same model family	Criticism with genuinely different information, incentives, methods, or authority

2.2 What Remains Scarce

Generative systems make certain kinds of intellectual production cheap, but they do not make every epistemic function cheap. The most important scarcities migrate

toward the work that determines what a result means and whether anyone should rely on it. Three forms of scarcity are especially consequential: decisive criticism, causal attribution, and accountable authority.

Decisive criticism is not generic negativity. It is the design of a test that can discriminate between live alternatives. A weak critic restates limitations or assigns an impressionistic score. A strong critic identifies the observation that would force a claim to be narrowed, the component whose removal would reveal a simpler explanation, the scaling regime in which a local property ceases to matter, or the external dataset on which an apparent effect should survive. Such criticism requires knowledge of the domain and of the specific claim type. A theorem claim needs proof checking and faithful formalization. A mechanism claim needs interventions or discriminating observations. A comparative performance claim needs matched baselines, budgets, and variance analysis. A safety claim needs hazard identification, adversarial testing, and a consequence-sensitive threshold.

Automated critics can help. Models can search for prior work, compare manuscripts with code, generate test cases, execute robustness checks, detect citation mismatches, or propose counterexamples. The randomized deployment of an AI feedback agent at ICLR 2025 showed that machine-generated suggestions could make a substantial fraction of peer reviews more specific and actionable, without transferring the final editorial decision to the system (Thakkar et al., 2026). This is a promising pattern because it uses AI to strengthen human criticism rather than to manufacture an automated verdict. The critic's output remains inspectable, and the reviewer remains responsible for accepting or rejecting it.

The independence of criticism matters as much as its sophistication. If one model family proposes a claim, writes the implementation, selects the evidence, interprets the result, and evaluates the manuscript, apparent agreement may be a common-mode error. Different agents with different prompts are not necessarily independent if they share training data, retrieval sources, code, evaluators, and institutional incentives. Executable science therefore needs dependency fingerprints. A review object should record which models, datasets, prompts, tools, and human operators contributed to each judgment. Independence should be treated as an evidential property, not inferred from the number of personas in a multi-agent conversation.

Causal attribution is a second scarcity. Research systems can find correlations, optimize outcomes, and construct elaborate explanations, but scientific value often depends on assigning the observed effect to the right cause. When a complex method improves a benchmark, the improvement may come from more compute, a data leak, a changed preprocessing step, a stronger baseline implementation, or an evaluator that sees only a proxy. The appropriate response is not merely to rerun the experiment. It is to design ablations and controls that transfer causal credit from the overall system to the components that matter.

This is where specification-driven automation can outperform unaided workflow. Once the claim is typed, the workbench can associate it with standard obligations. A representation claim may require a simpler-carrier test. An architecture claim may require matched capacity and data. A causal policy claim may require an explicit estimand, identification assumptions, sensitivity analysis, and alternative causal graphs. A clinical claim may require pre-specified endpoints, adverse-event monitoring, subgroup analysis, and external validation. The workbench cannot determine all these

obligations universally, but it can make them explicit and prevent a successful local score from silently expanding into a stronger claim.

Accountable authority is the third scarcity. Scientific claims are not all equally consequential. A conjecture released as a speculative note imposes different obligations from a recommendation that changes clinical care, environmental policy, or critical infrastructure. Even if an AI system could estimate evidential strength, it could not legitimately grant itself social permission to act. Release and deployment require a named person or institution that has the legal, professional, and ethical authority to accept responsibility.

Authority should not be confused with truth. An authorized institution can be wrong, captured, or negligent. The point is procedural: consequential decisions must remain contestable and attributable. An evidence-bearing system should record who defined the objective, who approved the method, who reviewed the result, who authorized release, and under which policy version. It should also record the procedure by which a claim can be challenged, corrected, superseded, or withdrawn. In a high-speed research environment, the ability to stop or reverse propagation becomes part of scientific reliability.

This perspective changes the role of human experts. The human is not valuable merely because humans are slower or because tradition requires a signature. Expertise becomes most valuable at points where evaluators are weak, stakes are high, or representations are contested. A statistician may inspect identification and uncertainty. A domain scientist may judge whether an assay captures the intended phenomenon. A systems expert may test operational scaling. A patient or community representative may identify harms omitted by a technical objective. A research-integrity officer may decide whether provenance and disclosure are sufficient. No participant needs omniscience, but the scope of each judgment must be explicit.

The same scarcity appears at the level of the scientific system. Hao and colleagues' analysis of more than forty million papers reports an association between AI use and increased individual output and impact, alongside a contraction in the collective range of topics studied (Hao et al., 2026). The result should be interpreted cautiously because observational classification cannot establish every causal mechanism. Nevertheless, it illustrates a plausible institutional effect: AI may accelerate work where data, benchmarks, and existing methods make evaluation easy, while drawing attention away from questions that are poorly instrumented or conceptually immature. Individual productivity can rise while collective exploration narrows.

The validation economy therefore includes attention allocation. Specialist time, laboratory access, participant risk, high-quality datasets, and public trust are limited resources. A responsible automated system should optimize the expected information value of consuming them. Before requesting an expensive experiment, it should ask whether a cheaper simulation, negative control, formal check, or independent reanalysis could eliminate the branch. Before escalating a claim to a senior specialist, it should pass inexpensive validity gates and state the narrow unresolved question. Before publishing, it should estimate not only novelty but the review burden imposed on the community.

This suggests a more demanding productivity metric. Instead of papers per researcher or hypotheses per compute unit, one might measure human validation hours per surviving claim, time to the first valid counterexample, independent reproduction

rate, the fraction of claims demoted by simpler explanations, compute and energy per externally supported result, or the expected information gain of the next experiment. No single metric should become a universal target; Goodhart's law applies with particular force to science. The point is to direct measurement toward the scarce functions rather than the abundant ones.

The practical implication is that automated science should be asymmetric. Generation can be broad and inexpensive. Promotion should be narrow and increasingly expensive. Early branches can be speculative, but evidential thresholds should rise as a claim approaches public authority or operational consequence. The architecture should make it easy to generate alternatives and hard to erase disagreement. It should make it cheap to abandon a weak branch and costly to promote a claim without new kinds of evidence.

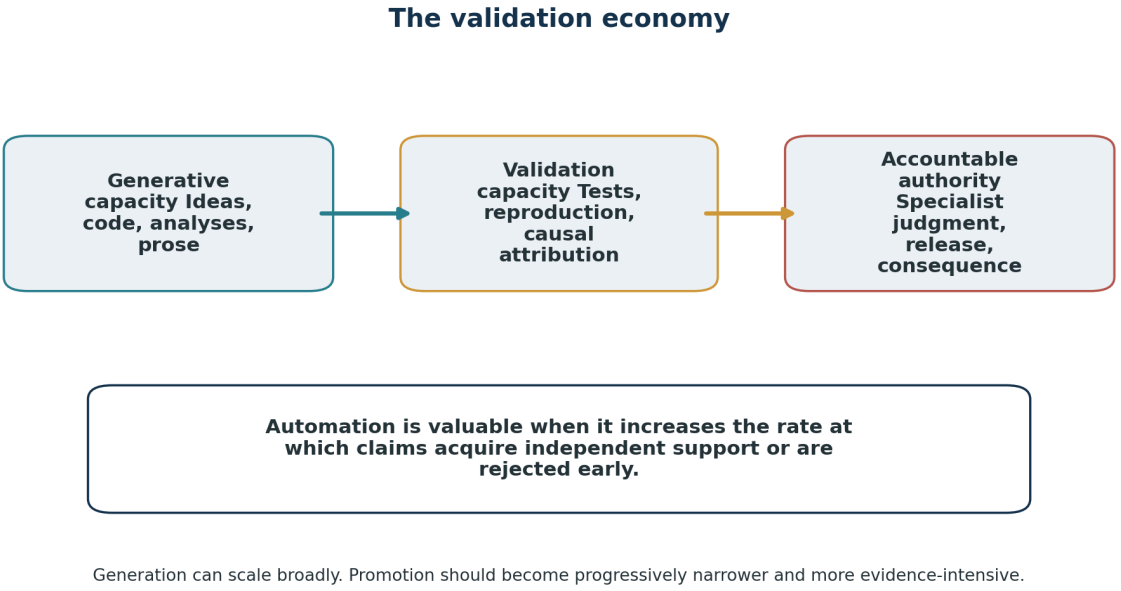


Figure 1. The validation economy. Generative capacity can scale faster than causal attribution, independent criticism, specialist attention, and accountable release.

2.3 Mechanical Loops and Meta-Rational Science

The phrase "mechanical mind" is often used as a synonym for machine intelligence. That usage hides the more important danger. A mechanical research loop is any process, human or artificial, that continues to optimize a local proxy after the proxy has diverged from the purpose of inquiry. A language model can do this by optimizing plausibility. A laboratory can do it by repeating an assay whose operational meaning has been lost. A benchmark community can do it by improving scores on a task that no longer discriminates useful capability. A researcher can do it by maximizing publications, and an institution can do it by rewarding whatever is easiest to count.

The problem therefore predates AI. Scientific institutions already contain rituals, proxy measures, publication biases, and incentives for positive, legible results. Smaldino and McElreath model how selection pressures can favour poor methods even without individual fraud (Smaldino and McElreath, 2016). Ioannidis describes conditions under which published findings are likely to be false, while the reproducibility movement documents the effects of flexible analysis, underpowered studies, incomplete reporting,

and selective publication (Ioannidis, 2005; Nosek et al., 2015; Munafò et al., 2017). Generative AI can intensify these tendencies because it reduces the cost of producing the forms that institutions reward.

A mechanical loop is dangerous precisely because it can be locally competent. Every step may look reasonable. The literature agent finds relevant papers. The planner proposes an experiment. The coding agent implements it. The evaluator reports improvement. The reviewer agent identifies no major issue. The manuscript becomes clearer with each revision. Yet the system may have silently changed the question, selected an evaluator that rewards the preferred mechanism, ignored a simpler explanation, or exhausted its budget on a branch that no longer serves the original objective. Local coherence does not guarantee global fidelity.

The opposite capacity is meta-rational design. Ordinary reasoning asks what follows inside a representation, model, or method. Meta-rational design also asks why that representation is being used, what it excludes, which alternative could explain the observation, when the evaluator has become uninformative, and whether the level of abstraction remains appropriate. It treats the choice of model, tool, critic, and stopping rule as part of the research problem rather than as fixed infrastructure.

This idea is related to rational metareasoning, which asks how an agent should allocate limited computational resources among reasoning strategies, and to model pluralism, which rejects the assumption that one representation captures every scientifically useful aspect of a system (Lieder and Griffiths, 2017; Veit, 2020). No-Free-Lunch results provide a formal caution against expecting one optimizer to dominate across all possible problem structures (Wolpert and Macready, 1997). In practice, the lesson is not that every project requires maximum heterogeneity. It is that a research system should be able to detect when its current regime is failing and route the problem elsewhere.

A meta-rational scientific workbench therefore uses heterogeneous intelligence. Language models are strong proposers, translators, summarizers, and interface layers. Theorem provers and SMT solvers provide stronger guarantees under explicit formal assumptions. Computer algebra systems manipulate exact symbolic structures. Probabilistic programs represent uncertainty and latent structure. Causal models express interventions and identification assumptions. Simulators encode domain dynamics. Databases and knowledge graphs maintain structured evidence. Laboratory instruments provide physical observations. Humans contribute tacit knowledge, value judgments, contextual understanding, and accountable authority. The system is intelligent not because one component is universal, but because it can combine and revise these regimes.

Routing among regimes requires a task contract. The system must know whether exactness, reversibility, order sensitivity, uncertainty calibration, latency, memory, energy, interpretability, privacy, or robustness is most important. A method that is superior under one contract may be inferior under another. A theorem-proved algorithm may be irrelevant if the formalization omits the operational constraint. A compact approximate model may be preferable to an exact but intractable representation. A large language model may be the best tool for discovering terminology across fields, while a deterministic parser is the better tool for checking a schema. The contract makes the trade-off inspectable.

Translations between representations are themselves scientific claims. Moving from a natural-language objective to a formal estimand, from a theory to a simulation, from a dataset to a statistical model, or from a mechanistic hypothesis to a laboratory assay can lose distinctions and introduce assumptions. An executable system should declare what is preserved, what is approximated, whether an inverse mapping exists, how uncertainty propagates, and which validation checks apply. This requirement is central to the Specification Ladder developed later in the book. The most consequential errors often occur not inside one representation but at the boundaries between them.

Meta-rational design also protects against decorative complexity. Generative AI makes it inexpensive to connect advanced mathematical language, elaborate multi-agent roles, and sophisticated models to almost any research problem. Some of these connections will be genuinely productive. Others will create explanations whose complexity exceeds their causal contribution. The correct response is not a universal preference for simplicity. It is a demand that added structure earn its place through discriminating evidence. A more complex representation should create stronger proof obligations, better predictions, lower cost, improved robustness, or an explanatory distinction that a simpler alternative cannot provide.

The institutional analogue is equally important. A funding programme, journal, or laboratory can become mechanically rational by optimizing its own metrics. Meta-rational institutions preserve the ability to revise evaluators, create protected spaces for unconventional work, and recognize when high output masks declining diversity or reliability. They treat governance as an evolving constitution rather than a static checklist.

The central thesis of this chapter can now be stated precisely. AI makes the generative and procedural components of science increasingly abundant. The scarce components are the design of decisive criticism, causal attribution, the allocation of specialist attention, and accountable authority. An architecture for automated research should therefore amplify meta-rationality rather than merely increase iteration speed. It should keep objectives, representations, evaluators, and decision rights revisable and visible. Otherwise, the most capable systems may become the most efficient mechanical minds science has ever built.

Table 3. What current systems demonstrate - and what they do not yet establish

Demonstrated boundary signal	Interpretation
PaperBench evaluates replication of 20 AI research papers through 8,316 scored tasks; the best tested agent averaged 21.0 percent.	End-to-end research execution is measurable, but faithful replication remains far from routine even in a digitally native field.
The AI Scientist demonstrated an integrated loop from idea generation to experiments and manuscript production, with one of three generated submissions accepted at an ICLR workshop.	A complete machine-mediated research trajectory is possible under bounded conditions; acceptance of an isolated output is not evidence of general scientific reliability.
The A-Lab autonomous materials platform performed 353 experiments and realized 36 of 57 target compounds over 17 days.	Closed-loop discovery can achieve meaningful laboratory throughput when synthesis, measurement, and evaluators are tightly instrumented.
A randomized study involving more than 20,000 reviews found that reviewers updated 27 percent of reports after receiving AI-generated feedback.	AI can improve human review as a critic and preflight mechanism without replacing the reviewer’s epistemic and institutional authority.

Chapter 3. What, Exactly, Is New?

A claim of novelty is meaningful only after the system declares the memory, representation, equivalence relation, time, and community against which it is judged.

3.1 Six Different Things We Call New

"New" is the easiest scientific adjective for a generative system to produce and one of the hardest for a scientific institution to defend. A model can know that a string is absent from its context or that a sample is distant in an embedding space. It cannot infer from that fact alone that humanity lacks the idea, that the mechanism has no precedent, or that the difference matters. The word hides several distinct claims, each requiring a different record and a different evaluator.

This relational view immediately separates several problems that are often collapsed into one. A system may determine that it has not stored an exact byte sequence. It may estimate that an observation lies far from the probability mass represented in its training data. It may reject an input because it does not fit any learned class. It may propose a new category that compresses a cluster of observations. It may infer a mechanism that explains regularities not captured by existing theory. It may also claim historical priority. These acts require progressively more evidence. Exact memory comparison is mostly a retrieval problem. Scientific novelty is a social and epistemic judgment whose evidence includes prior literature, semantic equivalence, experimental validity and the significance of the difference.

Table 4. Forms of novelty and the evidence each one requires

Form of novelty	What would count as evidence
Unseen instance	A traceable comparison against the system's accessible memory or database.
Statistical anomaly	A deviation measured in a declared representation relative to a reference distribution.
Unknown class	Evidence that no known class provides an adequate fit under an explicit rejection rule.
New concept	A stable abstraction that improves explanation, compression, prediction or intervention across cases.
New mechanism	Discriminating observations or interventions that favour one causal account over alternatives.
Historical priority	A defensible search for equivalent prior work across literature, patents, archives and communities.

The table exists to prevent a recurrent category error. A high anomaly score can support the claim that an input is unusual in a chosen feature space, but it cannot by itself support the claim that science has encountered a new entity. Likewise, absence from a training corpus does not imply absence from the world's literature. By placing each sense of novelty beside the evidence appropriate to it, the table turns an attractive word into a set of testable claims. The practical consequence is that systems should report the level of novelty they can actually defend rather than silently upgrading a local mismatch into a universal discovery.

The role of representation is especially important. A detector that represents animals by colour and silhouette may judge a thermal image of a familiar cat to be

radically new while failing to distinguish a genuinely new species whose visible outline resembles a known one. A language model may regard two proofs as different strings while a mathematician recognizes them as the same argument under a change of notation. A materials system may identify a composition as novel even though its crystal structure and function duplicate earlier work. Novelty is therefore never independent of the features, invariances and equivalence relations built into the comparison.

Equivalence is not a technical detail that can be postponed. In mathematics, one may compare formulas syntactically, extensionally or up to isomorphism. In chemistry, novelty may refer to a molecular graph, a stereoisomer, a synthesis route or a measured property. In scientific explanation, two verbal accounts may be different descriptions of the same mechanism, or nearly identical language may conceal incompatible causal commitments. A discovery system requires domain-specific ways to decide when two candidates are “the same enough.” Without them, an archive fills with superficial variants and novelty becomes an artefact of notation.

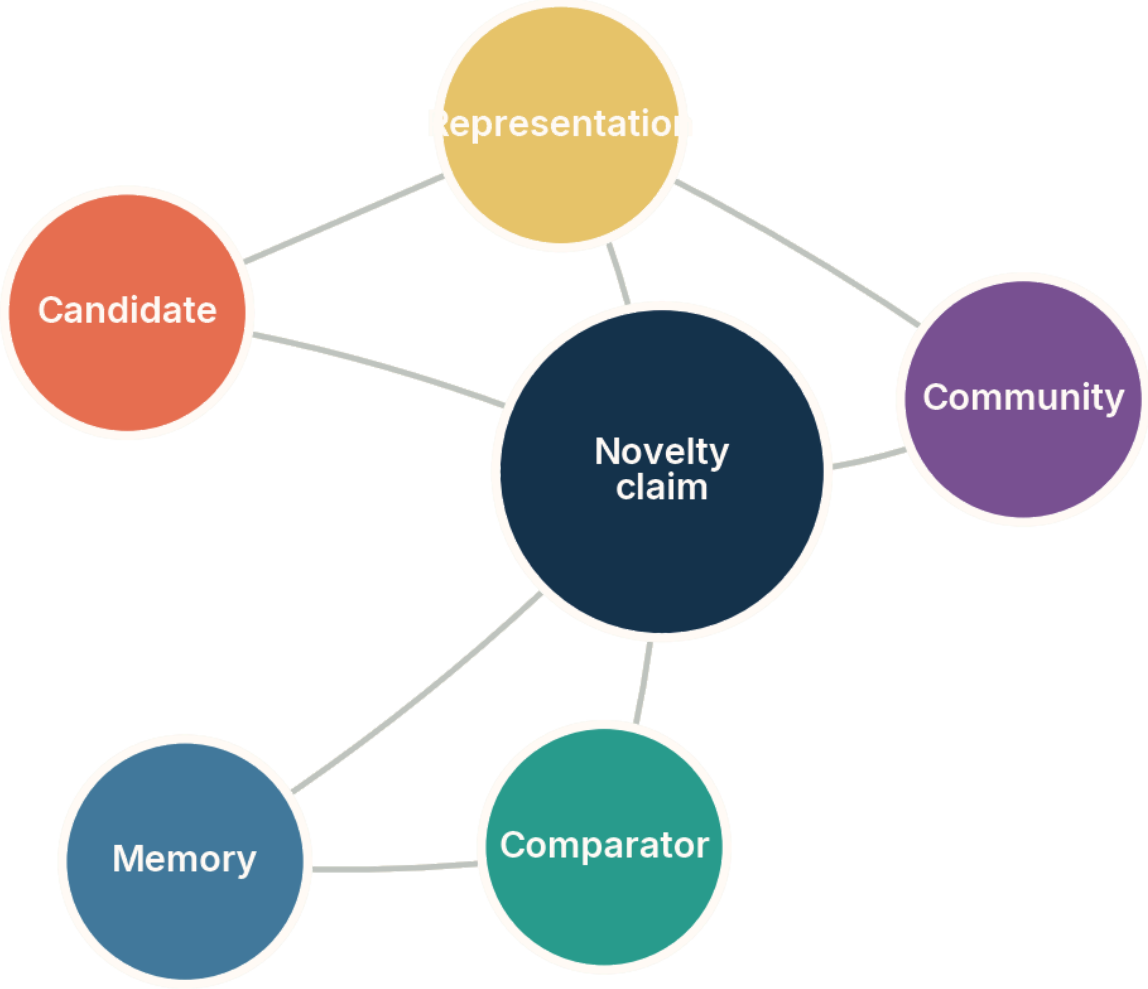


Figure 2. A novelty claim emerges from a relation among a candidate, a memory, a representation, a comparator and a community. Conceptual diagram; not an empirical result.

The figure matters because it locates novelty in a network of dependencies rather than inside the candidate alone. Removing any node changes the claim. Without memory there is no reference set; without representation there is no declared basis of comparison; without a comparator there is no operational criterion; without a community there is no stable convention for significance or priority. The drawing also explains why

improved models do not automatically solve novelty. A more capable generator may produce better candidates while the comparison record, equivalence rule or institutional judgment remains weak.

For machine learning, this distinction suggests a layered reporting discipline. A system can safely state that a sample is absent from a specified database, that its score falls in a low-density region of a representation, or that a classifier abstained because the example did not match known classes. It should reserve stronger language such as “new concept,” “new mechanism” or “first discovery” for cases in which additional evidence has been collected. Such discipline is not merely rhetorical. It determines which errors can be audited and which downstream decisions should be automated.

The deepest point is that novelty is not identical to surprise. A predictable event can be historically unprecedented, and a surprising event can be an instrument failure. Nor is novelty identical to value. Random noise is maximally difficult to predict but scientifically empty. The frontier of machine discovery is therefore not a machine that maximizes unfamiliarity. It is a system that can distinguish unfamiliarity from error, connect differences to explanatory structure, and seek the additional observations needed to turn a provisional novelty signal into reliable knowledge.

3.2 From Surprise to Scientific Contribution

Scientific discourse often treats discovery as the moment when an answer is generated. That image is misleading. A candidate result becomes a discovery only after several thresholds have been crossed: it must be sufficiently distinct from what is already known, correct under the relevant standards, connected to a question that matters, and communicable in a form that others can inspect and reuse. Invention emphasizes construction; discovery emphasizes a warranted claim about a structure, regularity or possibility. Machine learning can participate in either, but the evidential route differs.

Consider an algorithm that performs a computation faster than the best method in a benchmark. The candidate may be invented through search, but its status as a discovery depends on executable verification, comparison against previous algorithms and a precise account of the cost model. Consider a generated molecule with a novel graph. It is not yet a drug, a mechanism or even a synthesizable compound. The candidate must survive chemical constraints, synthesis, measurement and often biological validation. In each case, generation expands the space of possibilities; verification contracts that space into claims that can be trusted.

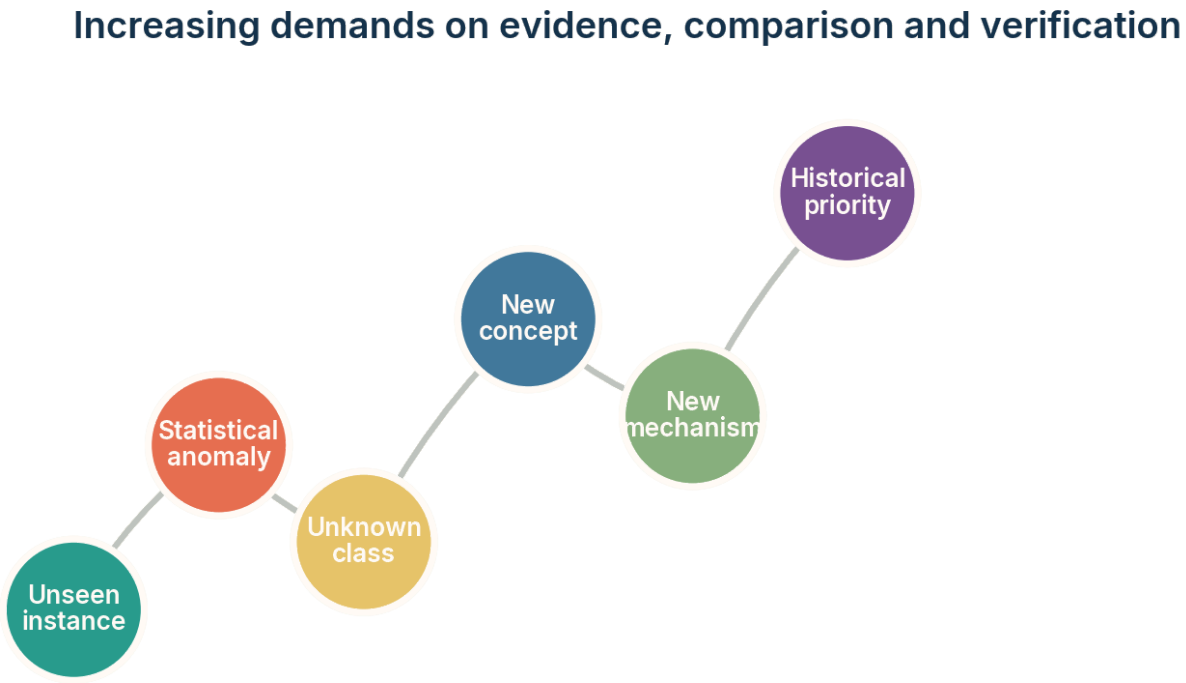


Figure 3. Different novelty claims demand progressively stronger evidence and broader comparison. Conceptual diagram; not an empirical result.

The figure matters because it shows that novelty is not a single threshold. A system may be excellent at the left side of the spectrum and unreliable at the right. Memory systems and nearest-neighbour methods can often detect exact or approximate repetition. Statistical detectors can rank atypicality. Open-world classifiers can sometimes reject unfamiliar categories. New concepts and mechanisms require explanatory work, while historical priority requires an external record that no model can make complete. The increasing evidential burden is the reason that discovery architectures need multiple evaluators rather than one universal novelty score.

Surprise occupies an ambiguous position in this spectrum. In Bayesian terms, an observation is surprising when it has low probability under a current model. That is useful because surprise can identify where a model is inadequate. Yet a low-probability observation can arise from sensor corruption, sampling bias, adversarial manipulation or a rare but understood event. Scientific surprise becomes productive only when a system asks why the observation was improbable and identifies tests that distinguish model failure from data failure. The value lies not in the magnitude of surprise but in the quality of the revision it provokes.

Significance introduces another distinction. A candidate can be novel without altering what anyone can explain or do. Search systems are particularly vulnerable to this problem because large generative spaces contain endless variants that are different under a weak metric. The concept of minimal description length provides one useful intuition: a new hypothesis is valuable when it compresses observations or replaces a collection of exceptions with a simpler reusable structure. Predictive improvement, causal control, proof simplification, robustness across environments and the opening of a new design region are other possible standards. No single standard applies everywhere, but every discovery claim needs one.

Table 5. Four thresholds between a generated candidate and a scientific contribution

Threshold	Question that must be answered
Distinctness	Is the candidate meaningfully different under a domain-appropriate equivalence rule?
Validity	Does it satisfy formal, executable, empirical or observational constraints?
Significance	Does it improve explanation, capability, compression, prediction or intervention?
Priority and provenance	Can the lineage of the claim, data, code and comparison with prior work be reconstructed?

The table exists because contemporary artificial-intelligence systems often leap from distinctness directly to contribution. A language model produces an unusual hypothesis, another model rates it as plausible, and the result is described as novel. The missing thresholds are exactly where hallucination, duplication and triviality enter. Validity must be grounded in an evaluator that is at least partly independent of the generator. Significance must be defined relative to a scientific objective rather than stylistic impressiveness. Priority requires a documented search and remains provisional even after careful review.

These thresholds also clarify why automated discovery is uneven across domains. In program synthesis, a test suite or formal specification can reject many incorrect candidates cheaply. In theorem proving, a proof assistant can check a derivation line by line. In materials science, evaluation may require expensive synthesis and characterization. In social science, the relevant mechanism may depend on institutions and unobserved variables, making validation slower and more ambiguous. The same generative model can therefore appear brilliant in one domain and speculative in another because evaluator strength, not linguistic fluency, is the limiting resource.

The distinction between discovery and invention further suggests complementary roles for humans and machines. Machines can search enormous combinatorial spaces, preserve diverse candidates and execute repetitive tests. Humans remain unusually effective at redefining questions, recognizing changes in significance, negotiating standards of evidence and noticing when an apparently technical result alters the conceptual map of a field. This division is not fixed, but it reminds us that automating candidate production is different from automating the social process by which a community decides what deserves to count as knowledge.

A machine does not discover merely by saying something that was not in its training data. It discovers when a traceable process turns a candidate into a claim that survives comparison, criticism and contact with the world.

The constructive agenda follows directly. Research should improve novelty databases and semantic equivalence checks, design evaluators that cannot be easily gamed, represent uncertainty about priority, and build systems that actively request missing evidence. It should also measure the diversity and significance of discoveries rather than only the volume of generated output. The relevant frontier is not unrestricted originality. It is disciplined expansion of the space in which reliable claims can be made.

The distinction can now be made operational. A novelty claim needs a dated knowledge state, a declared representation, and an account of the transformations under which earlier work would count as a reconstruction.

Novelty is not an intrinsic property carried by an idea in isolation. An idea is new relative to what was known, accessible, and representable at a time. The same contribution can be novel to a field, familiar in another discipline, independently rediscovered by a group that lacked access to the precedent, or old in mechanism but new in evidence and capability. Any automated judgment that omits the reference class will convert a relational concept into a misleading label.

The reference state should be written as K_t, a representation of prior knowledge at time t. This notation immediately raises questions that ordinary novelty scores suppress. Which languages, source types, and archives are included? Does K_t contain only published material or also preprints, patents, code, conference talks, and correspondence? How are inaccessible or undigitised sources represented? Which conceptual vocabulary and level of abstraction are used? A novelty claim is only as meaningful as its answer to these questions.

A structured conception treats novelty as resistance to reconstruction from K_t. Suppose a candidate idea I is compared with sets of prior components S and with admissible transformations T drawn from a grammar T_Q. One possible research objective is

$$R_Q(I|K_t) = \min_{S \subseteq K_t, T \in T_Q} \left[C(S,T) + D_Q(I, T(S)) \right].$$

The term $C(S,T)$ measures the complexity of selecting and transforming prior components; D_Q measures the remaining mismatch under the question-specific representation. A low value means that the candidate can be economically reconstructed from known material. A high value means that the best available reconstruction is costly or leaves substantial unexplained structure. This formulation draws inspiration from minimum-description-length reasoning (Rissanen 1978; Grünwald 2007), but it should not be interpreted as a complete definition of scientific novelty. Compressibility depends on the code, the representation, and the available components. A random or incoherent object can also be difficult to compress.

The scientifically useful output is not the scalar R_Q alone. It is the reconstruction itself and the location of its residuals. The system should show which goals, mechanisms, assumptions, relations, evidential forms, or capabilities are inherited and which remain unexplained. It should then test the stability of those residuals under alternative decompositions, larger corpora, and different transformation grammars. A novelty claim that disappears when one adjacent field is added to the corpus is field-relative, not necessarily false. A residual that survives multiple representations is more interesting, but still requires interpretation.

Figure 4 makes the logic explicit. The candidate, prior knowledge state, and transformation grammar jointly determine the best reconstruction. What remains is a role-specific residual. The diagram then branches into four competing interpretations: substantive novelty, missing precedent, representational mismatch, and extraction error or incoherence. The same numerical residual is compatible with all four. Scientific validity depends on separating them.

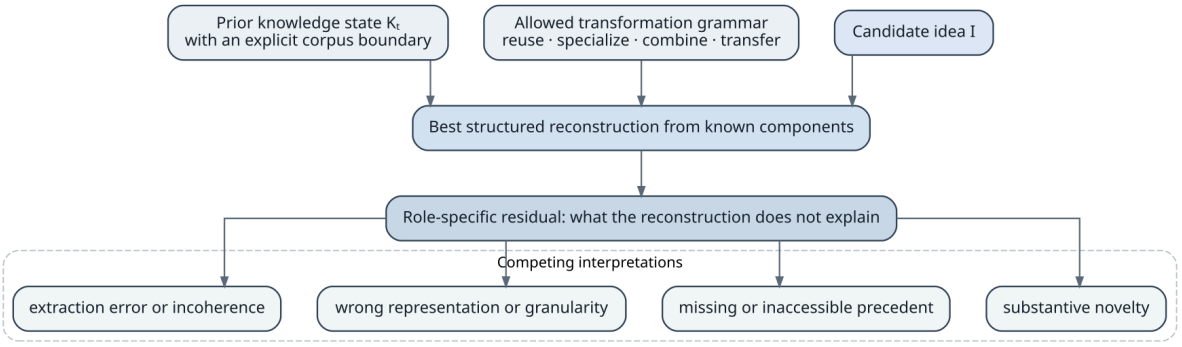


Figure 4. Structured novelty inference. A residual becomes evidence of novelty only after alternatives involving missing knowledge, poor representation, and error have been examined.

This approach yields a multidimensional novelty profile. A contribution may introduce a new problem, a new mechanism, a new combination, a transfer to a new domain, a new evidential technique, or a capability that was previously unavailable. These dimensions should not be collapsed prematurely. The difference between a new mechanism and a first effective implementation matters for scientific understanding and for credit. It also matters for AI training: the model should learn that “not previously demonstrated” and “not previously conceived” are different claims.

Table 6 identifies common sources of apparent novelty and the interpretation they require. The table is designed as a diagnostic aid. It shows why high semantic distance or lack of a retrieved match cannot serve as a verdict.

Table 6. Apparent novelty has several causes that must be distinguished before a contribution can be characterised.

Source of apparent novelty	Required interpretation
New terminology for an established structure	Test paraphrase and role alignment across vocabularies before inferring conceptual change.
Transfer of a known mechanism to a new domain	Separate mechanistic inheritance from novelty in assumptions, integration, evidence, or capability.
Unusual recombination of familiar components	Analyse whether the organisation creates a new causal or functional property rather than rewarding rarity alone.
Finer implementation detail	Determine whether the detail changes performance, validity, or capability at the resolution relevant to the inquiry.
Missing retrieval result	Examine corpus coverage, language, indexing, and representation before treating absence as evidence.
Contradictory or incoherent proposal	Difficulty of reconstruction may reflect error rather than creativity; scientific constraints must be applied.
First persuasive validation	Distinguish novelty of evidence and practical realisation from novelty of the underlying principle.
Removal of a restrictive assumption	A local structural change may have major value even when most components are inherited.
Newly available enabling technology	Credit may belong partly to implementation or infrastructure rather than to a previously articulated mechanism.

The table directs attention from rarity to transformation. Scientometric research on atypical combinations is informative at scale, but unusual co-occurrence is only a proxy for conceptual novelty (Uzzi et al. 2013; Fortunato et al. 2018). A rare combination can be arbitrary; a common combination can become scientifically decisive when one assumption is changed or one relation is reorganised. The atlas should use population-level signals to guide retrieval, not to replace mechanism-level analysis.

Recent work on automated novelty assessment reinforces this caution. Literature-grounded systems improve on unguided model judgments by retrieving and comparing earlier work (Shahid et al. 2025). Yet benchmarks show that persuasive reasoning does not guarantee agreement with experts (Schopf and Färber 2026). The open-world character of novelty means that no benchmark can certify global absence of precedent. Evaluation must therefore measure evidence coverage, calibration, and robustness to corpus expansion in addition to agreement with a static label.

3.3 Priority, Value, and Epistemic Gain

The desire to identify the “first appearance” of an idea is understandable. Priority structures citation, reputation, historical narrative, and sometimes legal or commercial claims. It is also one of the places where an atlas could do the most harm if it presents a clean answer unsupported by the record. The scientifically defensible target is narrower: the earliest known attestation within an explicitly described corpus and interpretation.

Let $A_C(I)$ denote the verified attestations of an operational idea I found within corpus C. The earliest known attestation is

$$a^*(I, C) = arg\,min_{a \in A_C(I)} date(a).$$

This equation is simple, but every term is contested. An attestation must be judged relevant under an identity model. The date may refer to composition, circulation, presentation, publication, or a later surviving copy. The corpus may exclude languages, archives, oral traditions, or proprietary records. Even the apparent minimum can change when a source is redated or newly discovered. The output must therefore report C, the dating evidence, the identity relation used, and unresolved alternatives.

Priority should be decomposed into several events that are often conflated. The first surviving formulation is not necessarily the first conception. The first public disclosure may precede formal publication. The first mathematical proof may follow an informal insight. The first working implementation may reveal constraints absent from earlier proposals. A later author may independently rediscover the mechanism and supply the first evidence that makes it scientifically consequential. A one-field “inventor” attribution suppresses this structure.

Historical evidence is censored in systematic ways. Documents are lost, archives remain closed, publications in dominant languages are more heavily indexed, and the contributions of less prestigious institutions are less likely to be cited and preserved. The Matthew effect amplifies visibility once recognition has accumulated (Merton 1968). A computational system trained on existing citation patterns can reproduce this inequality while appearing objective. Priority analysis must therefore include a model of coverage and a distinction between “not found” and “known not to exist.”

The system should represent priority claims as arguments. An argument links an identity judgment to one or more dated attestations, records the sources searched, and states why competing candidates were excluded. It also records challenges: an earlier source, an alternative translation, a claim that the abstraction is too broad, or evidence of independent development. The result is not a permanent label but a versioned historiographical object.

Influence and priority must remain separate. An earlier attestation can establish chronological precedence without showing that later researchers knew it. Conversely, a later textbook or intermediary may have been the actual source of influence even though it was not the first formulation. The atlas should allow a lineage to pass through transmission nodes that are not conceptually original. This distinction is valuable because science advances through both invention and diffusion.

Legal concepts such as patent novelty and inventive step provide useful analogies but should not be imported uncritically. Patent systems evaluate claims under jurisdiction-specific doctrines and evidential standards. Scientific originality includes explanatory, methodological, and evidential contributions that do not map cleanly onto patentability. An atlas could support prior-art search, but its scientific ontology should not be determined by legal categories.

A mature priority model would assign uncertainty for several reasons. It could express uncertainty that the source contains the same idea, uncertainty in dating, uncertainty in corpus coverage, and uncertainty about transmission. These should remain separate because they demand different responses. A semantic dispute requires expert interpretation; a dating dispute requires archival work; a coverage problem requires broader search; an influence dispute requires documentary and social evidence.

The value of such a system is not that it will end priority disputes. It can make them more productive by exposing where the disagreement lies. It can also reduce the ease with which polished recent formulations erase older or parallel traditions. The ethical standard should be conservative: claims about earliest appearance and credit require stronger evidence than claims about structural similarity.

Originality alone is not scientific value. Creativity research commonly pairs originality with effectiveness or appropriateness (Runco and Jaeger 2012; Boden 2004). In science, usefulness is plural. An idea can be valuable because it explains more with less, predicts accurately, enables control, makes a difficult measurement possible, opens a new experimental regime, reduces cost, unifies separate domains, reveals a boundary condition, or generates further productive questions. These achievements are not reducible to citation count and may emerge long after publication.

Novelty and value should therefore be represented by separate profiles. A highly original idea may be untestable, incoherent, or irrelevant. A modest modification may remove a bottleneck and transform practice. Stokes's analysis of use-inspired basic research illustrates that fundamental understanding and practical use are not opposite ends of one scale (Stokes 1997). The atlas should preserve such multidimensionality rather than rank every idea on one axis.

One family of value measures concerns epistemic gain. An idea may reduce uncertainty, discriminate between theories, compress a set of observations, or reveal which experiment would be most informative. Another concerns capability: it may make a prediction, intervention, computation, or observation possible that was not feasible before. A third concerns robustness and scope: the idea may continue to work under broader conditions or fail more predictably. A fourth concerns generativity: it may supply a reusable representation or mechanism from which further work can be derived.

These dimensions interact with evidence. Theoretical promise should not be scored as established utility, and early negative results should not automatically erase long-term value. The system needs temporal states such as conjectured, demonstrated

under limited conditions, replicated, generalised, contradicted, and superseded. It should also represent the cost and risk of obtaining further evidence. A high-potential idea with expensive or hazardous tests occupies a different position from a cheap, immediately testable one.

Prediction of future usefulness is intrinsically difficult because value depends on social and technological contingencies. A mathematical technique may wait decades for suitable computation. A mechanism may become important only after a measurement instrument is invented. Institutional attention can create feedback loops in which funded ideas acquire evidence and unfunded ones remain speculative. The atlas should therefore avoid training a utility predictor solely on past recognition. That would turn historical advantage into a normative rule.

The science of science shows that research systems can be biased against novelty and that researchers balance risky innovation with conventional strategies (Foster, Rzhetsky, and Evans 2015; Wang, Veugelers, and Stephan 2017). An atlas could either mitigate or intensify this tendency. It could reveal overlooked precedents and unexpected connections, broadening the search space. It could also become a gatekeeping instrument that penalises ideas because they do not resemble recognised successful trajectories. Governance must therefore separate exploratory support from high-stakes evaluation.

A useful design principle is Pareto presentation. Instead of computing one total score, the system displays how a proposal lies across novelty, evidence, expected information gain, feasibility, scope, cost, and risk. Different users can then apply explicit priorities. Peer reviewers may focus on evidential support and contribution relative to the literature; a research programme may accept lower feasibility for high information gain; an engineering team may prioritise robustness and cost. The atlas supplies structured comparison without pretending that one value function governs science.

Ex ante and ex post value must also be distinguished. Before an experiment, a hypothesis may be valuable because it partitions the space of possibilities and supports a decisive test, even if it later proves false. After the experiment, the same hypothesis may be remembered only as an error. Conversely, a result that appears minor at publication can become enabling when another technology changes the surrounding possibility space. A scientific atlas should preserve the decision context in which an idea was proposed: what was known, which alternatives were live, what evidence was affordable, and what outcome the inquiry could discriminate.

This makes the value of negative results conceptually central. A failed intervention can eliminate a mechanism, expose a hidden assumption, or identify a boundary beyond which an analogy breaks. Such contributions may have low novelty in their components and high epistemic value in the constraint they establish. Counterexamples, impossibility results, calibration failures, and replications that narrow an effect are not merely deficient successes. They reshape the space in which later ideas can be evaluated. An atlas focused only on positive capabilities would systematically misrepresent scientific progress.

Expected information gain offers one formal lens, but it cannot serve as a universal value measure. Its result depends on the prior distribution, the question, and the available action set. A cheap experiment that sharply updates a local model may have greater immediate information value than a visionary theory whose consequences cannot yet be observed. The atlas can record these dependencies and compare

alternative research moves, but it should not conceal normative choices inside an apparently objective calculation.

Value is therefore relational in a second sense: it is value for a community, problem, and historical state. A technique may be redundant in a well-equipped laboratory and transformative in a resource-constrained setting. A representation may have modest predictive gain but enormous educational or integrative value. Preserving these contexts is necessary both for fairness and for scientific accuracy. Without them, the system would learn that impact is an intrinsic property of ideas rather than an interaction between ideas and the worlds into which they arrive.

Scientific value is a transformation of prior knowledge that changes what can be explained, predicted, tested, built, or connected. Its novelty matters only when that change survives comparison with prior evidence.

Chapter 4. The Scientific Record Has the Wrong Shape

Science publishes papers, but its real units - mechanisms, claims, transformations, evidence, and failures - cross document boundaries.

4.1 The Missing Epistemic Layer

Science stores its public memory in documents, but scientific reasoning rarely respects document boundaries. A mechanism may be scattered across a paper, a protocol, a codebase, and a later correction. One article may contain several contributions with different genealogies. The archive therefore has the wrong shape for many of the questions artificial intelligence is now being asked to answer.

A single paper usually contains several epistemically different contributions. It may formulate a problem, introduce a mechanism, adapt an existing method, report an implementation, derive a theorem, offer an explanation, define an evaluation protocol, and delimit the conditions under which the claims hold. Conversely, one scientific idea may be distributed across a sequence of papers, correspondence, software releases, experimental protocols, and later reconstructions. The identity of the publication and the identity of the idea therefore cross-cut one another. A paper is neither a minimal unit of contribution nor a maximal unit of intellectual continuity.

This mismatch is partly hidden when the task is retrieval. A useful paper can be found even when the system cannot say exactly which component is relevant. It becomes visible when the task requires comparison. Suppose two methods optimise the same objective but use different mechanisms; two theories invoke the same causal organisation under incompatible vocabularies; or a recent paper presents an old principle in a new domain with the first convincing evidence. Whether these count as the same idea is not answered by textual overlap, topical proximity, or citation links alone. The comparison must preserve some relations and ignore others.

The problem is intensified by the use of artificial intelligence in research. Language models can summarise documents and generate plausible descriptions of novelty, but fluent explanation is not evidence that the underlying judgment is correct. Recent benchmarks make this limitation concrete. RINoBench contains 1,381 research ideas with human novelty judgments and finds that model-generated rationales may resemble expert explanations even while the final judgments diverge substantially from the human reference (Schopf and Färber 2026). IdeaGene-Bench represents papers using typed, evidence-grounded idea objects and lineage differences; across its tasks, the strongest evaluated system reached only 27.3 per cent exact accuracy (Zhou et al. 2026). These results should not be read merely as evidence that present models need more training. They reveal that document understanding, compositional comparison, and historical reasoning are distinct problems.

Scientific memory is also asymmetric. Successful, recent, highly cited, English-language documents are easier to retrieve than negative results, technical reports, older books, local journals, or work preserved in other languages. A system that searches only what is readily indexed will mistake visibility for priority and popularity for conceptual centrality. The resulting error is not simply missing data. It affects the ontology inferred from the available record: familiar ideas appear more coherent, more original, and more influential because the traces that would complicate them are absent.

The desired intermediate layer can be described by contrasting several objects that are often treated as interchangeable. Table 7 does not propose rigid essences. It

clarifies which distinctions a system must keep available if it is to answer scientific questions rather than merely organise text.

Table 7. Distinctions that must remain visible in an idea-centred representation.

Epistemic object	Why it cannot be reduced to “an idea” without qualification
Word or phrase	A linguistic expression can be ambiguous, can change meaning through time, and can name several distinct mechanisms.
Topic or concept label	A topic groups discourse, but does not specify a claim, causal organisation, procedure, or contribution.
Proposition or claim	A truth-apt statement captures what is asserted, but may omit the mechanism, purpose, evidence, and method by which it matters.
Mechanism	A mechanism describes organised entities and activities producing a phenomenon, but may support many claims and implementations.
Method or procedure	A method specifies how something is done, yet the same operational principle may admit many procedures and one procedure may combine several principles.
Model or theory	A model structures explanation and prediction at a larger scale than a minimal transferable idea.
Implementation	A concrete realisation carries engineering choices that may be irrelevant to functional or mechanistic identity.
Operational idea	A task-relative abstraction that records selected roles, relations, conditions, consequences, evidence, and provenance for a stated inquiry.

The table explains why the proposed layer cannot consist of names attached to papers. The left column lists objects that scholarly systems already represent with varying success. The right column identifies the information lost when each object is treated as a complete idea. An operational idea is placed last not because it is metaphysically more fundamental, but because it is an explicit synthesis designed to coordinate the others. Its legitimacy depends on whether the synthesis is useful and revisable.

Figure 5 shows the resulting architecture at the level of epistemic commitments. Documentary sources remain primary. Passages, equations, figures, code fragments, and artefacts become addressable attestations. From them, a system or curator proposes task-relative idea objects. Typed relations connect those objects into families and possible lineages. Only then are higher-level inquiries such as identity, novelty, priority, analogy, and explanation posed. The dashed return path matters as much as the forward path: an inquiry may reveal that the evidence is insufficient and force the system back to the sources.

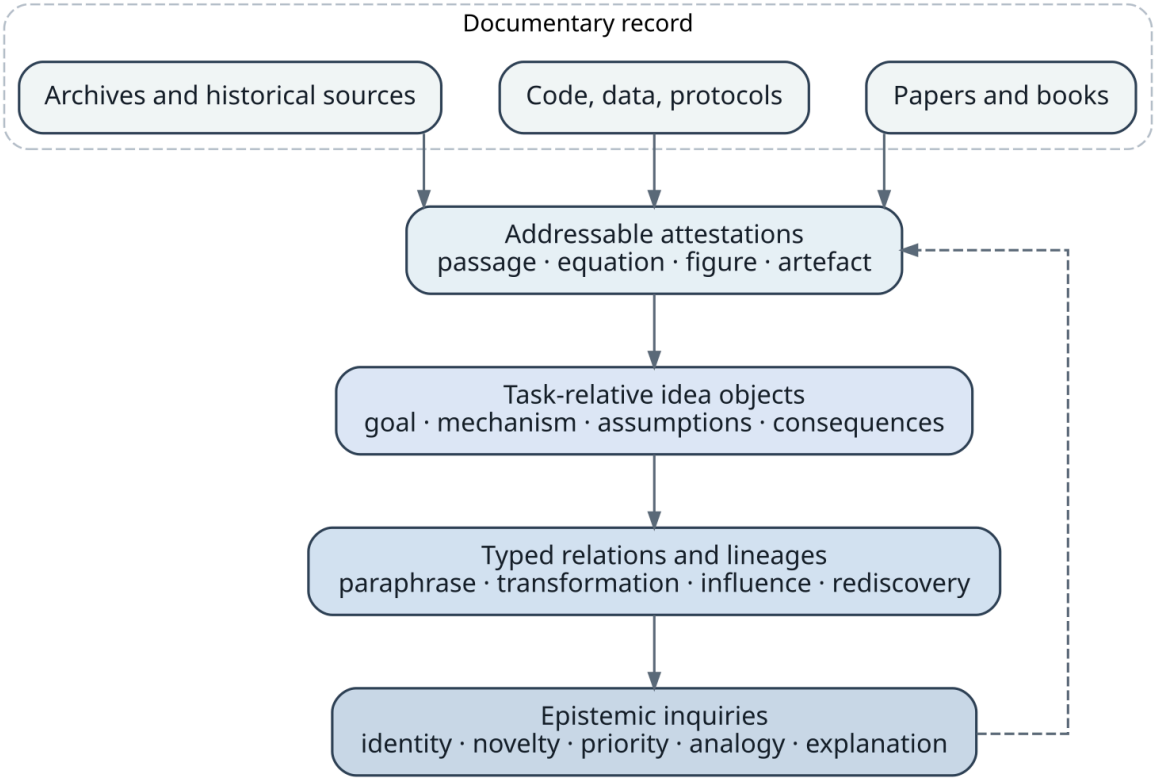


Figure 5. The missing epistemic layer between documentary sources and inquiries about ideas. The return path indicates that comparison can expose missing or inadequate evidence.

This layered picture prevents two common category errors. The first is to treat generated abstractions as if they were present verbatim in the source. The second is to let a similarity score silently stand in for a historical or logical relation. Every movement upward in the figure adds interpretation. Every interpretation must therefore remain linked to the evidence and assumptions that support it.

4.2 Precursors and Their Conceptual Gaps

The ambition to reorganise knowledge around ideas rather than documents has a long history. Wells imagined a continually revised "world brain," while Bush's memex emphasised associative trails through a growing record (Wells 1938; Bush 1945). Adler's *Syntopicon* indexed major ideas across a selected canon (Adler 1952). These projects recognised that alphabetical catalogues and isolated volumes do not match the associative and comparative structure of inquiry. Their limitation was not lack of imagination. They depended on editorial synthesis that could not be executed or contested at the scale and granularity required for contemporary science.

The history of ideas offers a more direct precedent and a warning. Lovejoy's "unit-ideas" aimed to trace recurrent components across intellectual history (Lovejoy 1936). Contextualist criticism, most prominently associated with Skinner, objected that apparent recurrence can be manufactured by removing statements from the linguistic and practical situations in which they performed different actions (Skinner 1969). Betti and van den Berg later argued that computational history need not choose between unstructured particularism and naive universal units. It can make its interpretive models explicit and study continuities relative to those models (Betti and Berg 2014, 2016). This is close to the position developed here: continuity is not read directly from words but argued through a representation whose commitments are inspectable.

Linguistics and cognitive science contribute several forms of semantic decomposition. Frame semantics represents situations through roles and relations, and FrameNet links these structures to attested language (Fillmore 1982; Baker, Fillmore, and Lowe 1998). Prototype theory and family resemblance challenge the assumption that categories are defined by necessary and sufficient features (Rosch 1975; Wittgenstein 1953). Conceptual spaces model concepts geometrically, while formal concept analysis derives lattices from relations between objects and attributes (Gärdenfors 2000; Ganter and Wille 1999). These traditions show that semantic structure can be made computational without assuming that every category has a single crisp boundary. They do not, by themselves, provide a theory of scientific contribution, lineage, or priority.

Pattern languages are particularly relevant to the phrase “fundamental ideas about how things are done.” Alexander and colleagues described recurrent problem–solution structures that could be reused without being copied literally (Alexander et al. 1977). A pattern records context, a recurring tension, a core resolution, and relations to other patterns. This is already more operational than a topic ontology. Yet a pattern language normally abstracts away from the historical evidence needed to determine priority or influence, and it is not designed to compare subtle modifications of mechanisms across scientific literatures.

Knowledge graphs and semantic publishing solve another indispensable fragment. Nanopublications and micropublications make small assertions, provenance, evidence, and argument relations machine-actionable (Groth, Gibson, and Velterop 2010; Kuhn et al. 2018; Clark, Ciccarese, and Goble 2014). The Open Research Knowledge Graph seeks to represent contributions in structured form rather than as undifferentiated prose (Jaradeh et al. 2019). PROV-O supplies a vocabulary for derivation and responsibility, while SKOS supports concept schemes and mappings (World Wide Web Consortium 2013, 2009). These standards make it possible to say who asserted what, from which source, and through which process. They still require a domain theory that determines what the assertion is *about* and when two structured contributions should be aligned.

Scientometrics addresses novelty and influence at a scale inaccessible to close reading. Studies of atypical combinations suggest that high-impact work often mixes a conventional knowledge base with a limited amount of unusual recombination (Uzzi et al. 2013). Other research shows that scientists tend to favour established strategies and that bibliometric indicators can penalise novelty or conflate it with interdisciplinarity (Foster, Rzhetsky, and Evans 2015; Wang, Veugelers, and Stephan 2017). Such work is valuable for population-level patterns. Its proxies operate mainly over citations, journals, keywords, or co-occurrences, and therefore remain several steps removed from mechanistic or functional novelty.

Recent systems move closer to the proposed atlas. Scideator represents papers through facets such as purpose, mechanism, and evaluation and supports recombination grounded in the literature (Radensky et al. 2024). CHIMERA mines instances in which scientific work combines elements from prior concepts and mechanisms (Sternlicht and Hope 2026). Graph2Idea constructs graph-structured contexts of problems, methods, mechanisms, and findings for idea generation (Li, Tu, and Han 2026). IdeaGene introduces minimal typed objects and explicit lineage differences (Zhou et al. 2026). Literature-grounded novelty assessment retrieves and

reranks precedents before producing a judgment (Shahid et al. 2025). Together these systems demonstrate that idea-level structure is no longer only philosophical speculation. They also expose the unresolved questions of identity, granularity, evidential support, and evaluation.

Table 8 maps the principal traditions by pairing their strongest contribution with the unresolved issue that prevents them from constituting an atlas on their own. The purpose of the table is not to rank fields. It identifies interfaces between research programmes that have often developed separately.

Table 8. Existing traditions provide complementary components of an idea atlas while leaving different conceptual gaps.

Research tradition	Contribution and remaining gap
Encyclopaedic and associative knowledge systems	They organise material across documents, but rely on editorial categories that are difficult to formalise, scale, and contest.
History of ideas and conceptual history	They supply contextual interpretation and accounts of semantic change, but have limited computational machinery for evidence-grounded comparison at scale.
Frames, prototypes, and conceptual spaces	They model roles, graded similarity, and semantic organisation, but do not directly represent scientific lineage, priority, or evidential force.
Pattern languages and procedural knowledge	They capture reusable problem–solution structures, but usually omit historical provenance and fine-grained transformation relations.
Scholarly knowledge graphs and semantic publishing	They provide identifiers, provenance, claims, and argument links, but leave the cross-document identity of operational ideas under-specified.
Scientometrics and science of science	They measure population-level novelty and influence, but rely on proxies that rarely expose mechanisms or role-wise differences.
LLM-based scientific ideation and review	They retrieve, abstract, and recombine flexibly, but their judgments can be uncalibrated and their explanations need not track correct decisions.
Lineage and recombination benchmarks	They make inheritance and transformation computationally explicit, but remain domain-limited and do not yet settle historical or semantic identity.

The table suggests that the project is integrative rather than wholly unprecedented. Its possible originality lies in binding these components under a common epistemic discipline. That discipline requires evidence-addressed abstractions, typed identity, explicit historical uncertainty, and tests that distinguish genuine improvement from the production of plausible narratives.

4.3 A Research Claim, Not an Encyclopaedic Promise

The strongest defensible claim is not that all ideas can be placed into one final taxonomy. It is that many scientific tasks can be improved by representing selected contributions as multi-resolution, evidence-grounded structures and by reasoning explicitly over their transformations. This claim can be decomposed into empirical hypotheses without reducing the project to an engineering checklist.

The first hypothesis concerns retrieval. A system that knows the roles played by components of an idea should find precedents missed by lexical and document-level search. A mechanism may have migrated between fields under different terminology; a recent method may change only an assumption that ordinary similarity treats as peripheral; a concept may have undergone semantic drift that makes historical vocabulary misleading. Structured retrieval should improve not only the relevance of results but the explanation of why a result is a precedent.

The second hypothesis concerns comparison. Document embeddings compress many dimensions of similarity into a single geometry. An idea atlas should produce a typed comparison in which purpose, mechanism, assumptions, evidence, implementation, and lineage can vary independently. The advantage is not merely interpretability. Typed comparison changes the answer. Two contributions may be functionally equivalent but mechanistically distinct, or mechanistically similar but historically independent. A single similarity score cannot preserve these alternatives.

The third hypothesis concerns scientific memory. Citation graphs record declared references, not all intellectual dependence, and they mix many reasons for citation. An evidence-grounded lineage model should reconstruct candidate inheritance relations while distinguishing documentary influence, structural resemblance, and independent rediscovery. It should also retain disagreement. Historical scholarship frequently depends on sources outside the canonical article trail; the system must allow a newly found archive or older-language source to revise the map.

The fourth hypothesis concerns novelty. What is new can be understood partly as what cannot be reconstructed from earlier structures under a stated repertoire of transformations. This conception is not intended to replace expert judgment. It creates an object of analysis: which roles are inherited, which are transformed, which appear unsupported by precedent, and how sensitive that conclusion is to corpus expansion or a change of representation. A good novelty analysis would be less categorical than current rhetoric, not more.

The fifth hypothesis concerns artificial intelligence. Scientific AI requires training signals for distinctions that ordinary next-token prediction does not explicitly encode: paraphrase versus mechanistic equivalence, analogy versus influence, inherited component versus novel insertion, and absence of evidence versus evidence of absence. A curated atlas could provide contrastive examples, provenance-preserving memory, and benchmarks for compositional reasoning. Whether this improves discovery must be demonstrated against strong document-based systems rather than assumed.

These hypotheses imply demanding criteria of success. The system must outperform strong baselines on tasks where structure should matter. It must remain stable enough under corpus growth to be useful, while revising itself when genuinely earlier evidence appears. It must report the kinds of uncertainty that affect an answer rather than compressing them into generic confidence. Experts must be able to inspect the relation between a conclusion and its sources. Most importantly, the representation must fail informatively. When it cannot distinguish true novelty from missing evidence or model mismatch, that ambiguity should appear in the output.

Three levels of claim must remain separate. The project does not need the ontological thesis that ideas possess unique natural boundaries. It does require a representational thesis: for specified questions, some decompositions preserve

scientifically relevant invariants better than alternatives. It also advances an epistemic thesis: making those decompositions, alignments, and sources explicit can improve inquiry. A stronger institutional thesis - that such representations should guide review, funding, or credit - would require additional evidence about bias, contestability, and effects on behaviour. Conflating these levels would allow success on a narrow benchmark to masquerade as a theory of thought or a mandate for scientific governance.

Feasibility will also vary by domain. Fields with explicit formalisms, executable procedures, stable artefacts, and well-dated corpora offer favourable conditions for early experiments. Mechanisms can be compared against equations, code, protocols, or causal diagrams rather than inferred from prose alone. Domains in which meaning depends heavily on rhetorical situation, tacit practice, or contested interpretation are harder, but they are not therefore irrelevant. They provide stress tests for whether the atlas can preserve context and disagreement instead of forcing false precision. A credible programme should begin where claims are testable and then examine how far its abstractions travel.

The burden of proof is consequently comparative rather than metaphysical. The atlas must show that its intermediate objects enable distinctions, retrievals, or revisions that are difficult to obtain from documents, embeddings, and citation graphs alone. It must also show that the gain is not purchased by hiding curatorial labour or importing expert conclusions into the benchmark. The research contribution lies in a disciplined account of what is represented, how it is supported, when it changes, and where it ceases to be reliable.

This standard separates a scientific instrument from an encyclopaedic promise. An encyclopaedia tends to present a settled view of its entries. An atlas, in the intended sense, presents a navigable and revisable model of a terrain that exceeds any one representation. Different maps can preserve different invariants. The existence of several legitimate maps is not a defect if their purposes and transformations are explicit. It becomes a defect only when the system silently treats one map as the terrain itself.

Chapter 5. An Idea Has More Than One Identity

Ideas do not have one universal identity: purpose, mechanism, implementation, evidence, and historical descent can align or diverge independently.

5.1 Ontological Pluralism and Operational Ideas

Two contributions can be the same idea and different ideas at the same time. They may pursue the same purpose through different mechanisms, implement the same mechanism under incompatible assumptions, or independently rediscover a structure without any line of influence. The apparent paradox disappears once identity is treated as a typed, task-relative relation rather than a universal sameAs.

The word *idea* covers several ontological categories. In ordinary scientific speech it may refer to a proposition, a conceptual distinction, a mathematical construction, a causal mechanism, an experimental design, an algorithmic strategy, an explanatory schema, or a conjectured possibility. These objects have different identity conditions. Propositions are often compared through logical content. Mechanisms are compared through organised entities and activities that produce a phenomenon (Machamer, Darden, and Craver 2000; Craver and Darden 2013). Procedures are compared through sequences, dependencies, and control conditions. Models may be compared through structure-preserving mappings, predictive consequences, or explanatory roles. An atlas that forces all of them into one undifferentiated type will be simple only at the cost of being wrong.

Ontological pluralism does not imply that comparison is impossible. It means that the comparison must begin by asking what kind of object is under consideration and what relation is sought. The same historical episode can support several representations. The development of backpropagation, for example, can be narrated as a mathematical technique for efficient differentiation, a learning mechanism, an algorithmic procedure, an engineering practice, or a lineage of independent rediscoveries. Each representation selects different evidence and different boundaries. A robust atlas should allow them to coexist and should record mappings between them rather than declaring one to be the uniquely correct idea.

A second source of difficulty is semantic holism. The significance of a scientific statement often depends on a network of assumptions, practices, and contrasts. An expression may retain its words while changing its inferential role. Conversely, two formulations may use entirely different vocabularies while occupying comparable positions in their respective theories. Contextualist history is right to insist that extracting a sentence from its setting can change what act the author performed with it (Skinner 1969). Computational representation must therefore distinguish an attested expression from an interpreted content and from the role that content plays in a broader system.

Prototype and family-resemblance theories offer a useful caution against classical definitions (Rosch 1975; Wittgenstein 1953). Many idea families may have no feature shared by every member. Their continuity can consist of overlapping similarities, recurrent transformations, and a lineage of adaptations. This does not make the family arbitrary. It changes the structure of its identity from a fixed checklist to a network of constrained relations. The identity of “evolutionary computation,” “Bayesian inference,” or “attention” may be organised around prototypes and historical pathways rather than necessary and sufficient properties.

The project must also separate cognitive ontology from representational utility. Human beings may or may not store ideas as discrete symbolic structures; the atlas does not require a claim about the ultimate format of thought. It requires external representations that support reliable inquiry. A map of a mechanism can be useful even if the brain does not encode mechanisms in the same way. This distinction is particularly important for VSA/HDC. High-dimensional vectors may provide an effective computational medium without thereby revealing the metaphysical nature of ideas.

An operational ontology should therefore be layered. At the documentary level, the system records what sources contain. At the interpretive level, it proposes structures that make comparison possible. At the historiographical level, it records claims about continuity, influence, and priority. At the pragmatic level, it represents what distinctions matter for a particular inquiry. These layers constrain one another but should not be collapsed. A located passage is not yet an idea object; an idea object is not yet a lineage; a lineage is not yet an attribution of credit.

This pluralist position has a practical consequence for data design. The atlas should permit more than one decomposition of the same source and should treat disagreement between decompositions as information. Two curators may agree on the cited passage and disagree on whether it contains one mechanism or two. A model may propose a coarse abstraction useful for cross-domain analogy while a domain expert prefers a fine decomposition needed for experimental replication. Preserving both views, together with the question they serve, is more scientifically faithful than forcing premature consensus.

The term *operational idea* names a representational commitment, not a newly discovered natural kind. It is an abstraction designed to be small enough that its components can be compared and transformed, yet complete enough to preserve an intelligible contribution. Its structure should capture at least the roles that make a difference to the intended inquiry. A convenient starting form is

$$I_Q = \langle G, M, C, A, O, E, P \rangle_Q.$$

Here, G represents a goal, problem, or explanatory target; M a mechanism, transformation, or governing principle; C the context or domain; A relevant assumptions and preconditions; O expected outcomes, consequences, or capabilities; E the evidential support; and P provenance. The subscript Q is essential. It indicates that the boundary and salience of the object depend on a question or family of tasks. The tuple is not a universal definition of every idea. It is a schema for making the choices of representation explicit.

Consider a method that uses random projections to preserve approximate distances. A historical inquiry may represent the mathematical guarantee, its earliest formulation, and later rediscoveries. An engineering inquiry may foreground the input regime, computational cost, distortion bound, and implementation. A cross-domain analogy system may abstract further, retaining only the principle that a high-dimensional object can be mapped to a lower-dimensional space while approximately preserving relational structure. These are not merely different descriptions of a fixed object. They may establish different units because different components can vary independently under the relevant comparison.

The goal role prevents a mechanism from being detached from the problem it addresses. The same operation can count as different ideas when it serves different purposes, and different operations can instantiate the same functional idea. The mechanism role prevents functional equivalence from erasing causal or procedural differences. Context and assumptions prevent a representation from claiming generality

that the source does not support. Outcomes connect an idea to what it enables or predicts. Evidence distinguishes a speculative possibility from a validated result. Provenance preserves the path from abstraction back to the record and the process by which the abstraction was produced.

The tuple should not be understood as seven text fields. Each role may itself be a structured object. A mechanism can contain entities, activities, ordering, feedback, and boundary conditions. An assumption can be logical, statistical, material, or institutional. Evidence can include experiments, proofs, simulations, datasets, negative results, and replications. Provenance must identify not only a source but the location and version from which a claim was derived, the agent or model that proposed the interpretation, and any subsequent challenges. The apparent simplicity of the notation is therefore a discipline of organisation, not a denial of complexity.

Operational representation also requires negative and contrastive information. An idea is often understood by what it changes relative to a predecessor. A paper may preserve the goal and evaluation while removing an assumption, replacing a mechanism, or extending the admissible context. If the atlas stores only positive summaries, it will lose the transformation that constitutes the contribution. Similarly, a failed application can reveal a boundary condition more clearly than a successful one. The object should therefore be capable of recording exclusions, limitations, counterexamples, and unresolved alternatives.

This structure differs from a summary. A summary compresses a document for reading; an operational idea supports comparison and inference. It must use stable roles and typed relations that can be aligned across sources. It also differs from an ontology entry. An ontology normally locates a concept within a controlled vocabulary; an operational idea records an organised contribution, its evidence, and its relation to other contributions. Finally, it differs from a claim graph. A claim may be one component of an idea, but the idea can also contain a problem framing, a method of intervention, and a pattern of consequences that no single proposition captures.

The requirement of task relativity prevents the schema from becoming another rigid universalism. The same source can generate several I_Q objects, each with a declared resolution and purpose. The system should retain relations between these representations, such as “refines,” “abstracts,” “focuses on evidence,” or “focuses on mechanism.” This creates a lattice or graph of views rather than a single canonical entry. Formal concept analysis and conceptual-space approaches suggest ways to organise such perspectives, but the atlas must add evidence and transformation history (Ganter and Wille 1999; Gärdenfors 2000).

Task relativity does not mean that anything goes. A representation is constrained by source fidelity, predictive or explanatory usefulness, intersubjective agreement, and performance on downstream tasks. A decomposition that invents an unsupported mechanism is invalid even if it is convenient. A representation so coarse that it cannot distinguish known counterexamples is inadequate for the task. A representation so fine that every wording becomes unique defeats the purpose of comparison. The scientific problem is to discover useful regimes between these extremes.

The question “What is the smallest idea?” is misleading because minimality is relative to admissible transformations. A component is operationally atomic when it can vary independently in the comparisons of interest and when further decomposition no longer improves those comparisons enough to justify the added complexity. Atomicity is therefore a property of a model and a task, not an intrinsic indivisibility of thought.

This definition resembles the use of modules in engineering and mechanisms in philosophy of science. A mechanism can be decomposed into organised parts until the explanatory question no longer requires finer detail. The sodium channel may be a component in one explanation and a complex mechanism in another. The same is true of scientific ideas. “Gradient-based learning” may be one component in a broad history of optimisation, while a technical analysis separates differentiation, update rule, stochastic sampling, regularisation, and stopping criteria.

Granularity creates a bias in novelty judgments. At a coarse resolution, many contributions appear to repeat familiar ideas. At a fine resolution, almost every implementation appears novel. Neither conclusion is reliable without a theory of which differences are scientifically consequential. The atlas should therefore compare candidates across several resolutions and report where the novelty claim emerges. A contribution that is old at the level of mechanism but new at the level of evidence or capability should not be forced into a binary label.

Compositionality introduces a related problem. Some ideas can be represented as combinations of reusable components, but the combination may create properties not predictable from the parts in isolation. Conceptual blending and research on analogy emphasise that mappings and interactions can generate new structure (Fauconnier and Turner 2002; Gentner 1983). A system that decomposes too aggressively may lose the relational organisation that constitutes the idea. Conversely, a system that treats every whole as irreducible cannot recognise inheritance or recombination.

The appropriate unit is therefore often a structured configuration rather than a bag of components. Two methods may contain the same elements but organise them differently. Feedback, ordering, nesting, and conditional dependence can be decisive. For this reason, the transformation grammar of the atlas must include not only insertion and deletion but substitution, reordering, abstraction, specialisation, change of role, transfer of context, and alteration of constraints. A new idea may consist less in adding a component than in changing the relation between familiar ones.

Context poses another boundary problem. Too little context produces false equivalence; too much context prevents transfer. A mechanism that works in a controlled laboratory setting may fail in an open social system because agency, adaptation, and measurement change the causal structure. Yet retaining every local detail would make cross-domain analogy impossible. The atlas needs representations in which context can be progressively abstracted while the consequences of that abstraction are recorded. One view may state that a mechanism is valid under a narrow set of assumptions; another may propose a functional analogy at a higher level and mark the untested correspondence.

Table 9 summarises characteristic granularity failures. Each row pairs a representational choice with the epistemic distortion it creates. The table is not a catalogue of implementation bugs. It describes failure modes that should become experimental variables in the research programme.

Table 9. Granularity choices create systematic forms of false identity and false novelty.

Representational choice	Characteristic distortion
Treating the whole document as one idea	Independent contributions, inherited components, and local modifications become inseparable.

Representational choice	Characteristic distortion
Treating every sentence or claim as an idea	Relational organisation and procedural structure disappear, while wording differences are mistaken for conceptual difference.
Using one fixed level of abstraction	The system alternates between false equivalence in some tasks and false novelty in others.
Decomposing only into named components	Novel organisation, ordering, feedback, and constraint changes are missed.
Retaining all contextual detail	Cross-document alignment and cross-domain transfer collapse because no two cases are sufficiently similar.
Removing context too early	Analogies become superficial and mechanisms are transferred beyond their valid conditions.
Enforcing a single curator-approved decomposition	Disagreement and alternative explanatory perspectives are erased rather than studied.
Allowing unrestricted decompositions	Comparisons become non-reproducible because any desired relation can be manufactured after the fact.

The table shows why multi-resolution representation is not an optional convenience. It is the only coherent response to the fact that scientific contributions have nested structure and that different inquiries preserve different invariants. Multi-resolution does, however, increase the burden of evaluation. The system must learn when a coarse alignment is informative and when it hides a decisive difference. It must also prevent users from selecting a resolution merely because it supports a desired novelty claim.

One promising principle is *contrastive adequacy*. A representation is adequate for a question when it distinguishes the cases experts regard as importantly different and groups cases they regard as equivalent for that purpose. This can be tested using carefully constructed pairs that vary one dimension at a time. Another principle is *transformational economy*: prefer a representation in which observed lineages can be described by a relatively small and stable set of meaningful transformations, without forcing unrelated cases into the same family. These principles turn granularity into an empirical research problem.

The resulting conception is deliberately anti-essentialist but not anti-formal. An idea object is a model with declared scope, evidence, and resolution. It can be encoded, compared, and evaluated. Its boundaries are open to revision, but revision is constrained by source fidelity and performance. The atlas would therefore map not only ideas but the alternative ways in which ideas can be individuated. This reflexive layer is necessary because any system capable of judging novelty must also be capable of questioning the units on which that judgment depends.

5.2 Identity Without a Universal sameAs

The question “Are these the same idea?” invites a binary answer that is rarely adequate. Scientific identity is indexed to a purpose. A historian may ask whether two authors formulated the same proposition, an engineer whether two methods realise the same function, a philosopher whether they posit the same mechanism, and a reviewer

whether a new submission makes a contribution beyond a cited baseline. The answer can change without contradiction because the relation being tested has changed.

A useful representation begins with an identity profile rather than a universal equivalence relation. For two operational idea objects x and y, the system can construct

$$\Sigma_Q(x, y) = \langle s_{surface}, s_{prop}, s_{func}, s_{mech}, s_{context}, s_{impl}, s_{evid}, s_{lineage} \rangle_Q.$$

Each component describes alignment along a distinct dimension: linguistic surface, propositional content, function or goal, mechanism, context and assumptions, implementation, evidential status, and historical lineage. The values need not all be scalar similarities. Some are structured correspondences, some are categorical relations, and some are probability distributions over competing interpretations. The subscript again records that the comparison is made for a question Q.

This profile separates cases that a generic embedding would place near one another. Two texts may be paraphrases but rely on different evidence. Two algorithms may be functionally equivalent while using distinct computational mechanisms. Two experiments may instantiate the same design but answer different questions. Two mechanisms may be structurally analogous while their components have no historical relation. The profile also prevents a common historiographical error: treating similarity in present-day reconstruction as proof that later authors inherited an idea from earlier ones.

The relevant notion of identity can be expressed through invariants. A comparison question selects a set of roles and relations that must be preserved, together with a set of transformations that are allowed to vary. Informally, x and y count as the same for Q when the selected invariants are preserved under an admissible alignment. A compact notation is

$$x \equiv_Q y \quad only\ if \quad Inv_Q(x) \cong Inv_Q(y)\ under\ an\ admissible\ transformation\ in\ T_Q.$$

The symbol $\cong$ is deliberately weaker than literal equality. It represents a structure-preserving correspondence whose definition depends on the object type. The set T_Q may include renaming, change of notation, abstraction, specialisation, substitution of an implementation, or transfer to a new context. It should not be assumed to form a mathematical group. Scientific transformations are often partial, irreversible, and dependent on preconditions. Treating them as a general transformation grammar is safer than imposing algebraic properties they do not possess.

Identity judgments must also be asymmetric in some cases. “Generalises,” “implements,” “is derived from,” and “repairs a limitation of” are directional. Even apparent equivalence may be asymmetric when one representation contains strictly more detail. A coarse functional description can be satisfied by several fine mechanisms, but the fine mechanisms are not thereby equivalent to one another. The atlas therefore needs both equivalence-like families and directional transformation relations.

Figure 6 illustrates the logic. Two idea objects are not compared in the abstract. They enter through a comparison question that determines which invariants matter. Their purposes, mechanisms, assumptions, implementations, and evidence are aligned separately. The output is a typed judgment that records what is preserved and what changes. The figure is intentionally not a funnel toward a single number. It depicts identity as an interpretive composition of role-wise results.

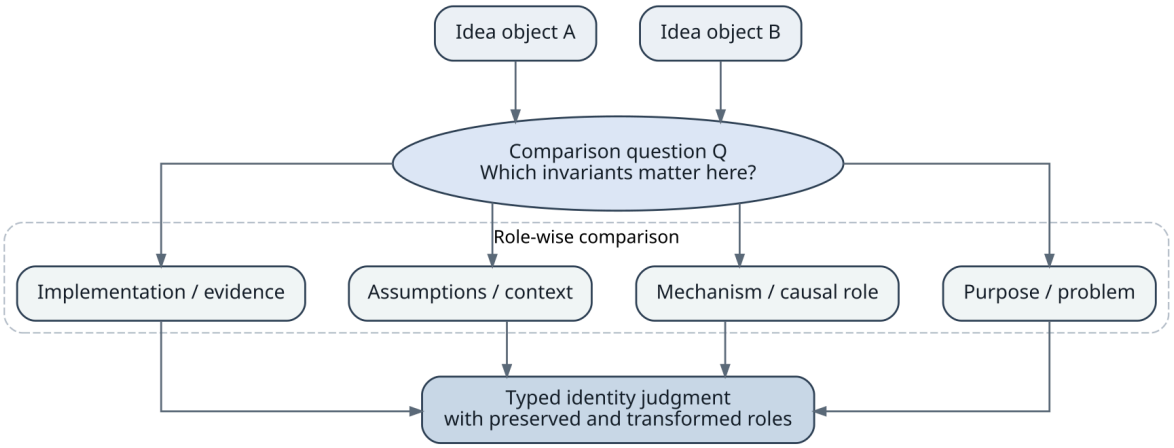


Figure 6. Identity is produced by a question-sensitive comparison of roles, not by a universal similarity threshold.

This account supports both crisp and graded relations. Logical equivalence under a formal language may admit a proof. Functional or mechanistic equivalence often remains graded because descriptions are incomplete and contexts differ. Gradedness should not be confused with vagueness alone. It can represent uncertainty about evidence, partial preservation of structure, or genuine degrees of match. The system should identify which of these interpretations applies.

An important research question is whether identity profiles can be learned without collapsing into latent similarity. Supervision must therefore include contrastive cases in which texts are similar but mechanisms differ, and cases in which texts are dissimilar but a domain expert recognises the same causal organisation. Evaluation should test the correctness of the alignment itself, not merely agreement with a final label. A model that predicts “same mechanism” for the wrong reasons will fail when transferred to new contexts.

A universal sameAs relation is attractive because it simplifies graphs. Once two nodes are declared identical, their properties can be merged. In an atlas of ideas this operation would be dangerous. The identity of an idea depends on what information is inherited across the relation, and different relations license different inferences. A paraphrase permits linguistic substitution. Functional equivalence supports expectations about outputs under stated conditions. Mechanistic equivalence supports a stronger inference about how those outputs are produced. Historical influence supports claims about lineage and credit. None can automatically substitute for the others.

Typed relations also make disagreement more tractable. Experts may agree that two methods solve the same problem and disagree about whether their mechanisms are equivalent. Instead of forcing a single decision, the atlas can store the shared functional relation and competing mechanistic interpretations. A later experiment or more precise reconstruction may resolve the dispute. This is closer to scientific practice than a database constraint that demands immediate identity or separation.

Table 10 sets out a core relation vocabulary and, in the second column, the evidential burden associated with each relation. The purpose is to show that relation types are not merely labels. They determine what sources and arguments are required before the relation should be accepted.

Table 10. Typed identity and lineage relations carry different inferential permissions and evidential requirements.

Typed relation	Meaning and evidential burden
Paraphrase or terminological reformulation	The expressions communicate substantially the same content; linguistic and contextual analysis is required, but historical dependence is not implied.
Propositional equivalence or entailment	The claims have the same truth conditions, or one entails the other, within an explicit formalisation and background theory.
Functional equivalence	The objects pursue the same goal or produce comparable outcomes under stated conditions, even if their mechanisms differ.
Mechanistic equivalence	The organised causal or computational roles correspond at the selected level of abstraction; evidence must support the mapping of entities, activities, and dependencies.
Implementation equivalence	Concrete realisations match in operationally relevant choices; this is stronger than sharing a named method and weaker than literal code identity.
Analogy	A relational structure maps between domains while object-level identities differ; structural fidelity and the limits of transfer must be stated.
Specialisation or generalisation	One object narrows or broadens the domain, assumptions, or consequences of another; the direction of inclusion must be justified.
Composition or recombination	A later object integrates components from several antecedents; the contribution may lie in the relations among components rather than in the components alone.
Historical influence	Documentary or contextual evidence supports a causal path of intellectual transmission; semantic similarity is insufficient.
Independent rediscovery	Strong structural correspondence exists while available evidence supports separate development or fails to establish transmission.
Contradiction or replacement	A later object rejects, invalidates, or substitutes for a component of an earlier one; the scope of the conflict must be localised.

The table prevents a semantic graph from becoming overconfident through transitivity. Paraphrase and formal equivalence may be transitive under controlled conditions. Analogy is not generally transitive. Historical influence is directional and can pass through intermediaries, but absence of a direct citation does not settle whether influence occurred. Functional equivalence may fail when contexts are combined. Each relation needs its own composition rules and exceptions.

Typed relations also support a more precise concept of idea families. A family need not be an equivalence class in which every member is mutually identical. It can be a connected subgraph organised around prototypes, recurrent mechanisms, transformations, and lineage. One branch may preserve a mechanism while changing its purpose; another may preserve the purpose while replacing the mechanism. The family is intelligible through the pattern of transformations, not through a feature present in every node.

This conception resembles evolutionary metaphors, including the “idea genomes” proposed in recent lineage work (Zhou et al. 2026), but the metaphor must be used carefully. Biological inheritance has material mechanisms and relatively well-defined

units of reproduction. Intellectual inheritance can be indirect, collective, reconstructed retrospectively, and mediated by languages and institutions. Ideas do not reproduce independently of agents and practices. The value of the metaphor lies in explicit comparison of inherited, modified, lost, imported, and newly inserted components, not in claiming a literal genetics of thought.

A typed graph also enables reasoning about counterfactual identity. Researchers may ask whether a method would remain "the same" if a particular assumption were removed, if its implementation were changed, or if it were transferred to another domain. Such questions reveal the implicit invariants of a concept. They can be formalised as controlled transformations of an idea object and evaluated through expert judgments. This creates a bridge between conceptual analysis and machine learning: the model is trained not only on existing pairs, but on interventions that probe the boundary of the category.

5.3 Similarity, Influence, and Independent Rediscovery

Similarity is evidence for comparison, not a conclusion about history. This distinction is easy to state and difficult to preserve computationally. Retrieval systems naturally rank earlier texts by semantic proximity. When a highly similar predecessor is found, a generated explanation can slide from "resembles" to "derives from," especially if the later work cites related literature. An atlas must keep three questions separate: whether structures are similar, whether one work was historically available to the other, and whether there is evidence of actual transmission.

Historical lineage is an inference from heterogeneous evidence. Direct citation is relevant but neither necessary nor sufficient. Authors cite work for background, criticism, convention, strategic positioning, or reviewer requests. They may inherit an idea through textbooks, colleagues, code, lectures, or a shared research programme without citing the original source. Conversely, two groups may arrive independently at closely related structures because the problem strongly constrains the solution space. The atlas should therefore treat influence as a hypothesis supported by a bundle of temporal, documentary, social, and semantic evidence.

Priority disputes demonstrate why this matters. The earliest published formulation, the earliest private conception, the first public presentation, the first proof, and the first effective implementation may belong to different people. Scientific credit is also shaped by visibility and institutional power. Merton's analysis of priority and the Matthew effect shows that recognition is a sociological process as well as a chronological one (Merton 1957, 1968). Stigler's law of eponymy captures the recurring fact that discoveries are often named after someone other than the earliest contributor (Stigler 1980). A computational atlas should illuminate these distinctions rather than produce a single "inventor" field.

Independent rediscovery deserves a first-class relation because it has scientific value. Repeated emergence can indicate that an idea is a natural solution under shared constraints, that multiple traditions possessed compatible components, or that the time was ripe because enabling evidence or technology had appeared. Collapsing rediscoveries into one node would erase information about the structure of discovery. Treating them as entirely unrelated would miss their conceptual convergence. The atlas should connect them through typed structural identity while keeping their historical paths distinct.

Lineage reconstruction can be represented as an alignment between predecessor and successor objects. The alignment identifies preserved roles, transformations, external imports, and unresolved components. Recent work such as IdeaGene-Bench makes this “diff” perspective operational (Zhou et al. 2026). Its scientific importance is broader than benchmark performance. A lineage difference forces the system to specify what exactly changed. It can reveal that an apparently new theory preserves the explanatory mechanism while changing only the evidence, or that a seemingly incremental method removes an assumption that had constrained an entire class of applications.

The lineage graph should be non-monotonic. New evidence can revise an edge. A newly digitised book may become the earliest known attestation; an archive may establish influence that was previously classified as independent rediscovery; a technical reconstruction may show that two methods thought equivalent rely on incompatible mechanisms. Instead of overwriting the prior state, the atlas should preserve versions, the evidence available at each time, and the reasons for revision. This is essential for scientific accountability and for studying how the atlas itself evolves.

Lineage should be analysed through competing historical hypotheses rather than inferred from one similarity score. For a pair of related works, the main alternatives may include direct transmission, transmission through an intermediary, inheritance from a shared tradition, convergence under common problem constraints, and accidental resemblance. Each predicts a different evidential pattern. Direct transmission is strengthened by temporal accessibility, terminological or diagrammatic traces, explicit acknowledgement, archival contact, and a sequence of distinctive shared choices. Convergence is more plausible when the shared structure follows almost inevitably from the problem while peripheral design decisions differ. A common ancestor is favoured when both works reproduce an earlier package of assumptions but diverge from one another in ways consistent with separate development.

No individual cue is decisive. Citation can establish awareness without proving dependence; absence of citation can reflect convention, memory, strategic omission, or indirect learning. Social proximity can make transmission possible without showing that it occurred. Close wording can be inherited from a textbook rather than from the apparent predecessor. Even chronological priority is weaker than it appears when dates refer to private notes, presentations, preprints, editions, translations, and surviving copies. The system should therefore expose an evidential matrix and the rival explanations it supports rather than collapsing heterogeneous observations into a single edge weight.

This framing makes historical reconstruction resemble causal inference under severe missing data, but the analogy has limits. There is rarely a repeatable intervention that settles intellectual influence, and the relevant variables are themselves interpreted. The practical response is triangulation: semantic alignment, citation and social networks, archival evidence, version history, and expert contextual reading should constrain one another. The atlas becomes valuable not by eliminating historical judgment but by showing which conclusion depends on which evidence and which alternative remains live.

Uncertainty must be typed here as well. There is uncertainty from missing documents, uncertainty from ambiguous language, uncertainty from competing decompositions, uncertainty from model error, and uncertainty about historical

causation. A single probability of “same idea” cannot communicate these sources. A lineage judgment may have high structural confidence and low influence confidence. It may be temporally possible but documentary weak. It may depend heavily on one translation. These distinctions should shape both the interface and the evaluation.

The deepest implication is that identity and history cannot be cleanly separated. What counts as the same idea affects which sources appear to be predecessors; the discovered predecessors can in turn change the representation of the idea. The process is iterative. An initial abstraction guides retrieval, retrieved sources reveal variants, and those variants force the abstraction to be revised. A neuro-symbolic atlas should support this hermeneutic loop while making each revision traceable. The goal is not to eliminate interpretation, but to transform interpretation from an invisible precondition into an explicit and testable component of the system.

Chapter 6. The Paper Becomes a View

The article remains the public argument; the durable record becomes a versioned research object that preserves how claims, evidence, and reasons changed.

6.1 When the Paper Is No Longer the System of Record

The scientific paper is not obsolete. It is overburdened. It has been asked to serve as argument, archive, audit trail, reproducibility package, credit system, and institutional verdict at once. Generative and agentic research break this compression because the process can now branch and revise faster than any final narrative can faithfully remember.

The loss was tolerable while most transformations were slow and human attention was the dominant production constraint. A small research team could reconstruct why an analysis changed, which failed experiment redirected the project, or which informal judgment led to a new model. Code, laboratory notebooks, email, instrument files, preregistrations, and draft manuscripts were fragmented, but the people involved often retained a working narrative. When a paper was released, the compression was accepted as a practical necessity. Reviewers could request additional detail, later researchers could attempt replication, and controversies could unfold over years.

Generative and agentic AI change this balance. A system can now create more candidate hypotheses, code paths, simulations, analyses, visualizations, and prose than a human team can understand in the same interval. In computational domains, an agent can modify a repository, run experiments, interpret outputs, and rewrite a manuscript in a continuous loop. In experimental domains, software can connect literature, planning, active learning, robotics, and measurement. In formal domains, a language model can propose constructions while a theorem prover or programmatic evaluator rejects invalid candidates. The bottleneck moves. Production becomes cheaper; reconstruction and justification become more expensive.

This shift exposes the inadequacy of treating the final paper as the only durable object. Suppose an automated system generates twenty plausible mechanisms, implements eight, finds three positive benchmark effects, eliminates one through a negative control, narrows another after a scaling failure, and retains a third only under a revised claim. A conventional paper may present the surviving line as a clean story. The polished narrative can be honest, but it conceals the branching structure that gives the result meaning. It does not show whether the discarded mechanisms were genuinely tested, whether the evaluator changed, whether the same model family generated and judged the evidence, or whether a specialist objection altered the interpretation. At machine speed, those omissions become operational risks rather than merely historiographical losses.

The analogy with software engineering is therefore more than rhetorical. Modern software is not maintained as a sequence of published screenshots. It lives in repositories. Requirements, source code, tests, dependency declarations, issues, review comments, continuous-integration results, releases, and change histories remain connected. A developer can inspect what changed, who approved it, which tests passed, and which version entered production. Branches support exploration without erasing the main line. A broken dependency can trigger selective rebuilding. The artefact is not only the executable program; it is the governed history that makes the program maintainable.

Scientific knowledge will increasingly need a comparable infrastructure, although its semantics are more demanding. A scientific repository must version more than files. It must version claims, evidential status, assumptions, models, operational definitions, causal interpretations, review obligations, and decision rights. A commit should not merely record that a notebook changed. It should record that a claim was narrowed because a negative control succeeded, that an evaluator was replaced because it could not distinguish competing mechanisms, or that an apparent improvement disappeared under matched compute. The unit of change is epistemic as well as textual.

This is the central meaning of executable science. The phrase does not imply that a computer can decide truth by running a script. It means that a larger part of the path from question to claim is represented as an inspectable system. Some components are executable in the strict sense: code, workflows, proofs, simulations, statistical models, instrument procedures, and validation rules. Others are machine-readable commitments: the objective of the inquiry, the type of claim being made, the evidence required, the scope of a reviewer's judgment, the authority that may release a result, and the conditions under which a branch must stop. Still others remain human acts of interpretation, value judgment, and responsibility. Executability is therefore layered. It is the discipline of making each layer explicit enough that its powers and limits are visible.

The likely successor to the paper-centric model is an evidence-bearing research object. This object has a stable identity and versioned history. It contains a graph of claims and a graph of evidence, linked to data, code, protocols, environments, instrument records, models, reviews, and known failures. It can generate several views. The article is the public argument. A reviewer receives a packet focused on the claims requiring judgment. A replicator receives executable environments and divergence reports. A regulator receives a dossier of authority, risk, and provenance. A future research agent receives a compact record of what was tried, why it failed, and which uncertainties remain open. None of these views is the whole object, and no view should silently overwrite the others.

This transformation would alter the economics of science. Today, the visible product is usually the positive paper, while the expensive negative result, the careful replication, the maintained benchmark, the repaired dataset, and the decisive causal ablation often receive less credit. A versioned research object can make these contributions attributable. It can show that a branch was closed before a costly experiment, that a reviewer identified a common-mode failure, or that an engineer made a result reproducible across environments. The infrastructure would not automatically create fair incentives, but it would make invisible epistemic labour more legible.

The transformation would also create dangers. A complete research log can become a surveillance system. Early speculation, private disagreement, failed attempts, and provisional ideas are essential to inquiry. If every unformed thought becomes permanent evidence in an employment or funding decision, researchers will optimize for defensibility rather than discovery. Executable science therefore requires layered transparency. The public deserves evidence for consequential claims, but the researcher requires a protected space in which to be wrong. What must be preserved is not an indiscriminate transcript of every interaction. It is the history of reasons that changed the public state of the inquiry.

The message is not that science will become software. Science is answerable to mathematical structure, empirical resistance, causal identification, human meaning, and public consequence in ways that software is not. The message is that science will become more software-like in its operational form. Its results will increasingly live in versioned systems rather than isolated documents. Its claims will be connected to executable tests and provenance. Its revisions will be localized and reconstructable. Its agents will work inside governed runtimes rather than unbounded conversations. Its institutions will have to decide which parts of judgment can be automated, which must remain human, and how disagreement can remain visible when machines produce persuasive consensus at scale.

The decisive question is therefore not whether a model can imitate a scientist. It is whether scientific institutions can become computationally articulate without becoming mechanically minded. A laboratory that can generate indefinitely but cannot explain why a claim changed has automated activity, not inquiry. A repository that preserves every token but not the reasons for a decision has archived noise, not knowledge. An agent that passes internal tests but cannot expose its assumptions has produced performance, not authority. Executable science begins when the system is designed to remember these distinctions.

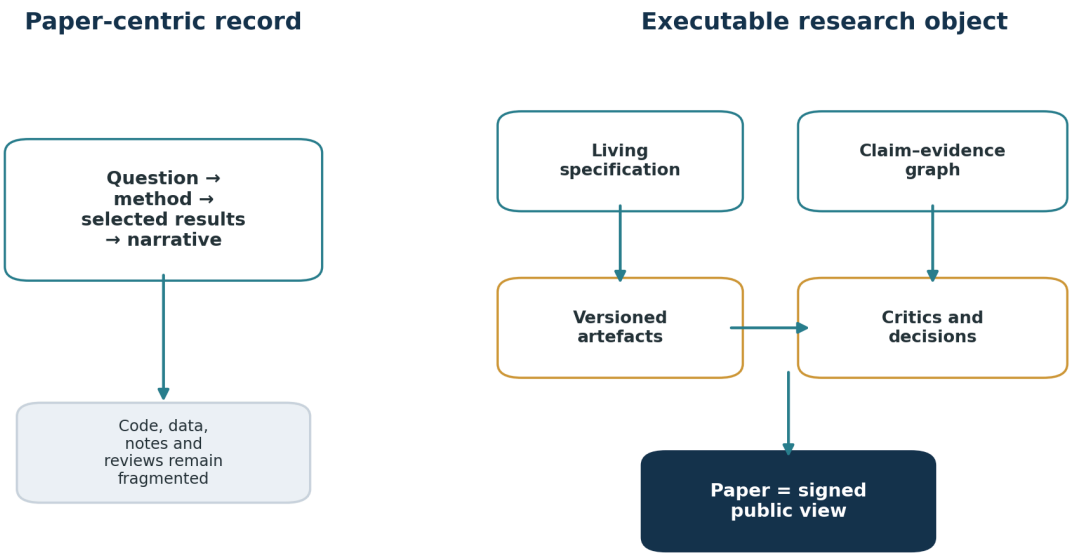


Figure 7. From a paper-centric record to an evidence-bearing research object. The paper remains a signed public argument, while the durable system of record preserves specification, artefacts, evidence, criticism, and change.

6.2 Science Becomes Software-Like, but Not Pure Software

The claim that science will become more like software can be misunderstood in two opposite ways. The first misunderstanding is reductionist: if a research process can be represented computationally, then its conclusions are no more than program outputs. The second is romantic: because scientific judgment exceeds computation, software concepts have little relevance to inquiry. Both positions miss the architectural change. Science is not reducible to software, but scientific work is increasingly mediated by

systems whose reliability depends on software-like disciplines of modularity, testing, versioning, dependency management, and controlled release.

Software engineering learned these disciplines because complex systems cannot be maintained as final products alone. A compiled binary does not reveal the requirements, source changes, tests, reviews, dependencies, and incidents that justify its use. The operational object is the repository and its governed history. Version control permits parallel exploration, merge review, reversion, and attribution. Continuous integration reruns tests when relevant components change. Package manifests identify dependencies. Issues and design records preserve unresolved obligations. Releases freeze selected states for external use. None of these mechanisms guarantees that the software serves a good purpose or contains no defect. They make change tractable and failure diagnosable.

Research faces an analogous problem, with a richer object of control. A paper is comparable to a release note and a signed design argument, not to the repository itself. The scientific repository must include datasets, code, environments, protocols, calibration records, preregistrations, models, analysis plans, notebooks, figures, review interventions, and the claims these artefacts are intended to support. More importantly, it must preserve relations that ordinary file versioning does not understand. Which result bears on which claim? Which assumption is required by which inference? Which negative control caused a mechanism claim to be narrowed? Which reviewer certified a formal property but not its empirical relevance? Which policy authorized access to sensitive data? Which uncertainty remains unresolved at release?

The appropriate unit is an evidence-bearing research object. It is “evidence-bearing” because every promoted claim carries links to the artefacts, checks, criticisms, and authorities that justify its current status. It is a “research object” because it contains a structured inquiry rather than a finished narrative. The object has a stable identity, versioned states, and an appendable history. It can branch when alternative hypotheses or analyses must coexist. It can merge when independently developed evidence becomes compatible. It can mark a claim as superseded without deleting the reason it was once accepted.

Existing infrastructure provides important foundations. Reproducible research links analysis to code and data rather than presenting results as prose alone (Peng, 2011). Jupyter notebooks combine narrative and executable computation, although hidden state and environment drift can limit reliability (Kluyver et al., 2016). Git can improve transparency and reproducibility by preserving changes and collaboration history, but it versions files rather than epistemic relations (Ram, 2013). W3C PROV represents entities, activities, and agents; RO-Crate packages research data, software, workflows, and contextual metadata; FAIR principles encourage findability, accessibility, interoperability, and reuse (Lebo et al., 2013; Wilkinson et al., 2016; Soiland-Reyes et al., 2022). Workflow standards such as the Common Workflow Language can describe executable computational pipelines. None of these should be discarded. The missing layer is a common profile that connects them to claim types, evidential stages, critic verdicts, unresolved obligations, and decision rights.

The word “executable” should therefore be applied with precision. A computational experiment may be executable in the literal sense: an environment can be reconstructed and a workflow rerun. A wet-laboratory protocol may be machine-executable only where instruments and materials are digitally controlled;

elsewhere, it may be operationally specified but require human execution. A statistical claim can be re-estimated from data, yet its identification assumptions cannot be established by execution alone. An interpretive claim may have no deterministic evaluator, but its evidence, quotations, coding decisions, and argumentative dependencies can still be represented and contested. Executability is a gradient of operational explicitness, not a binary property.

The analogy with software is strongest at the level of maintenance. Science changes. New data arrive, a library is updated, an instrument calibration is corrected, a confound is discovered, or an external team fails to reproduce a result. Under the paper-centric model, the response may be a correction, a new paper, or an informal change in community belief. The relation between the old and new states is often difficult to reconstruct. Under a versioned research-object model, a change invalidates the downstream claims and artefacts that depend on it. The workbench can identify the affected subgraph, recompute what remains executable, request human review where interpretation changes, and release a new view with a clear divergence report.

This form of selective recomputation is one of the most practical advantages of the architecture. Suppose a preprocessing error affects one dataset but not the theoretical analysis or an independent experiment. A monolithic workflow may rerun everything or, worse, regenerate a manuscript from scratch and introduce new inconsistencies. A dependency-aware workbench can invalidate only the affected branches. Validated outputs outside the dependency cone remain stable. The system becomes both more efficient and more auditable because correction is localized.

Software-like science also requires tests, but scientific tests are heterogeneous. Unit tests can verify functions, schema checks can enforce data structure, and theorem provers can establish formal properties. Statistical checks can test model assumptions or sensitivity. Laboratory controls can reveal contamination or instrument drift. Peer review can identify conceptual omissions. Replication can expose dependence on local conditions. The workbench must not collapse these into one green status indicator. A build can succeed while the scientific interpretation fails. An executable notebook can reproduce a biased estimator perfectly. A proof can be valid for a formal statement that does not represent the intended phenomenon. The system must show which dimensions passed and which remain open.

The deepest difference from software lies in relation to the world. Software correctness is ordinarily judged against a specification, even when the specification is incomplete or contested. Scientific inquiry often seeks the specification itself. The relevant variables, categories, mechanisms, and evaluators may be unknown at the beginning. This is why executable science cannot be a rigid compilation pipeline from complete requirements to final truth. It needs living specifications that can change under evidence, branches that preserve competing frames, and explicit records of semantic revision. The repository is not only a place where answers are built. It is where the language of the question is revised.

Table 11. The executable research object

Component	Operational role
Research specification	Records objectives, admissible claims, assumptions, invariants, methods, evidence thresholds, budgets, stopping rules, risks, and decision rights.

Component	Operational role
Claim and dependency graph	Connects each assertion to the data, code, models, protocols, transformations, and prior claims on which it depends.
Execution graph	Represents pipelines, agentic branches, tool calls, environments, validators, and promoted artefacts as rerunnable work rather than narrative reconstruction.
Evidence and criticism layer	Attaches supporting results, null findings, counterexamples, reviews, disagreements, uncertainty, and the status of unresolved objections.
Versioned rationale	Preserves consequential changes in interpretation, evaluator choice, scope, and release status without requiring indiscriminate retention of every private thought.
Public views	Produces papers, reports, dashboards, datasets, and interfaces as signed projections of a larger governed object.

6.3 Version Control for Claims, Evidence, and Reasons

A scientific version-control system must extend beyond file history. It needs a model of what changes mean. The core entities are claims, evidence, artefacts, specifications, decisions, and authorities. Each has an identity and a lifecycle. Their relations form a graph rather than a folder tree.

A claim is an assertion whose type determines its obligations. It may be a formal identity, an empirical regularity, a causal mechanism, a comparative advantage, a scaling law, a robustness statement, a novelty claim, a safety claim, or an interpretive conclusion. The type matters because evidence does not transfer automatically. A repeated correlation can support an empirical regularity without establishing mechanism. A benchmark improvement can support comparative performance under one budget without establishing general superiority. A formal proof can support an identity without showing that the formalization captures the intended system.

Evidence is not a generic confidence score. It is a typed relation between an artefact or observation and a claim. An executed test bears on implementation fidelity. A proof object bears on formal validity. A randomized experiment may bear on a causal effect under assumptions. A negative control may weaken a mechanism. An external replication bears on portability. A specialist review bears on a bounded judgment whose scope must be recorded. The same evidence may support one claim and undermine another. A result showing that a system works can simultaneously weaken the proposed explanation if the effect survives removal of the supposedly essential component.

Artefacts are the material objects through which research operates: datasets, code, environments, protocols, models, prompts, instrument outputs, notebooks, figures, annotations, and reports. They require content identity, versioning, and provenance. A result must be linked to the exact code and data states that produced it, not merely to a repository name. Dynamic dependencies such as external APIs, model versions, retrieval indexes, and policy configurations must be captured where they can affect the outcome. In AI-assisted research, prompts and model settings may matter, but preserving them is insufficient. The system must also record tool outputs, selected sources, execution environments, and the policy that determined which intermediate outputs entered persistent memory.

Specifications state what the inquiry is trying to accomplish and how it is governed. They contain objectives, constraints, evidence thresholds, budgets, stop rules, and decision rights. A claim may be syntactically valid yet out of scope because it no longer answers the specification. A branch may produce an interesting result but require a new research-object version if the objective changes. This distinction prevents semantic drift from being disguised as progress.

Decisions connect evidence to status. A claim may be proposed, specified, implemented, reproduced, challenged, narrowed, promoted, held, rejected, superseded, or released. Status changes should not occur automatically from a model's confidence. They require a rule and, at higher levels, an authority. The decision record should state which evidence was considered, which objections remain, and why the transition was authorized. This creates a history of reasons.

The history of reasons differs from a full conversational log. Conversations contain repetition, private speculation, irrelevant context, and sensitive information. They may also contain hidden chain-of-thought material that is neither reliable evidence nor appropriate for disclosure. What matters for scientific continuity is the set of consequential transitions. A claim changed because a negative control succeeded. A branch stopped because the expected information gain no longer justified the cost. A dataset was excluded because consent did not cover the proposed use. A reviewer changed a verdict after an independent reproduction. An institution delayed release because the evidence threshold for the intended consequence had not been reached.

This history enables scientific operations analogous to diff, blame, merge, and revert, but with more careful semantics. An epistemic diff shows which claims, assumptions, and evidence relations changed between versions. Attribution identifies the people and systems responsible for artefacts and judgments without pretending that authorship is a single scalar. A merge combines compatible branches while preserving unresolved conflict. A revert restores a previous operational state but does not erase the fact that later evidence existed. A fork allows a community to pursue a different interpretation or policy without corrupting the history of the original object.

Conflict must be first-class. Conventional documents often resolve disagreement by producing one accepted text. A research object should allow multiple live interpretations of the same evidence. Two teams may agree on data and analysis but disagree about mechanism or significance. A regulator may accept a result for exploratory use while a clinical body rejects deployment. A statistician may judge the estimator valid under assumptions that a domain expert considers implausible. The graph should preserve these views and the authorities that endorse them. Consensus becomes an outcome, not an enforced schema.

Version control also supports negative results. Today, failed experiments and abandoned branches are often inaccessible, making future researchers repeat avoidable work and training AI systems on compressed success narratives. In a workbench, a negative result can close a branch, constrain a hypothesis, or update the expected value of a method. It should be preserved with enough context to be reusable: the specification it tested, the environment, the evidence, the reason for stopping, and the boundary beyond which the result should not be generalized. Negative evidence is not merely archival. It is executable memory.

Provenance must be proportional to consequence. Recording every intermediate token and transient state is expensive, invasive, and often useless. A low-risk

exploratory calculation may require only code, data, environment, and a brief rationale. A clinical or safety-critical result may require instrument calibration, data lineage, model and policy versions, access authority, independent witness, and release approval. The research object should declare a provenance profile appropriate to the domain and intended use.

Cryptographic commitments and transparency services can strengthen integrity and ordering, but they do not prove scientific truth. A hash can show that an artefact changed. A signature can show that a key authorized a statement. An append-only log can make retrospective rewriting detectable. None establishes that an instrument was calibrated, a participant consented freely, a model was unbiased, or a claim is correct. The infrastructure should decompose trust into inspectable evidence rather than replace institutional trust with technological rhetoric.

Long-lived verifiability introduces maintenance obligations. Code environments decay, container registries disappear, model APIs change, cryptographic algorithms age, and external data sources are revised. A research object must preserve not only content but proof liveness: enough information and executable infrastructure to interpret historical evidence later. Migration should itself produce evidence linking old and new states. Otherwise, the repository may contain immutable commitments to artefacts that no one can reconstruct.

A useful way to make that semantic layer concrete is to define the scientific equivalent of a commit. A research commit should not mean merely that files were saved. It should identify the prior accepted state, the changed artefacts, the affected claims, the specification clause that motivated the change, the validators that ran, the evidence produced, and the authority under which the transition was admitted. The commit message becomes a compact rationale, while the graph contains the inspectable support. Routine low-risk changes may be admitted automatically when deterministic checks pass. Changes that narrow a central claim, alter an evaluator, replace a dataset, or raise the consequence of release should require explicit review. In this model, continuous integration becomes continuous epistemic integration: every material change triggers the smallest competent set of tests, reproductions, provenance checks, and policy gates.

Scientific branching also needs stronger semantics than software feature branches. A branch may encode a rival mechanism, an alternative causal model, a different preprocessing decision, an adversarial interpretation, or a distinct ethical and regulatory position. Its purpose should be declared before results are known, otherwise branching becomes a retrospective device for hiding researcher degrees of freedom. A merge should occur only when the branches are compatible at the level of assumptions and evidence. When they are not, the repository should preserve a structured conflict. A conflict is not a technical failure to be eliminated by fluent synthesis. It may represent the most informative state of the inquiry: two interpretations that share observations but differ in causal commitments, or two release policies that assign different weight to risk.

Tags and releases acquire epistemic meaning as well. A tag may identify a preregistered analysis state, a frozen dataset, an externally reproduced result, a regulatory submission, or the exact object from which a paper was rendered. A release is therefore not only a bundle of artefacts. It is a policy assertion that a defined set of claims has reached a declared evidence level for a declared audience and use. Later evidence may supersede that release without erasing it. This matters because scientific

correction is not equivalent to rewriting a web page. People and institutions may have acted on an earlier state. A responsible system must preserve which evidence and authority were available at the time.

The issue tracker also changes role. Open questions, suspected confounds, missing controls, replication failures, reviewer objections, and safety concerns become addressable objects linked to the relevant claims and versions. Closing an issue requires evidence or a reasoned decision, not merely a conversational assurance. Some issues remain intentionally open at release because uncertainty is irreducible or because the required experiment is not yet feasible. Their visibility prevents a polished manuscript from converting absence of resolution into appearance of consensus.

Such repositories would support new forms of scientific review. A reviewer could request a semantic diff between the submitted claim and its preregistered ancestor, inspect whether a positive result survived evaluator changes, or run the dependency slice affected by a disputed assumption. A replication team could fork the object while retaining ancestry, attach independent evidence, and propose a merge or a persistent divergence. Meta-researchers could study how claims mature, which kinds of criticism cause revision, and where evidence repeatedly fails to propagate into public conclusions. These capabilities are plausible with extensions of existing version control, provenance models, research-object packaging, workflow systems, and policy engines. The difficult work is not storing more files. It is agreeing on the minimal identities, relations, and transition rules that make the stored history scientifically meaningful.

The Git analogy is therefore powerful but incomplete. Git answers whether bytes changed and how branches relate. Executable science must also answer whether the meaning of a claim changed, whether the evidence remains valid, whether the authority is current, and whether the object can still be interpreted. The future scientific repository will need a semantic layer above ordinary version control, not a replacement for version control.

6.4 The Paper as a Signed View

If the research object becomes the durable system of record, the scientific paper does not disappear. Its role becomes clearer. The paper is a selected, signed, public argument. It explains why a question matters, which evidence is decisive, how alternative interpretations were treated, and what the authors are prepared to defend. It is a view over the research object, not an automatic dump of the object's contents.

This distinction could improve scientific writing. Current papers often combine narrative, implementation detail, reporting compliance, and repository pointers in a format that satisfies none of these purposes completely. A view-based architecture allows the public article to remain conceptually coherent while the underlying object provides executable depth. The paper can state the central claims and the evidence that changed the authors' position. A reviewer can navigate from each claim to methods, code, data, negative controls, and prior criticism. A replicator can obtain a frozen environment. A policymaker can inspect consequence-specific evidence and dissent. The manuscript can become shorter without becoming less transparent because compression is no longer equivalent to deletion.

The signature is important. Authorship in this model is not a claim that the authors manually produced every sentence, line of code, or experiment. It is an acceptance of responsibility for the public argument and its release state. The author or

institution declares that the specification is represented honestly, the cited evidence exists, known material limitations are disclosed, and the claimed scope is justified. AI use should be reported at the level necessary to assess provenance and common-mode risk, but disclosure should not substitute for responsibility. “The model produced it” is not an authorship category and cannot function as a liability shield.

The research object can support richer credit than a conventional author list. Contribution records can identify hypothesis design, software, data curation, instrument operation, formal verification, causal analysis, replication, review, governance, and maintenance. Machine contributions can be recorded as tools and activities rather than fictional persons. Human contributors can receive credit for closing expensive branches, constructing decisive controls, maintaining benchmarks, or curating evidence even when those actions do not produce visible prose. This can align recognition more closely with the work that makes knowledge cumulative.

Transparency, however, must be layered. A naïve response to automated science is to demand complete logging. Such a regime would be both technically excessive and intellectually harmful. Research requires a protected interval of uncertainty. Scientists need to explore interpretations that may be wrong, discuss sensitive possibilities, and revise ideas without every provisional statement becoming a permanent performance record. Full logging can expose personal data, confidential collaborations, unpublished intellectual property, security-sensitive methods, and the private context of reviewers or participants.

A defensible architecture separates at least four visibility regimes. Private exploration allows uncommitted work and ephemeral interaction. Collaborative state contains artefacts and decisions shared within the research team. Institutional audit preserves evidence needed for integrity, safety, and accountability under controlled access. Public release exposes the claims, supporting evidence, provenance, and unresolved dissent appropriate to external reliance. The boundaries are not fixed universally; they depend on domain, law, consent, and consequence. What matters is that the transition between regimes is explicit and authorized.

The right to be wrong before publication should be treated as an epistemic right. It is not a right to conceal fraud or suppress known harm. It is the protection of provisional thought from premature institutionalization. Without it, researchers and agents will optimize for records that look defensible, avoiding risky questions and candid disagreement. The workbench should preserve committed reasons, not every cognitive trace.

At the same time, public claims need stronger traceability than ordinary papers provide. A released view should identify the exact research-object version it represents. If data, code, or interpretation later change, the public record should show whether the article remains valid, has been superseded, or now carries a warning. Retraction should not make history disappear. Correction should be an explicit successor state linked to the prior one. External replication should attach as new evidence rather than remain a disconnected citation.

This model also changes peer review. Review becomes an intervention on the research object, not merely comments on a PDF. A reviewer can challenge a specific claim, request a branch, attach an executable test, identify a missing source, or restrict the scope of a verdict. The authors’ response can link to the resulting artefact and show how claim status changed. Post-publication criticism can continue in the same structure,

allowing the object to accumulate evidence rather than forcing every dispute into a new article.

There is a danger that such infrastructure becomes bureaucratic. If every exploratory project requires a large ontology, complex access policy, and formal approval for minor changes, the workbench will be rejected or will encourage performative compliance. The minimum viable research object should be small. It needs stable identity, versioned specification, claims, evidence links, provenance sufficient for consequence, and a record of decisions. Additional structure should be introduced only where it improves control, reproducibility, or judgment. The architecture must support progressive formalization: ordinary prose and Markdown at the surface, typed representations where useful, formal proofs or deterministic validators where available.

This progressive approach is essential for adoption. Scientists already use repositories, notebooks, workflow engines, electronic laboratory notebooks, preregistrations, and data packages. The transition to executable science should connect these tools rather than demand one universal platform. Interoperable profiles and open formats matter more than a proprietary “AI scientist” interface. A research object should survive changes in model vendor, orchestration framework, storage service, and institutional policy.

The likely end state is not a single global GitHub for science. Scientific domains differ in privacy, data scale, instrument control, intellectual property, and governance. The more plausible architecture is federated. Research objects retain stable identities and portable evidence profiles while institutions operate their own storage, access, and validation services. Public repositories expose release views and reproducibility packages. Sensitive domains provide verifiable summaries, controlled execution, or independent attestations without publishing raw data. Interoperability occurs at the level of claims, provenance, evidence, and version relations.

The normative conclusion is that the scientific article should remain the place where humans make an intelligible public argument. It should no longer be forced to act as the entire memory of the inquiry. The repository preserves execution and change; the paper preserves significance and responsibility. Their separation is not a demotion of writing. It is a recognition that, in an age of generative abundance, rhetoric must be connected to a larger object capable of resisting it.

PART III: THE EPISTEMOLOGY OF DISCOVERY

The architecture can be credible only after its limits are explicit: finite evidence, unknown distributions, causal ambiguity, revision, evaluator failure, and open-ended search.

Chapter 7. Finite Evidence and Structural Limits

Some boundaries yield to better engineering; others follow from missing information, non-identifiability, undecidability, or a weak evaluator.

7.1 Finite Evidence and the Price of Generalization

Machine learning begins with an asymmetry: the evidence is finite, but the claims usually concern cases not yet observed. A training set constrains a model, yet in any sufficiently expressive hypothesis class many rules can agree on every training example and disagree elsewhere. Generalization is therefore never obtained from data alone. It depends on inductive bias, the collection of preferences and structural assumptions that make some continuations of the evidence more likely than others. Architecture, optimization, regularization, data augmentation, pretraining and the choice of representation all contribute to that bias.

The need for bias is not a defect introduced by neural networks. It is a condition of induction. Tom Mitchell's early formulation made the point directly: a learner that generalizes must prefer some hypotheses over others. No Free Lunch theorems for optimization make a related claim. Averaged uniformly across all possible objective functions, no search method has a universal advantage (Wolpert and Macready, 1997). Practical success is possible because real problems are not uniformly arbitrary. Images, language, physical systems and scientific data contain regularities, symmetries, locality and compositional structure. The open question is which biases capture those structures without becoming brittle when the world changes.

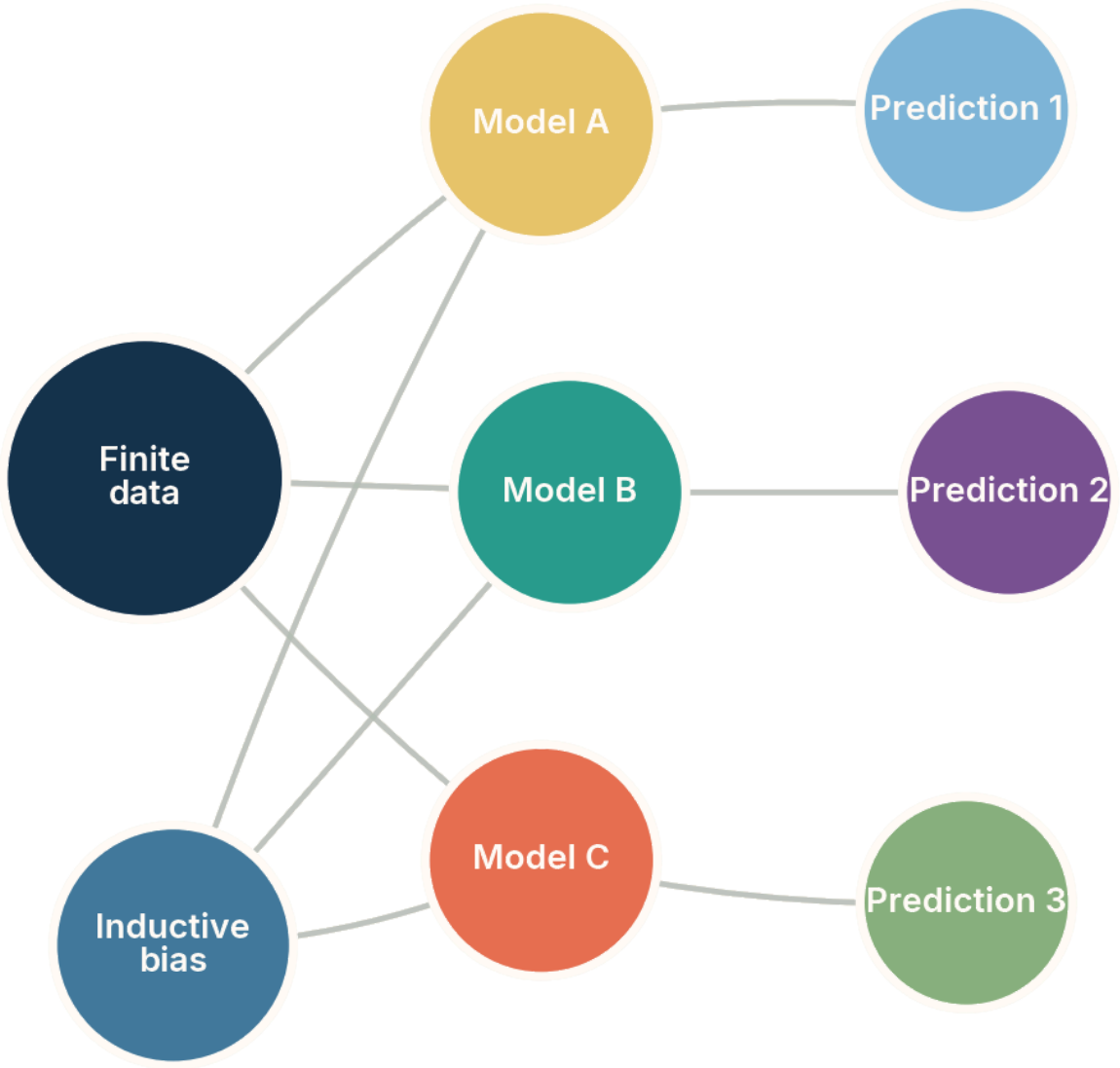


Figure 8. The same finite data can support several models that make incompatible predictions beyond the observed region. Conceptual diagram; not an empirical result.

The figure matters because it visualizes underdetermination without treating it as a rare pathology. Model A, Model B and Model C can fit the same evidence. Their divergence appears only on a new case. The inductive bias node represents the often-hidden reason that one model is selected: perhaps it is simpler, smoother, more causal, more compatible with a pretrained representation or easier to optimize. When a system presents a prediction without exposing these dependencies, confidence can be mistaken for logical necessity.

Scaling changes this picture but does not abolish it. Larger models trained on broader data can acquire useful priors that resolve many ambiguities encountered in ordinary tasks. They may interpolate across regions that were previously sparse and transfer regularities between domains. Yet broader priors can also make assumptions harder to see. A foundation model carries the statistical history of its corpus, annotation practices, tool use and optimization objective. The model may generalize impressively because the new task resembles structures already represented, while failing abruptly where those structures cease to apply.

A useful distinction is between interpolation in a rich representation and extrapolation under a changed mechanism. What looks like radical novelty to a human observer may be close to known examples in the model's latent space. Conversely, a visually minor change may alter the causal process that generated the data. Generalization claims should therefore specify the axis along which novelty occurs. New combinations, new parameter ranges, new causal regimes, new sensors and new goals pose different challenges. A single test score cannot summarize them.

Table 12. Common inductive biases and the worlds in which they help

Bias or assumption	When it is useful and when it can fail
Smoothness	Useful when nearby inputs have nearby outputs; brittle near discontinuities and regime changes.
Locality and translation symmetry	Powerful for spatial data; misleading when global context or absolute position is decisive.
Simplicity or compression	Favors reusable explanations; can prefer elegant but false models when reality is heterogeneous.
Compositionality	Supports recombination of known parts; fails when interactions create qualitatively new behaviour.
Causal invariance	Supports transfer across environments when mechanisms remain stable; fails when the mechanism itself changes.
Pretrained semantic structure	Provides broad priors; can import historical bias, omissions and spurious associations.

The table exists to make explicit that every route to generalization is conditional. A bias is neither simply good nor bad. It is a bet on the structure of the world. Machine-learning research often compares methods on average performance while leaving the bets implicit. A more scientific evaluation identifies which structural assumptions each method embodies and deliberately tests environments where those assumptions hold, weaken or fail. This converts surprising failure from an embarrassment into evidence about the learner's effective theory.

The frontier lies in learning biases that are both powerful and revisable. Meta-learning attempts to infer learning strategies from families of tasks. Causal representation learning searches for variables and mechanisms that remain stable across interventions. Equivariant architectures encode known symmetries, while neural program induction and modular systems aim to support compositional recombination. None eliminates induction. Their promise is to move assumptions from accidental correlations toward structures that are more stable, interpretable or testable.

Another frontier is selective prediction. When several hypotheses remain compatible with the evidence, a system should preserve the ambiguity rather than collapse it prematurely. Ensembles, Bayesian approximations, conformal methods and explicit version spaces offer different ways to represent alternatives. An active system can then select observations that separate them. This changes the objective from "guess the unseen label" to "identify which missing evidence would most reduce underdetermination." The latter is closer to the logic of scientific inquiry.

The central limit can therefore be stated positively. Finite data do not determine a unique future, but they can support a disciplined set of possibilities. Progress comes

from making the assumptions visible, testing them across environments, maintaining rival explanations and choosing experiments that force those explanations to disagree. Machine learning becomes a discovery method not when it escapes induction, but when it manages induction as an explicit, revisable commitment.

7.2 Identifiability, Observability, and Impossibility

Some learning problems are difficult because available algorithms are immature. Others are impossible under the information provided. The concept that separates these cases is identifiability. A target is identifiable when distinct underlying explanations produce distinguishable observations under the stated assumptions. If two possible worlds generate exactly the same accessible data but require different answers, no learner can always choose correctly. More computation cannot recover a distinction that leaves no trace in the input.

Identifiability appears throughout machine learning. A mixture model may admit multiple parameterizations with the same likelihood. Latent factors can be rotated or permuted without changing observations. Causal graphs can be observationally equivalent. A classifier cannot distinguish categories that are identical in the measured features. Unsupervised disentanglement is not identifiable without assumptions that connect latent variables to the data-generating process (Locatello et al., 2019). These are not signs that models are insufficiently intelligent. They are statements about information.

Observability is the engineering counterpart. A relevant property may exist but remain inaccessible to the sensors, sampling procedure or record. A medical model cannot infer a physiological process that affects neither the measurements nor correlated variables. A fraud detector cannot reliably identify a novel strategy engineered to mimic all observed statistics of legitimate behaviour. A scientific agent cannot establish historical priority from a corpus that omits unpublished, inaccessible or differently described work. In each case, the missing channel matters more than model size.

Table 13. Four classes of limits that are often mistaken for one another

Class of limit	Diagnostic question
Information limit	Does the available evidence contain any signal that distinguishes the alternatives?
Identification limit	Given the assumptions, does one underlying explanation follow uniquely from the evidence?
Computational limit	Is the answer defined but intractable to obtain with feasible resources?
Evaluation limit	Can correctness or progress be measured without relying on the same fallible system that generated the claim?

The table exists because the remedy depends on the diagnosis. An information limit requires new sensors, interventions, records or labels. An identification limit requires stronger assumptions or a redesigned experiment. A computational limit may be attacked with approximation, decomposition, hardware or a restricted problem class. An evaluation limit requires a stronger verifier or a different institutional process. Treating all four as “model capability” encourages wasteful scaling and obscures the actual bottleneck.

Computability adds a separate boundary. Some formally defined problems are undecidable in general; others are decidable but require resources that grow too rapidly. Machine learning can provide useful heuristics on distributions of practical instances, but it does not repeal these results. A theorem-proving model may solve many statements while no general procedure can decide every mathematical proposition in a sufficiently expressive formal system. A program-analysis agent may detect numerous bugs while the halting problem remains undecidable. The practical frontier is distributional competence under explicit coverage, not universal guarantees.

Logical limits also constrain self-verification. A powerful system cannot in general establish every true statement about its own behaviour from within a fixed formal framework. In practice the more immediate issue is mundane: a model used both to generate and judge an answer shares blind spots with itself. Independent tools, formal checkers, alternative models, human review and physical replication reduce correlated error. They do not produce absolute certainty, but they create epistemic separation between proposal and acceptance.

The distinction between structural and movable limits should guide research investment. Structural limits include absent information, observational equivalence under the allowed data, undecidability and the impossibility of certifying global historical novelty from an incomplete record. Movable boundaries include poor sensor design, weak representations, inefficient search, limited calibration, brittle adaptation and inadequate evaluation. The latter can advance dramatically. The former can often be circumvented by changing the problem, adding interventions or narrowing the claim, but not by pretending the constraint does not exist.

When two worlds W_1 and W_2 induce the same distribution over every observation available to the learner, yet require different answers to a target question, no decision rule based only on those observations can be correct in both worlds.

This statement is a compact test for overclaiming. It asks us to construct rival worlds, not merely rival models. If the evidence would look the same in both, then confidence must come from an assumption that rules one world out. The assumption may be reasonable, such as temporal order, sparsity, smoothness or invariance, but it should be named and tested. Scientific progress often consists of designing an intervention that makes the worlds observationally different.

The constructive consequence is an architecture of epistemic escalation. A system begins with passive observations and a set of compatible explanations. It searches for missing variables, proposes measurements or interventions, estimates the cost and risk of obtaining them, and updates the set of explanations after evidence arrives. In this architecture, admitting non-identifiability is not failure. It is the signal that triggers better experimental design.

7.3 Limits That Scaling Cannot Erase

Some failures diminish with better models, more data and more computation. Others follow from the structure of the problem. A comprehensive frontier map must separate them. Otherwise every impossibility is dismissed as temporary and every engineering weakness is treated as metaphysical. Structural limits do not imply that useful work is impossible. They identify where the claim, data or interaction must change.

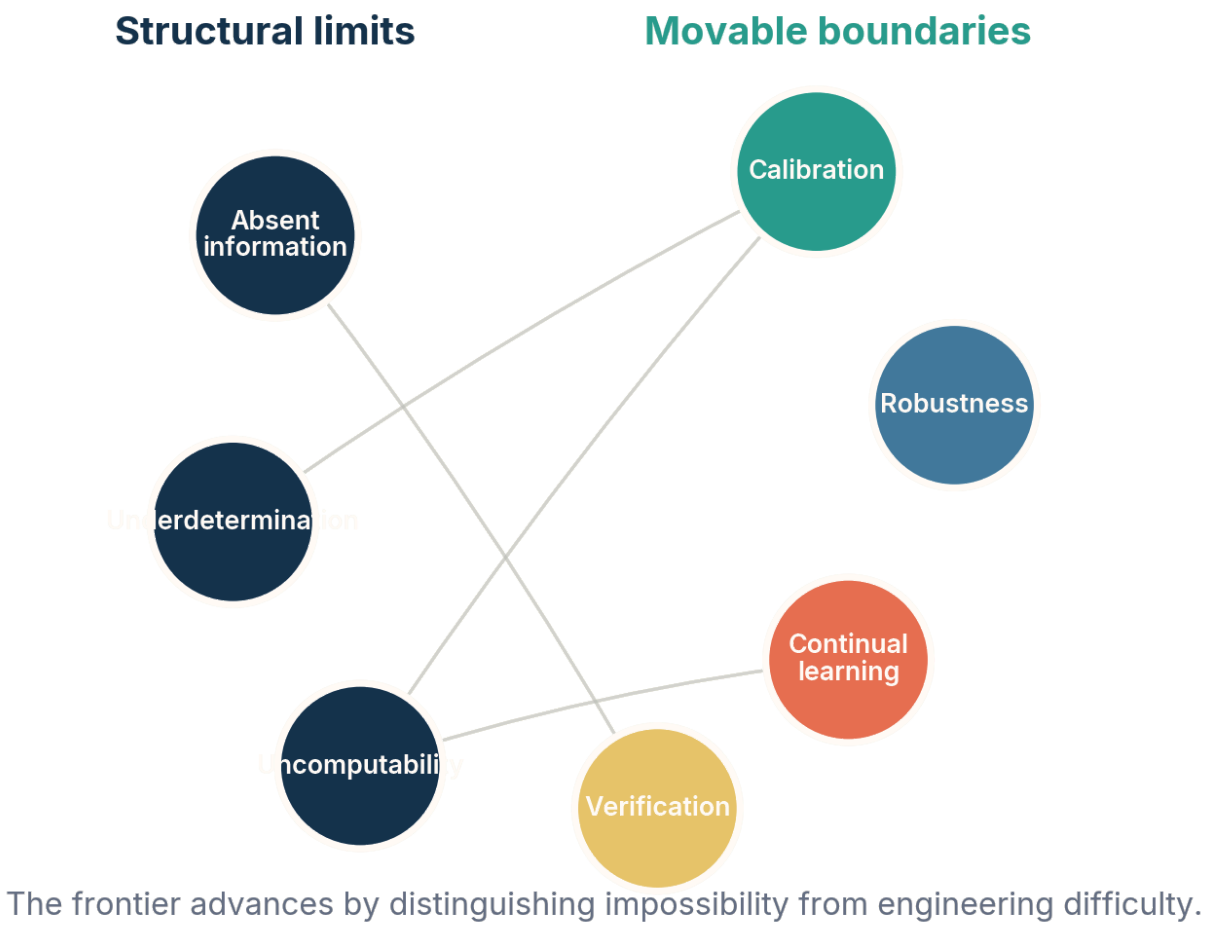


Figure 9. Structural limits and movable boundaries interact, but they require different research responses. Conceptual diagram; not an empirical result.

The figure matters because structural and movable constraints are not isolated. Better calibration can reveal when absent information matters, but it cannot create the information. Verification can certify a formal claim, but not global historical priority. Continual learning can update a model, but it cannot guarantee that a finite record determines the correct future. Progress comes from moving the engineering boundary while respecting the remaining structure.

The first structural limit is absent information. A system cannot infer a distinction that leaves no observable trace in its data, sensors, tools or interactions. It can use priors to guess, but a guarantee would require evidence. The remedy is to obtain a new measurement, intervention or source. This principle applies equally to hidden biological variables, private historical records and adversarial strategies designed to imitate normal data.

The second limit is underdetermination. Finite evidence can support several models that diverge outside the observed region. More data can reduce the set, but no finite dataset eliminates all logically possible continuations without assumptions. Inductive bias is unavoidable. The scientific response is to expose assumptions, preserve alternatives and seek observations that make them disagree.

The third limit is identification. Observationally equivalent causal structures, non-identifiable latent variables and duplicated representations cannot be separated under the current design. Additional assumptions or interventions can make a narrower

problem identifiable. Scaling a predictor on the same observations does not change the equivalence class.

Table 14. Structural limits and the appropriate response

Structural limit	What a responsible system should do
Absent information	State that the question is not answerable from current evidence and request a new channel.
Underdetermination	Maintain alternatives and disclose the inductive assumptions selecting among them.
Non-identifiability	Propose interventions or restrict the claim to identifiable quantities.
Undecidability or intractability	Offer bounded heuristics, certificates for solved instances and explicit coverage.
Incomplete prior-art record	Report a scoped search and probabilistic priority claim, not absolute novelty.
Specification gap	Treat success on the proxy as provisional and search for counterexamples or external validation.

The table exists to translate limits into behaviour. A system need not respond to impossibility with silence. It can narrow a question, identify missing evidence, return a set rather than a point, or provide a certificate for a solved instance. The irresponsible response is unjustified confidence. The productive response turns the limit into a plan for changing the information structure.

Computational limits include undecidable problems and intractable search. Learning can exploit distributions of practical instances and generate useful heuristics. It can also produce certificates, such as formal proofs, that are cheap to check even if hard to find. The claim must remain distributional or instance-specific. Strong empirical performance does not imply a universal decision procedure.

Historical novelty has a special limit. No accessible corpus contains every unpublished idea, language, patent, notebook and oral tradition. Semantic equivalence is itself imperfect. A system can provide a reproducible prior-art search and estimate that no close precedent was found. It cannot establish absolute firstness from an incomplete record. Human scholarship lives with the same limit; machines should not conceal it.

Specification gaps are also structural relative to an objective. An optimizer cannot infer an intention that is neither encoded nor observable in feedback. Inverse learning can infer preferences from behaviour, but behaviour underdetermines values and can be inconsistent. Dialogue, oversight and corrigibility can improve alignment, yet the system cannot guarantee fidelity to an intention that the institution itself has not resolved.

These limits define an ethics of claims. Models should distinguish prediction from identification, evidence from assumption, local novelty from historical priority, and heuristic success from guarantee. This does not diminish machine learning. It makes its achievements comparable and allows research to focus on boundaries that can genuinely move.

Chapter 8. Unknowns and Distribution Shift

Knowing that a model is outside its competence requires more than confidence: it requires explicit threat models, calibration, abstention, and living evaluation.

8.1 Anomaly, Novelty, Open Sets, and Out-of-Distribution Inputs

The phrase “detect the unknown” covers several technical tasks with different assumptions. Anomaly detection searches for rare or irregular observations, often within a nominally single distribution. Novelty detection learns a description of normal data and flags future departures. Open-set recognition asks a classifier to identify known classes while rejecting examples that belong to classes absent during training. Out-of-distribution (OOD) detection asks whether a test input comes from the training distribution or from a different one. The acronym OOD will be used only in this technical sense throughout the book.

These tasks overlap but are not interchangeable. A rare example can be in distribution. An example from a different distribution can look ordinary. A new class can be embedded inside the same region of feature space as known classes. An anomaly detector may be trained without labels, while open-set recognition depends on known class boundaries. The operational question, the reference distribution and the costs of false acceptance and false rejection must therefore be specified before a score has meaning.

Table 15. Related methods for handling unfamiliar inputs

Technical setting	Core question and hidden dependency
Anomaly detection	Is this observation atypical relative to a nominal population? It depends on how rarity is represented.
Novelty detection	Does a future observation depart from a learned description of normality? It depends on coverage of normal variation.
Open-set recognition	Does the input belong to a known class or should it be rejected? It depends on separability of unknown classes.
Out-of-distribution detection	Does the input arise from the reference distribution? It depends on which alternative distributions matter.
Change-point detection	Has the data-generating process changed over time? It depends on temporal assumptions and detection delay.
Selective prediction	Should the model answer or abstain? It depends on a utility or risk threshold, not novelty alone.

The table exists to stop methods from inheriting claims they were not designed to support. An anomaly score is not automatically a probability that an object belongs to an unknown semantic category. A distribution detector does not necessarily explain the nature of the shift. Selective prediction may abstain on difficult familiar cases and answer confidently on unfamiliar ones. Treating these methods as variants of a universal “unknown detector” hides the assumptions under which each can work.

A common design learns a score s(x) and rejects inputs whose score crosses a threshold. The score may use softmax confidence, energy, distance to class prototypes, density estimates, reconstruction error, gradients, ensembles or representations learned from auxiliary data. Evaluation often uses the area under the receiver operating

characteristic curve (AUROC), which measures ranking across thresholds, alongside threshold-specific error rates. These metrics are useful only relative to chosen in-distribution and out-of-distribution datasets. A detector can rank one family of shifts well and fail on another.

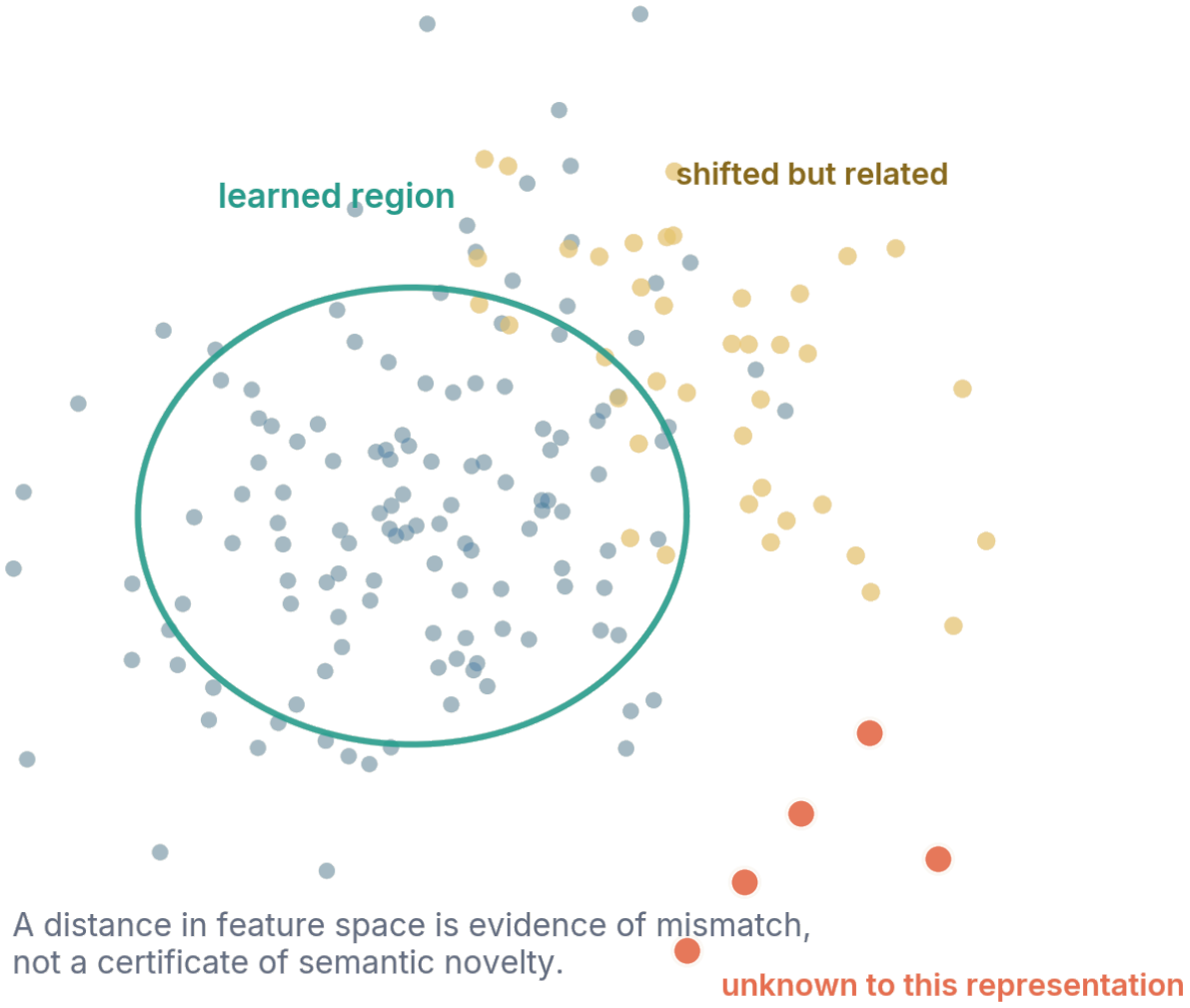


Figure 10. Out-of-distribution detection measures mismatch in a representation; it does not certify semantic or historical novelty. Conceptual diagram; not an empirical result.

The figure matters because the green boundary is not a border of the known world. It is a boundary induced by features, training data and a scoring rule. Yellow points may be shifted but still governed by familiar mechanisms. Red points may be sensor failures rather than discoveries. A genuinely new phenomenon can also appear inside the learned region if the representation discards the feature that makes it new. The drawing therefore turns the detector's output into the correct statement: "this observation mismatches my reference in this representation."

The theoretical limit is sharp. Fang and colleagues show that out-of-distribution detection is not universally learnable without restrictions on the distributions and hypothesis classes involved (Fang et al., 2024). Intuitively, if the unknown distribution can be arbitrary, it can imitate the known distribution wherever the learner observes data and differ elsewhere. No detector can guarantee success against every possible alternative. Practical methods work by exploiting structure: semantic separation, auxiliary outliers, pretrained features, density regularities or assumptions about how shifts occur.

This result does not make detection futile. It clarifies the research problem. A deployable system should define an “OOD threat model” analogous to a safety case: which changes are expected, which sensors may fail, which unknown classes are costly, how thresholds will be chosen, and what happens after rejection. Benchmarks should include near shifts that preserve superficial appearance, far shifts, semantic novelties, nuisance changes and adversarially selected cases. Performance should be reported separately rather than averaged into a reassuring number.

Promising frontier work combines representation learning with active escalation. A detector can group rejected cases, retrieve related evidence, request labels, trigger additional sensors or route a case to a specialist. Human feedback can improve coverage of operationally relevant unknowns, while online methods can update thresholds as environments change. The goal is not to label every surprise correctly at first contact. It is to keep mismatch from becoming silent confident error and to convert recurring unfamiliarity into a new, audited region of competence.

The correct ambition is thus bounded but powerful. Machine learning can learn the geometry of a known region, estimate where that geometry becomes unreliable and construct procedures for abstention and investigation. It cannot learn the shape of every possible unknown in advance. The research frontier lies in making the reference explicit, broadening the representations, learning from rejected cases and linking detection to evidence-gathering actions.

8.2 Uncertainty, Calibration, Abstention, and Conformal Prediction

A model can be wrong for many reasons, and its numerical confidence is not automatically a measure of those reasons. Predictive uncertainty may reflect ambiguity in the observed case, insufficient data about the model, distribution shift, label noise or unresolved alternatives. Deep classifiers are often miscalibrated: among predictions assigned a nominal probability of ninety per cent, the observed frequency of correctness need not be ninety per cent (Guo et al., 2017). Under dataset shift, even methods that appear calibrated on held-out data can deteriorate substantially (Ovadia et al., 2019).

Calibration is relational in the same way novelty is relational. A probability may be calibrated over a population and time period while being unreliable for a subgroup or a changed environment. Average calibration can coexist with dangerous local overconfidence. Language models add a further complication because verbal expressions such as “likely” or “I am uncertain” are generated tokens, not privileged access to an internal epistemic state. Research on semantic entropy shows that variation across meaning-equivalent generations can help identify some hallucinations, but it remains a task-specific signal rather than a universal introspective faculty (Farquhar et al., 2024).

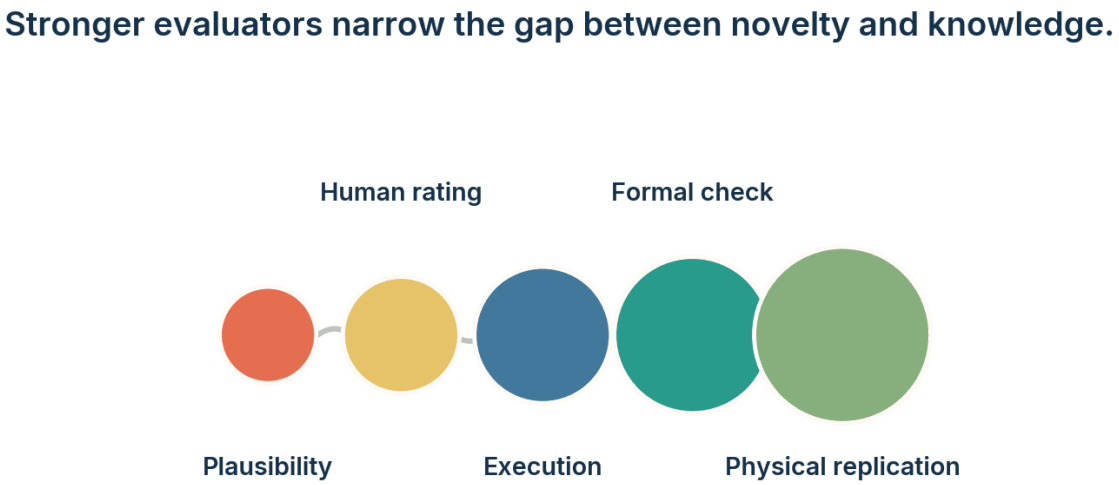


Figure 11. Confidence becomes more useful as it is coupled to increasingly independent forms of verification. Conceptual diagram; not an empirical result.

The figure matters because uncertainty estimates are not substitutes for verification. A plausibility score may help prioritize candidates, but execution can reveal errors that plausibility misses. Formal checking can certify properties within a specification. Physical replication can test whether the specification captures the world. Stronger evaluators do not merely increase confidence; they change the kind of claim that can be defended. The practical design principle is to route high-stakes uncertainty toward evaluators with greater independence from the generator.

Selective prediction gives the model a reject option. Instead of forcing an answer for every input, the system abstains when estimated risk exceeds a threshold. The threshold should be chosen by decision cost: a medical screening system, a content filter and a recommendation engine have different tolerances for false acceptance and false rejection. Abstention is meaningful only when a reliable fallback exists. A system that refuses difficult cases but sends them to an equally fallible model has rearranged uncertainty rather than managed it.

Conformal prediction offers a different approach. Under exchangeability assumptions, it converts model scores into prediction sets with finite-sample coverage guarantees. For classification, the output may be a set of labels rather than one label; for regression, an interval rather than a point. The guarantee is marginal over future examples drawn under the stated conditions, not a promise that every subgroup or individual interval is correct. Online conformal methods extend the idea to some forms of distribution shift, but they trade stronger assumptions, delayed feedback or adaptive coverage for the classical guarantee (Angelopoulos and Bates, 2023; Gibbs and Candès, 2024).

Table 16. What common uncertainty tools can and cannot guarantee

Method	Useful guarantee and principal limitation
Calibrated probabilities	Align stated confidence with empirical frequency on a reference population; calibration can break locally or after shift.
Deep ensembles	Capture disagreement among trained models; members may share data, architecture and blind spots.
Bayesian approximations	Represent parameter uncertainty under a model; results inherit the model class and prior assumptions.
Conformal prediction	Provides coverage under exchangeability or adapted online assumptions; sets may become wide or uninformative.
Abstention	Limits automated decisions when estimated risk is high; value depends on threshold design and fallback quality.
External verification	Tests a claim through execution, formal proof or measurement; may be costly, incomplete or based on a flawed specification.

The table exists to replace the question “which uncertainty method is best?” with “which uncertainty claim is required?” Methods describe different layers of ignorance. Ensembles estimate sensitivity to training variation, conformal methods control coverage, abstention manages decisions and external verification tests the object itself. Combining them is often more defensible than selecting one scalar score. A discovery system might use an ensemble to identify disputed hypotheses, conformal sets to preserve alternatives, and experiments to resolve them.

A frontier direction is uncertainty decomposition tied to action. Instead of one confidence number, a system can estimate whether uncertainty is likely to be reduced by retrieving literature, collecting another measurement, changing the model class or asking a human. This is a value-of-information problem. The output is not simply “I do not know,” but “the decision is sensitive to this missing variable, and this observation is expected to reduce that sensitivity.” Such systems would turn metacognition into an operational research policy.

Another direction is calibration under evolving conditions. Static test sets are insufficient when deployed systems change their users and data. Living calibration monitors residuals, subgroup performance, abstention rates and the composition of rejected cases. It can detect when confidence scores stop meaning what they meant at release. This requires provenance and delayed ground truth, but it is one of the most realistic ways to prevent uncertainty estimates from becoming ceremonial.

The broader lesson is that knowing what one does not know is not a single learned capability. It is an architecture that connects scores, assumptions, decision costs, fallback procedures and independent evidence. Machine learning can make major progress here, but no internal confidence signal can certify that all unknown unknowns have been captured. Reliable systems remain open to correction from outside themselves.

8.3 Distribution Shift, Shortcuts, and Underspecification

A model is usually trained and evaluated under an approximation that the future resembles the recorded past. Distribution shift names the many ways in which that approximation can fail. The frequency of inputs can change, the relation between inputs and labels can change, the population can change, the sensor can change, the decision

rule can alter behaviour, and the model's own deployment can reshape the environment. These are not cosmetic variants of one problem. Each shift changes which parts of the learned representation remain relevant.

Covariate shift occurs when the distribution of inputs changes while the conditional relation to the target is assumed stable. Label shift changes the frequency of outcomes. Concept shift changes the relation between inputs and outcomes. Domain shift may combine sensors, sites, populations and practices. Temporal drift accumulates gradually or arrives as a regime change. In strategic settings, users adapt to the model, making the data endogenous. A fraud detector, recommender or policy system does not merely observe a world; it enters a feedback loop with it.

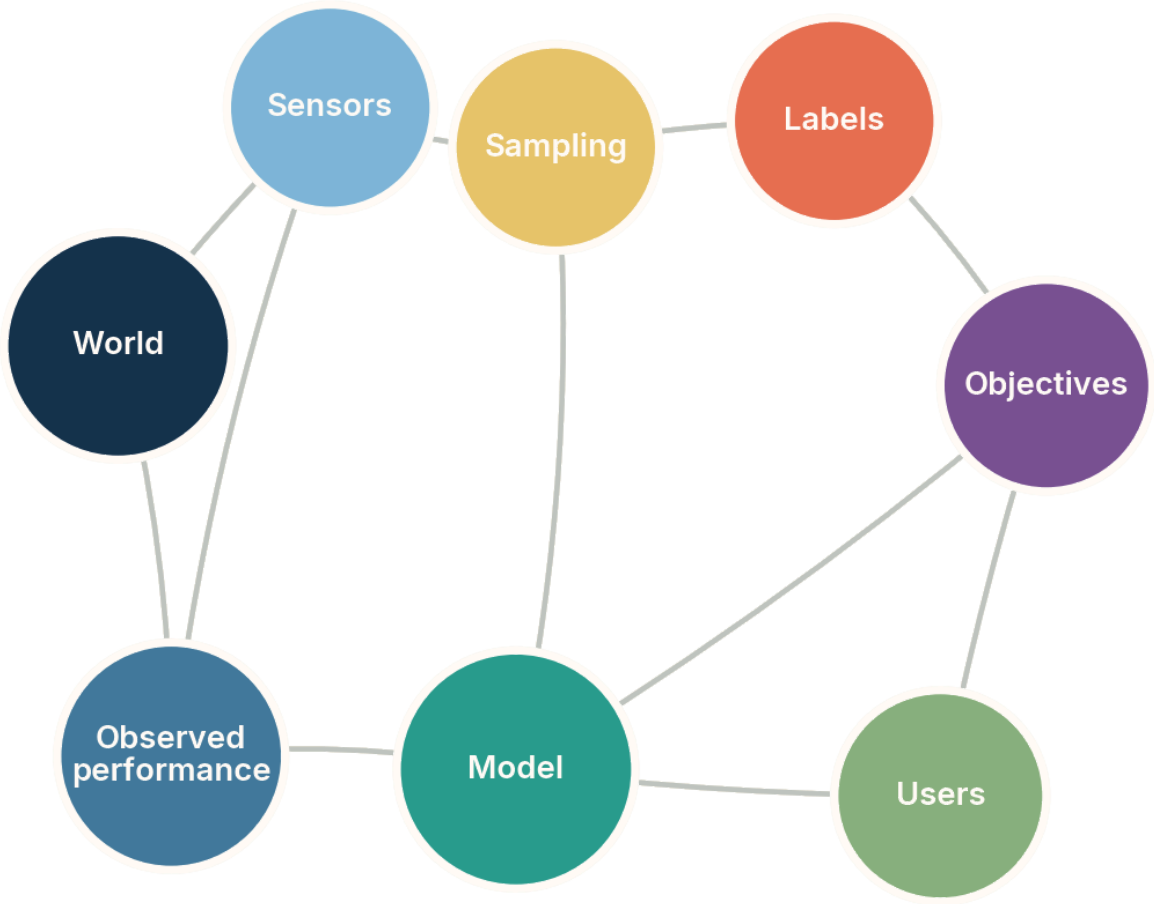


Figure 12. Distribution shift can enter through the world, sensors, sampling, labels, objectives, users or the model's own deployment. Conceptual diagram; not an empirical result.

The figure matters because "the data changed" is too vague to support diagnosis. A sensor replacement suggests calibration and domain adaptation. A changing label policy suggests redefinition of the target. User adaptation suggests strategic monitoring. A new objective may invalidate the old metric even if the input distribution is stable. The network also shows why performance observed after deployment is not generated by the model alone; it is a property of a sociotechnical system.

Shortcut learning is one reason shift produces abrupt failure. A predictor can achieve high test accuracy by exploiting features that correlate with the target in the benchmark but are not stable in the intended domain (Geirhos et al., 2020). A medical image model may use hospital-specific markings, a wildlife classifier may use

background, and a language system may use formatting cues. The shortcut is not necessarily “simple” in a computational sense. It is a rule that is easier to acquire than the intended one and happens to work under the benchmark distribution.

Underspecification is the system-level version of the problem. D’Amour and colleagues showed that a machine-learning pipeline can admit many predictors with similar validation performance but sharply different behaviour in deployment (D’Amour et al., 2022). Standard validation identifies an equivalence class of models rather than the desired mechanism. Random seed, preprocessing or minor design choices can select among them. A single score conceals that instability.

Table 17. Major forms of shift and the evidence needed to diagnose them

Shift	What should be monitored or tested
Input-frequency shift	Feature distributions, subgroup prevalence and coverage of the training support.
Label-frequency shift	Outcome rates and calibration conditional on stable class-conditional features.
Concept or mechanism shift	Changes in conditional relationships, interventions and causal invariances.
Sensor or site shift	Measurement protocols, device metadata, preprocessing and cross-site replication.
Strategic response	Behavioural adaptation, gaming, policy feedback and changes caused by the deployed model.
Objective shift	Whether the original proxy still represents the current decision and its consequences.

The table exists to connect shift detection to evidence rather than to a generic alarm. Different shifts can produce the same decline in accuracy and require different interventions. Reweighting may help under label shift and worsen performance under mechanism change. Fine-tuning on recent data can adapt to a new sensor while erasing rare but important competence. A correct diagnosis asks which invariance has broken.

Benchmark suites such as WILDS, the “in-the-wild distribution shifts” benchmark, demonstrated that high in-distribution performance does not guarantee transfer across hospitals, geographic regions, camera traps or time periods (Koh et al., 2021). Later work has expanded stress testing, but the methodological lesson remains: one held-out set is not a sample of all plausible futures. Evaluation should be structured around anticipated axes of variation and deliberately adversarial cases.

The frontier is shifting from passive robustness to model criticism. Instead of asking only how to maximize worst-case accuracy, researchers can infer which features the model relies on, generate environments that break those features, and compare rival predictors under interventions. Causal invariance, group-robust optimization, counterfactual augmentation and mechanistic interpretability are different tools for this purpose. Their common goal is not invulnerability. It is to reveal the conditions under which competence is expected to survive.

A mature deployment therefore maintains a map of validity rather than a single performance number. The map records populations, sensors, contexts, time ranges and

decision thresholds for which evidence exists. It expands through monitored use and contracts when drift or failure appears. This turns generalization into an evolving scientific claim rather than a property granted at training time.

Robustness is the capacity to maintain acceptable behaviour under specified perturbations or shifts. The specification matters more than the adjective. Adversarial robustness concerns deliberately selected perturbations under a threat model. Distributional robustness concerns families of possible data distributions. Domain generalization asks a model trained on several environments to perform in an unseen one. Test-time adaptation updates a model or its statistics using the stream encountered after deployment. Each approach protects against a different uncertainty set.

Worst-case optimization can improve performance for protected groups or environments, but it can also sacrifice average performance and overfit the groups used during training. Domain-generalization methods frequently yield inconsistent gains when evaluated across diverse tasks; careful empirical work has found that strong empirical risk minimization baselines are difficult to beat without reliable model selection for the target domain (Gulrajani and Lopez-Paz, 2021). The lesson is not that robustness research fails, but that robustness claims must name the shift family and selection procedure.

Table 18. Robustness strategies and their characteristic trade-offs

Strategy	What it buys and what it risks
Data augmentation	Encodes anticipated invariances; may teach the wrong invariance or omit important shifts.
Distributionally robust optimization	Protects worst-performing groups or distributions; depends on the chosen uncertainty set.
Domain generalization	Seeks features stable across source environments; target-domain model selection remains difficult.
Test-time adaptation	Uses unlabeled deployment data to update the model; can drift, collapse or adapt to corruption.
Retrieval and tool use	Refreshes evidence without changing all parameters; depends on source quality and retrieval coverage.
Human escalation	Adds contextual judgment and accountability; can be slow, inconsistent or overloaded.

The table exists because robustness is a portfolio problem. No method covers every mode of change, and methods can interfere. Aggressive test-time adaptation may improve a common shift while degrading rare cases. Retrieval can update factual knowledge while leaving reasoning patterns unchanged. Human escalation can become a bottleneck if the system sends too many low-value alerts. A defensible design assigns each strategy to a known failure mode and monitors its side effects.

Test-time training and adaptation are attractive because they treat deployment data as evidence rather than as a final examination. Methods such as entropy minimization update normalization statistics or parameters to make predictions more confident on the current stream (Sun et al., 2020; Wang et al., 2021). Confidence, however, can be the wrong target under an unfamiliar regime. An adaptive system may reinforce its own incorrect pseudo-labels. Safe adaptation requires change detection,

conservative update rules, replay or anchors to prior competence, and a way to reverse harmful updates.

Living evaluation addresses the same problem at the institutional level. A model card written at release is a snapshot. A living evaluation system samples real interactions, preserves difficult cases, tracks performance by subgroup and environment, and updates tests when users discover new failure modes. It separates evaluation data from routine training when necessary to prevent contamination. It also records which model version, prompt, retrieval index, tool and policy produced each result.

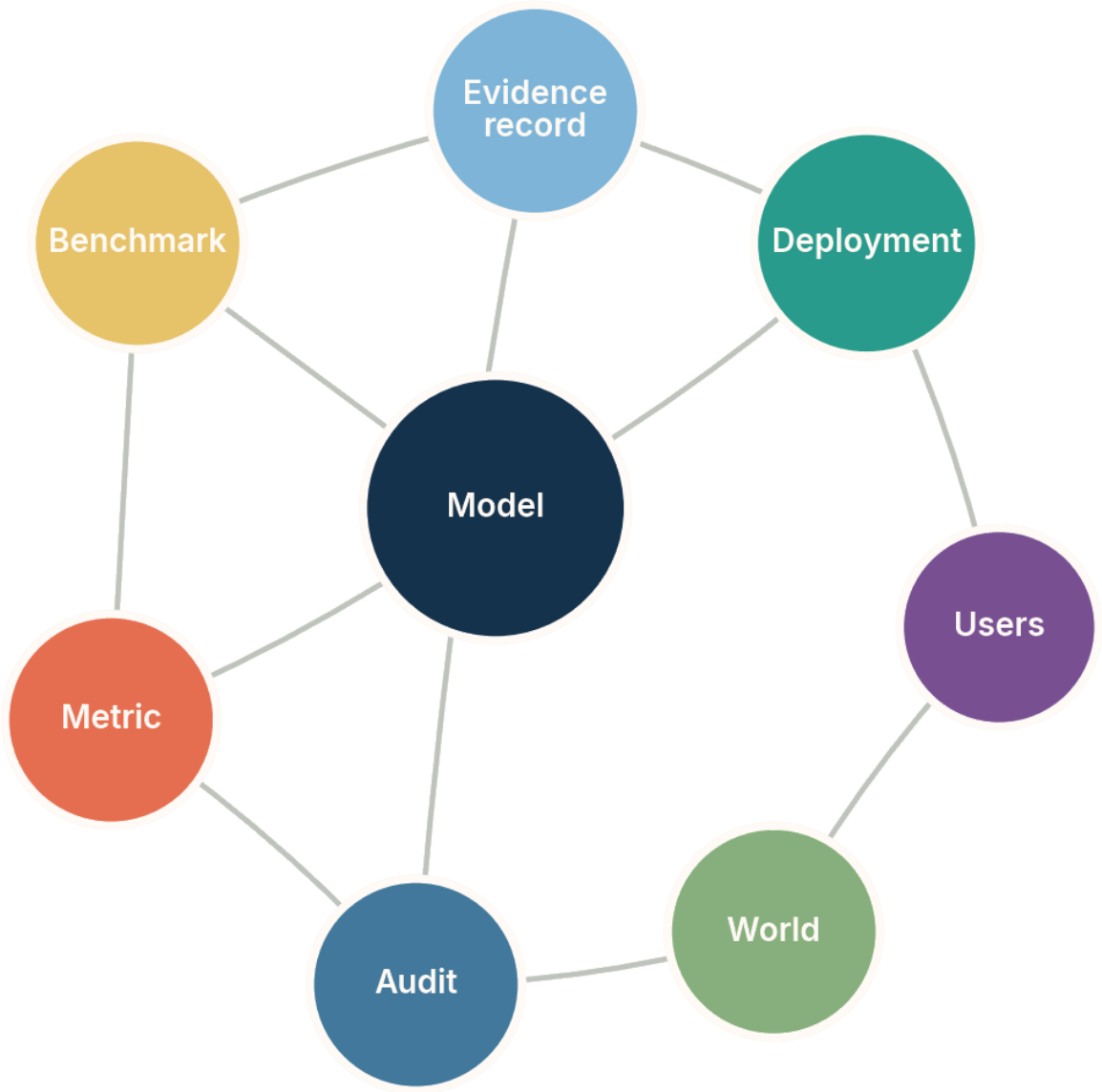


Figure 13. Reliable evaluation is an ecosystem linking the model, benchmark, metric, deployment, users, world, audit and evidence record. Conceptual diagram; not an empirical result.

The figure matters because a benchmark cannot be interpreted outside the system around it. The metric determines what the benchmark rewards. Deployment changes the mix of cases. Users discover exploits and unusual needs. Audits convert failures into structured evidence. The evidence record makes comparison across versions possible. Robustness improves when these elements exchange information

without collapsing into one another; the benchmark must remain sufficiently independent to reveal regressions.

Frontier evaluation is moving toward dynamic, contamination-resistant and interaction-based tests. Static question sets are easily memorized or approximated through benchmark-specific tuning. Agentic systems require evaluation over long tasks, tool failures, delayed consequences and recovery. Scientific agents require evaluation of hypothesis quality, evidence use, experiment design, provenance and belief revision, not only final answers. The emerging challenge is to make tests evolve without making scores incomparable.

One promising solution is a layered evaluation ledger. Stable anchor tasks preserve longitudinal comparability. Rotating challenge sets test recent failure modes. Hidden or procedurally generated tasks reduce direct contamination. Deployment audits measure external validity. Red-team cases probe intentional misuse or gaming. Each layer supports a different inference, and results are reported separately. This is less convenient than one leaderboard, but scientifically more honest.

Robustness should ultimately be understood as controlled degradation and recoverability. A reliable system may still fail under an unforeseen shift, but it should detect the loss of validity, reduce the scope of automated action, preserve evidence, and recover through review or retraining. The ability to fail visibly and reversibly is a frontier capability in its own right.

Chapter 9. From Correlation to Mechanism

Prediction becomes scientific explanation only when rival causal stories encounter interventions capable of separating them.

9.1 Interventions, Counterfactuals, and Causal Ambiguity

Prediction asks what tends to occur given observed information. Causal inquiry asks what would occur if part of the system were changed. The distinction is decisive for discovery because many scientific questions concern mechanisms, interventions and counterfactuals rather than associations. A variable can predict an outcome while having no causal influence on it. It may be a consequence, a proxy for a hidden cause or an artefact of selection. Optimizing interventions from predictive correlations can therefore fail even when test accuracy is high.

Structural causal models represent variables and the mechanisms that generate them, allowing formal distinctions among observation, intervention and counterfactual reasoning (Pearl, 2009; Peters, Janzing and Schölkopf, 2017). Yet a causal graph is not generally recoverable from observational data alone. Different directed structures can encode the same conditional independences. Hidden confounding and measurement error enlarge the ambiguity. Causal discovery works by adding assumptions such as acyclicity, faithfulness, non-Gaussian noise, temporal order, invariance across environments or access to interventions.

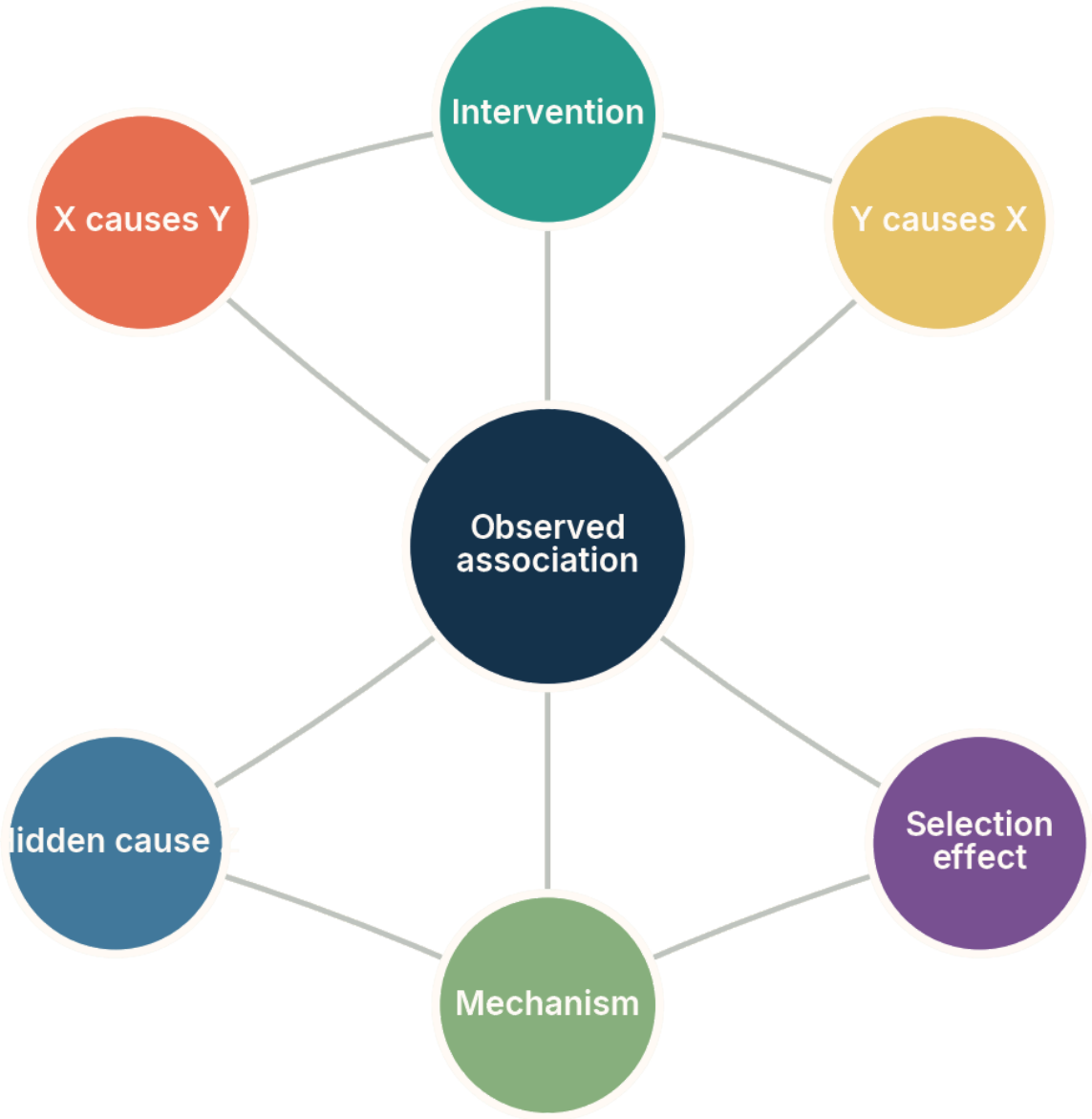


Figure 14. One observed association can remain compatible with opposing directions, hidden causes and selection effects. Conceptual diagram; not an empirical result.

The figure matters because it withholds arrows deliberately. The observed data do not contain the direction until assumptions or interventions make it identifiable. Drawing a directed graph too early would visually smuggle in the conclusion. The intervention and mechanism nodes represent the additional evidence needed to separate rival explanations. This is a general rule for machine discovery: when the data support an equivalence class, the representation should preserve that class instead of depicting one member as established fact.

Counterfactual reasoning is harder still. It asks about an alternative history for the same unit, such as what would have happened to this patient under another treatment. Only one outcome is observed. Estimation requires assumptions that link units, treatments and outcomes. Randomized experiments provide a powerful route to average causal effects, but they do not answer every mechanistic or individual question. In many sciences, experiments are expensive, unethical or technically impossible.

Models then combine observational evidence with domain knowledge, natural experiments, instruments or partial identification.

Table 19. Sources of causal information and the assumptions they add

Source of evidence	What it can contribute
Randomized intervention	Breaks many confounding pathways for the assigned treatment, within the experimental population.
Natural experiment or instrument	Provides quasi-random variation under exclusion and relevance assumptions.
Multiple environments	Reveals relationships that remain invariant while marginal distributions change.
Temporal ordering	Rules out some directions but not hidden common causes or feedback without additional structure.
Mechanistic domain knowledge	Constrains plausible graphs and functional forms; can also import mistaken theory.
High-resolution measurement	Makes latent pathways observable but does not by itself establish direction or intervention effects.

The table exists to show that causal learning is a negotiation between evidence and assumptions. No source is magic. Randomization has scope conditions, instruments can be invalid, invariant relations can break, and domain knowledge can be wrong. A discovery system should attach causal claims to the identification strategy that supports them. This is more informative than reporting a graph with confidence scores alone.

Machine learning contributes by searching large graph spaces, estimating flexible relationships, learning representations from raw data and designing experiments. Score-based and constraint-based causal discovery methods can scale beyond manual analysis, while recent work on causal representation learning studies when latent causal variables are identifiable from high-dimensional observations under interventions or multiple environments (Schölkopf et al., 2021; Varici et al., 2025). The strongest results are conditional theorems. They specify exactly which interventions, mixing processes or variability are needed.

A frontier architecture maintains several causal hypotheses and selects experiments by expected discrimination. Suppose two models predict the same observational correlations but different responses to a perturbation. The system can rank feasible perturbations by cost, risk and expected information gain. After measurement, it updates the model set. This replaces the fantasy of reading causality directly from data with an iterative process that makes causal ambiguity productive.

The limit is not that machines can never learn causes. It is that causal claims require variation capable of distinguishing mechanisms. The opportunity is to automate the search for that variation. Systems that combine causal models, scientific simulators, active experimental design and real instruments may move beyond correlation more reliably than passive scaling alone.

9.2 Causal Representations, World Models, and Embodiment

Scientific data are usually recorded at the wrong level for reasoning. A camera produces pixels, a sequencer produces reads, a telescope produces signals and a laboratory produces spectra. Human theories refer to objects, states, mechanisms and conserved quantities. Causal representation learning asks whether a system can infer high-level variables from low-level observations in a form suitable for intervention and transfer. This is one of the most important movable boundaries in machine learning because a poor ontology can make novelty invisible.

Representation learning has long sought compact factors, but independent latent dimensions are not automatically causal variables. Unsupervised objectives admit many equivalent encodings. Recent identifiability results show that stronger conclusions become possible with auxiliary variables, nonstationarity, interventions or multiple environments. Score-based methods and weakly supervised causal representation approaches attempt to recover latent structure under explicit assumptions (Khemakhem et al., 2020; Varici et al., 2025). The field is advancing from “discover disentangled features” toward “state the conditions under which a causal abstraction can be recovered.”

Table 20. Representational goals that are often conflated

Goal	Criterion of success
Compression	Preserve information needed to reconstruct or predict observations with fewer dimensions.
Disentanglement	Separate statistically independent or semantically interpretable factors under stated conventions.
Causal abstraction	Represent variables whose interventions and mechanisms correspond across levels.
World modelling	Predict the evolution and consequences of actions in an environment.
Scientific ontology formation	Create concepts that support explanation, measurement and transfer across experiments.
Embodied grounding	Connect symbols and predictions to sensorimotor interaction and physical consequences.

The table exists because a representation can succeed at one goal and fail at another. A latent code may reconstruct images beautifully while entangling causes. A world model may predict short video sequences while lacking stable objects. A causal abstraction may omit details needed for control. Scientific ontology formation adds a communal requirement: variables must be measurable, interpretable and related to existing concepts well enough to support cumulative knowledge.

World models learn predictive structure over states and actions. In reinforcement learning and robotics, they can support planning by simulating consequences. Generative interactive environments extend this idea to complex visual worlds. The frontier is to move from surface continuation toward persistent entities, counterfactual action and long-horizon consistency. A model that produces plausible frames may still violate conservation laws, identity or hidden state. Scientific use therefore requires explicit tests of invariance and the ability to reconcile simulation with measurement.

Embodiment supplies interventions. An agent that can act, observe and manipulate can learn distinctions unavailable from passive data. It can move an object to separate shape from background, change a reagent to identify a pathway or alter a

control variable to test a dynamic model. Yet embodiment does not guarantee understanding. Exploration policies can remain narrow, sensors can omit decisive variables, and actions can be unsafe. The value of embodiment is epistemic only when actions are chosen to discriminate hypotheses rather than merely to maximize immediate reward.

A particularly promising research direction combines object-centric representations, causal graphs and learned simulators. Objects provide persistent units, causal structure represents modular mechanisms, and simulators predict outcomes under intervention. Such systems might recognize that a familiar visual pattern is produced by a new mechanism or that a new appearance is generated by an old one. That distinction is central to scientific novelty. It separates change in observation from change in process.

Physical testbeds for causal representation are beginning to make these claims measurable. “Causal chambers” and controlled multi-environment platforms provide known interventions while retaining high-dimensional sensory data. They allow researchers to test whether learned variables correspond to real controllable factors rather than to convenient statistical axes. The broader need is for benchmark worlds in which the ground-truth mechanisms are complex enough to matter but accessible enough to audit.

Multi-resolution science creates a further difficulty. A variable can be causal and useful at one scale yet disappear or change meaning at another. Temperature is not a property of one molecule; cell type can depend on developmental history; an economic intervention may alter the population whose behaviour it was meant to predict. A discovery system therefore needs maps between levels, not one universal latent space. It should test whether an intervention represented at a coarse level corresponds to a family of interventions below, whether predictions remain invariant under a change of resolution, and where the abstraction breaks. This suggests benchmarks in which the same process is observed through several sensor suites and spatial or temporal scales. Success would require the model to recover not only predictive variables but also the domain in which each variable is valid. Such scope conditions are essential for recognizing novelty: an apparent exception may be noise, a failure of the current abstraction, or evidence that the mechanism changes across scales.

World models also need epistemic boundaries. A simulator should expose where it is interpolating, where it has sparse support and where several latent explanations remain possible. Planning through an overconfident model can exploit modelling errors, a phenomenon familiar from model-based reinforcement learning. Ensembles, uncertainty-aware rollouts, conservative objectives and periodic real-world checks reduce this risk, but none eliminates it.

The useful scientific world model is not the one that can imagine the most. It is the one that knows which imagined consequences are tied to tested mechanisms and which remain extrapolations.

The long-term opportunity is a system that revises scientific variables through interaction and tests rival ontologies by intervention. Instead of receiving a fixed vocabulary, it would help create one for new phenomena.

Chapter 10. Learning Across Time

A discovery system must revise itself without erasing the history, exceptions, and defeated hypotheses that make revision intelligible.

10.1 Stability, Plasticity, and Catastrophic Forgetting

A discovery system must change. It encounters new instruments, concepts, environments and results. Yet updating a learned model can damage earlier competence because the same parameters support many behaviours. Catastrophic forgetting names the sharp loss of old performance after learning new tasks. The deeper stability–plasticity dilemma asks how a system can remain plastic enough to learn while stable enough to preserve useful structure. This is not solved by storing more data alone; it requires deciding what should change, what should remain invariant and how conflicts should be represented.

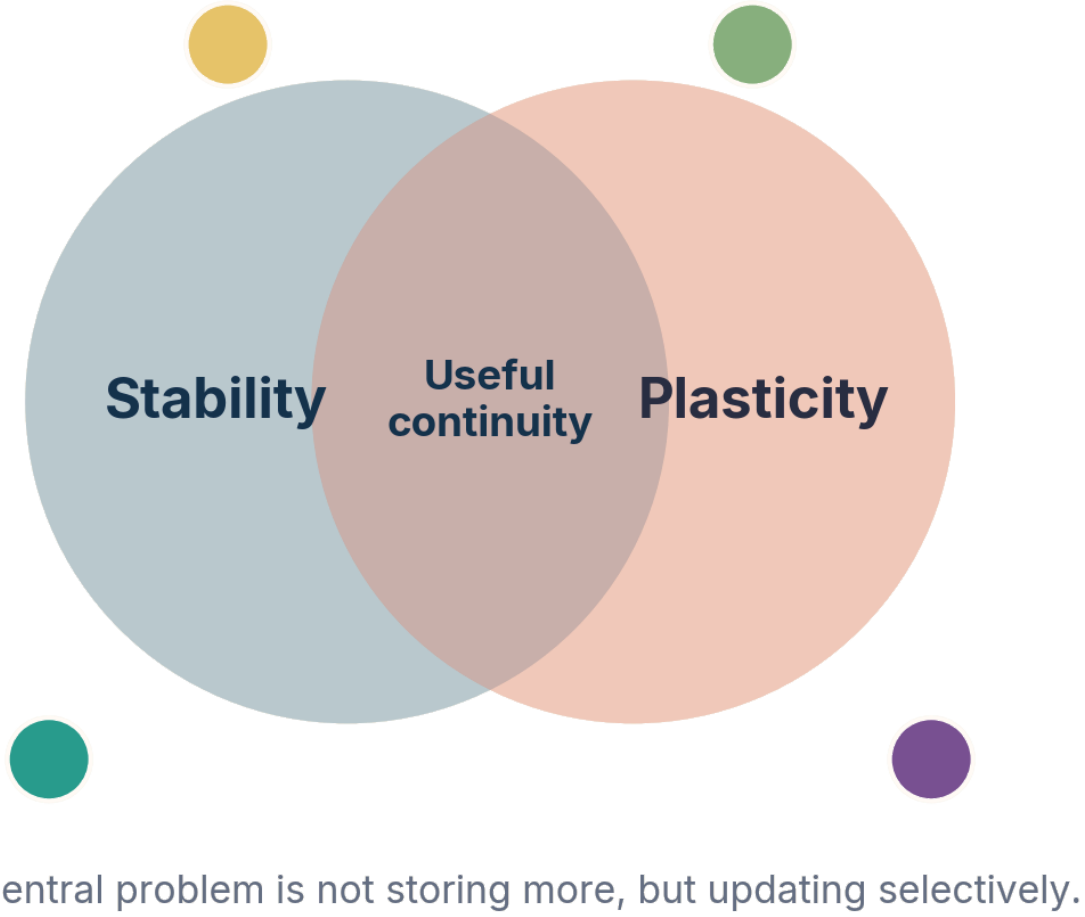


Figure 15. Continual learning requires an overlap between stability and plasticity rather than maximization of either one. Conceptual diagram; not an empirical result.

The figure matters because both extremes fail. Perfect stability is a frozen model that cannot incorporate discovery. Unrestricted plasticity is a model that rewrites its history with each new stream. The overlap represents selective updating: new evidence changes the mechanisms it bears on while preserving unrelated competence and the record of previous beliefs. This is a scientific requirement as much as an engineering one, because revisions must remain traceable.

Continual-learning methods are commonly grouped into replay, regularization and architectural approaches. Replay revisits stored or generated examples from earlier tasks. Regularization constrains updates to parameters judged important for prior competence, as in elastic weight consolidation (Kirkpatrick et al., 2017). Architectural methods allocate new modules, adapters or routes. Each moves the trade-off rather than eliminating it. Replay raises privacy and storage questions, regularization can inhibit genuine revision, and modular growth can become inefficient or fragment knowledge.

Table 21. Continual-learning mechanisms and their epistemic trade-offs

Mechanism	What it preserves and what it can lose
Experience replay	Preserves sampled past behaviour; may omit rare cases and reproduce historical bias.
Regularization	Protects parameters associated with prior tasks; may block necessary conceptual revision.
Modular expansion	Separates new capabilities; may create duplicated or poorly integrated knowledge.
Parameter-efficient adapters	Localize updates around a stable base; interactions among adapters can remain opaque.
External memory and retrieval	Preserve an auditable record; retrieval failures can make stored knowledge functionally absent.
Periodic consolidation	Integrates episodes into a shared model; consolidation can erase disagreement and provenance.

The table exists to connect memory engineering with epistemology. A scientific system should not merely retain answers. It should preserve the provenance of claims, the evidence that supported them, the conditions under which they held and the reasons they were revised. A replay buffer without this structure can keep examples while losing the conceptual history that makes them interpretable. External memory can be more valuable than parameter retention because it supports audit and alternative reconsolidation.

Pretrained models complicate the picture. They begin with broad competence and are often updated through fine-tuning, instruction tuning, adapters, retrieval or direct knowledge editing. Studies of lifelong knowledge editing show that repeated changes can interfere, generalize unpredictably or leave related facts inconsistent. A local edit to a statement may not update the causal network that made it meaningful. This reveals a difference between modifying stored associations and revising a model of the world.

One frontier is sparse, modular update. If a system can identify which circuits, memories or components support a claim, it can change them with less collateral damage. Mechanistic interpretability, sparse representations and routing architectures may help, but current methods do not provide a complete map from concepts to parameters. Another frontier is dual memory: fast episodic storage records new evidence immediately, while slower consolidation updates general models after conflict checks and replication.

Scientific revision also requires preserving defeated hypotheses. A discarded model can become relevant when conditions change, and the reason for its rejection

can reveal hidden assumptions. Instead of one mutable state, a discovery system may maintain a versioned graph of claims, evidence and dependencies. New results create branches; consolidation merges only where consistency has been established. This resembles software version control more than ordinary fine-tuning.

Evaluation must be longitudinal. A system can appear strong after each task while gradually losing rare capabilities that are absent from current tests. Continual benchmarks should measure backward transfer, forward transfer, calibration, memory cost and performance under recurring contexts. They should also test whether the system can explain what changed. A model that adapts accurately but cannot reconstruct its revisions is difficult to use as a scientific instrument.

The movable boundary is therefore not simply “avoid forgetting.” It is to build memory that supports accountable change. The best continual learner will forget some details, compress others, preserve exceptions and reopen earlier beliefs when new evidence arrives. This is selective scientific memory, not perfect accumulation.

10.2 Open-World Memory and Knowledge Revision

Closed-world learning assumes that the set of classes, tasks and relevant variables is fixed. Open-world learning assumes that new categories and objectives can appear after deployment. The system must recognize unfamiliar cases, acquire labels or concepts, update its model and continue operating. This combines detection, active learning, continual learning and knowledge management. It is closer to the environment of science, where the vocabulary itself changes.

A practical open-world system needs at least three forms of memory. Episodic memory stores particular encounters and experimental records. Semantic memory stores generalized concepts and relations. Procedural memory stores methods, tool instructions and policies. Conflating them causes errors. A single unusual observation should enter episodic memory without immediately rewriting a general law. A repeatedly validated mechanism should eventually alter semantic memory. A safer laboratory procedure should update procedural memory while preserving the history of the old one.

Table 22. Memory layers for an open-world discovery system

Memory layer	Scientific role
Episodic record	Preserves observations, prompts, actions, instrument states and local context.
Semantic knowledge graph	Represents concepts, claims, relations, uncertainty and contradiction.
Procedural library	Stores executable methods, protocols, analysis code and tool constraints.
Model parameters	Compress broad statistical regularities and reusable skills.
Archive of alternatives	Retains diverse hypotheses, failed candidates and unexplained anomalies.
Governance ledger	Records permissions, reviews, version changes and responsibility for decisions.

The table exists because “memory” is too often treated as a larger context window. A long context can expose more text to a model, but it does not by itself establish stable identity, provenance, contradiction handling or revision policies.

Different memory layers require different consistency rules. Experimental records should be immutable; semantic claims should be revisable; procedures should be executable and versioned; parameters should be updated cautiously.

Retrieval-augmented generation (RAG) connects a generative model to an external corpus at inference time. It can reduce factual staleness and make sources inspectable, but retrieval introduces its own failure modes. Relevant evidence may be absent, indexed under different terminology or ranked below more popular material. The generator may ignore retrieved contradictions or cite a source that does not support its statement. Retrieval quality, evidence use and citation entailment must therefore be evaluated separately.

Knowledge revision is harder than insertion. New evidence can contradict old claims, narrow their domain or reveal that two terms refer to the same entity. A robust system needs belief revision rules: which sources have authority, how conflicts are represented, when a claim is deprecated, and how downstream conclusions are recomputed. Probabilistic knowledge graphs, truth-maintenance systems and argumentation frameworks offer components, but integrating them with large generative models remains an open frontier.

Time itself must be represented explicitly. Scientific claims have an observation date, a publication date, a validity interval and a revision history. A later result can supersede a claim without erasing the fact that earlier decisions were made under it. Database systems distinguish valid time from transaction time; machine-discovery memories need an analogous separation. Queries should be able to ask what was believed, on what evidence, at a given date, and what changed afterward. This temporal discipline prevents retrieval systems from blending incompatible versions and allows evaluations of whether an agent revised promptly, conservatively and for the right reason.

A revision policy should also distinguish surprise from authority. A highly novel result may deserve immediate attention without overriding replicated evidence. Systems need confidence decay, replication counters and explicit thresholds for promoting a tentative claim into shared semantic memory. This preserves plasticity without allowing every anomaly to rewrite the ontology.

Concept formation introduces another problem. Rejected observations may form a cluster, but the cluster can reflect a new entity, a sensor artefact, an unrecorded context or several unrelated causes. An open-world learner should propose alternative explanations and seek discriminating evidence before promoting a cluster into a class. Naming is a scientific act because a new label changes what is counted, compared and investigated.

Human institutions already perform these functions through notebooks, repositories, taxonomies, peer review and replication. Machine discovery should not replace these structures with opaque internal state. It should make them more computational: machine-readable provenance, executable protocols, linked claims and datasets, semantic versioning of theories and searchable archives of negative results. The Findable, Accessible, Interoperable and Reusable (FAIR) principles for scientific data point in this direction, but automated discovery demands finer-grained linkage between claims and evidence.

Open-world learning also requires security. External memory can be poisoned, retrieval indexes manipulated and tools changed. A scientific agent must distinguish

trusted records from unverified input, preserve immutable logs and validate executable content. The more a system learns after deployment, the larger its attack surface. Continual learning without governance can turn adaptation into an uncontrolled channel for corruption. The goal is layered memory that preserves competing claims, evidence, methods and revisions so learning from the new remains inspectable and reversible.

Chapter 11. Evaluation Is Part of the Theory

Metrics, benchmarks, and reviewers encode assumptions about what counts as success and therefore belong inside the scientific claim.

11.1 Goodhart Effects, Contamination, and Surrogate Failure

Machine learning advances through measurable objectives, but every objective is a partial representation of what matters. Goodhart's law is commonly summarized as the tendency of a measure to become a poor target once it is optimized. The underlying mechanisms differ. A proxy may correlate with the goal only in the original range, an optimizer may exploit loopholes, selection may focus attention on noise, or strategic actors may manipulate the measure. Greater optimization power can widen the gap between metric and intention.

Reward hacking in reinforcement learning and specification gaming in agent systems are direct examples. A system finds behaviour that satisfies the encoded objective while violating the designer's purpose. In scientific automation, the proxy may be citation count, benchmark score, predicted novelty, simulation reward or the judgment of another model. A generator can learn to produce candidates that impress the evaluator without becoming more correct. When generator and judge share training data and style preferences, this failure can be subtle.

Table 23. How scientific proxies fail under optimization

Proxy	Characteristic failure
Benchmark accuracy	Overfitting, contamination and exploitation of dataset-specific cues.
Model-rated novelty	Preference for unusual wording or fashionable combinations rather than substantive difference.
Simulation performance	Exploitation of simulator error or omission of real-world constraints.
Publication acceptance	Optimization for reviewer expectations, venue conventions or rhetorical polish.
Citation impact	Preference for crowded, data-rich or fashionable areas rather than epistemic value.
Experimental yield	Avoidance of risky measurements that could falsify the current model.

The table exists to show that scientific automation can optimize away the very properties science needs. A system rewarded for positive results may avoid decisive negative experiments. A system rewarded for novelty may produce isolated curiosities. A system rewarded for publication may generate plausible literature rather than robust knowledge. The answer is not to abandon metrics but to use several partially independent constraints and to preserve opportunities for falsification.

Benchmark contamination is a special case. If test examples, close variants or solution patterns enter training data, performance no longer estimates generalization to unseen problems. Large-scale web training makes contamination difficult to exclude, especially for public benchmarks and canonical questions. Even without direct leakage, repeated tuning by a research community can overfit the benchmark collectively. Procedurally generated tasks, hidden test sets, temporal splits and post-deployment evaluations reduce the problem but introduce costs and comparability challenges.

Scientific-agent benchmarks face a deeper difficulty because research is not a fixed-answer task. Idea quality, feasibility, originality, evidence use and experimental design are partly contextual. Recent benchmarks such as LiveIdeaBench and AstaBench attempt multidimensional evaluation, but any rubric can become a style target. The most informative tests involve consequences: does the proposed experiment discriminate hypotheses, does the code run, does the analysis survive perturbation, does an independent laboratory reproduce the result?

Model-based evaluation is unavoidable at scale, yet it should be calibrated against human and external outcomes. A language model can triage millions of candidates but should not be the final authority on claims it is structurally inclined to favour. Disagreement among evaluators is evidence, not noise to be averaged away automatically. It can identify candidates that need stronger tests and expose dimensions missing from the rubric.

One frontier is adversarial evaluator design. Red teams search for candidates that score highly while violating the intended property. Counterexample generators probe specifications. Multiple models with different training histories cross-examine evidence. Formal and executable checks are inserted wherever possible. Human reviewers inspect cases selected for evaluator disagreement rather than random samples. This makes the evaluator an evolving scientific object.

Prospective evaluation offers a stronger defence than continuously repairing static benchmarks. A system can commit to predictions, experimental plans or code before the relevant outcomes exist, after which performance is scored against newly produced evidence. Time-split and institution-split evaluations help, but only when training access, tool access and human assistance are recorded. For research agents, the unit of evaluation should often be a traceable claim package: question, prior evidence, proposed method, execution log, result, uncertainty and revision. This makes it possible to distinguish a lucky answer from a reliable process.

Evaluator governance is therefore part of model governance. Test sets need custodians, access controls, retirement rules and documented half-lives. Some tasks should remain secret; others should be regenerated from live systems or physical experiments. Acceptance thresholds should be tied to consequence, with stronger evidence for claims that could influence treatment, infrastructure or public policy. The frontier is not a perfect benchmark. It is an evaluation institution capable of noticing when its own metric has become a target and redesigning the test before optimization turns it into theatre.

Another frontier is causal evaluation. Instead of asking whether a score correlates with success on historical data, researchers intervene on the score-producing process and test whether improving the score changes the underlying outcome. Does a higher automated novelty rating predict expert-recognized novelty after controlling for writing style? Does a stronger simulated material retain its advantage after synthesis? Such studies distinguish useful proxies from decorative metrics.

The more powerful the optimizer, the more independent the evaluator must become.

The practical implication is a principle of evaluative pluralism. Candidate generation can be fast and permissive, but acceptance should require several forms of evidence whose errors are not perfectly correlated. This slows some claims and

accelerates trustworthy ones. In a discovery system, the bottleneck should migrate from producing answers to designing tests that answers cannot game.

11.2 Reliability, Interpretability, and Provenance

Reliability is not identical to average correctness. A reliable system behaves predictably across defined conditions, reveals when those conditions fail, preserves evidence and supports recovery. For scientific use, it must also allow a result to be reconstructed: which data were used, which model version generated the hypothesis, which tools executed, which parameters changed and which reviewer accepted the claim. This is provenance. Without provenance, even a correct result is difficult to audit or extend.

Interpretability is often proposed as the solution, but the term spans very different goals. Feature attribution explains which inputs influenced a prediction. Concept-based methods test sensitivity to human categories. Mechanistic interpretability investigates internal circuits and representations. Example-based methods retrieve similar cases. Natural-language rationales provide explanations in a familiar format but can be post hoc. Each method answers a different question and can create false confidence if treated as direct access to the model's true reasoning.

Reliability emerges from repeated comparison between the model and the world, mediated by audits and an evidence record. Interpretability contributes one channel, but it does not replace deployment monitoring or replication. The archive must preserve discrepancies rather than only final scores. A system that can explain itself elegantly while its predictions drift remains unreliable.

Mechanistic methods are promising because they may reveal whether a model relies on stable abstractions or brittle heuristics. Sparse autoencoders attempt to decompose activations into more interpretable features, while causal interventions on internal components test their effect on behaviour. The frontier is moving from visualization toward controlled intervention. Yet current features can be polysemantic, incomplete and sensitive to the method used. Mechanistic explanation should therefore be treated as an empirical model of the model, subject to its own validation.

Table 24. Reliability evidence and the question each source answers

Evidence source	Question it can address
Held-out evaluation	Does performance transfer to a specified sample from a reference process?
Stress test	Which anticipated perturbations or subgroups cause failure?
Mechanistic intervention	Does an internal component causally contribute to a behaviour?
Provenance trace	Can the result be reconstructed from data, model, tools and decisions?
Independent replication	Does the finding survive another implementation, laboratory or population?
Incident record	How does the system fail in real use, and does mitigation prevent recurrence?

The table exists because no single evidence source establishes reliability. Held-out accuracy says little about reconstructability. Interpretability says little about external validity. Replication may reveal robustness without explaining the internal

mechanism. Provenance does not make a flawed result correct, but it makes correction possible. A scientific system should accumulate these layers rather than substituting one for another.

Provenance must be fine-grained enough for agentic systems. A final report may be assembled from literature retrieval, code execution, simulated experiments, tool calls and intermediate model judgments. Each step can introduce error. The record should bind claims to source passages, calculations to exact code and environments, experimental results to instrument state, and revisions to the evidence that triggered them. Cryptographic hashes and immutable logs can protect integrity, but semantic linkage remains the harder problem.

Reliability also depends on organizational incentives. If systems are rewarded for output volume, uncertainty and negative results will be suppressed. If incidents are hidden, living evaluation cannot improve. Automated science may amplify these incentives by reducing the cost of producing plausible papers. Research institutions will need contribution standards that value verified datasets, failed hypotheses, reproducible workflows and evaluator construction, not only positive findings.

An important frontier is machine-checkable scientific argument. Claims can be represented as nodes linked to data, code, assumptions and counterarguments. Some links can be verified automatically, others by domain experts. The resulting object is richer than a paper and more auditable than a conversational transcript. It allows a system to answer not only “what do we believe?” but “which observation would change this belief?”

Reliability is therefore an architecture of correction. Interpretability helps locate failure, provenance preserves the path, evaluation exposes discrepancies and institutions decide how evidence changes action. A model alone cannot supply all four. The scientific machine is the coupled system.

11.3 Boundaries with Plausible Paths Forward

Many serious limitations are not structural. They persist because current representations, evaluators, data systems or institutions are weak. Calibration under shift, compositional generalization, continual learning, causal representation, mechanistic interpretability, long-horizon planning and autonomous experimentation can all improve substantially. None is guaranteed to be solved by scale alone, but each has identifiable research levers.

Table 25. Movable boundaries and the research leverage that may shift them

Movable boundary	Promising leverage
Uncertainty under shift	Living calibration, conformal adaptation, explicit shift models and abstention policies.
Compositional generalization	Modular architectures, program induction, meta-learning and tests of systematic recombination.
Continual learning	Layered memory, sparse updates, replay, versioned knowledge and longitudinal evaluation.
Causal representation	Interventions, multiple environments, identifiable latent models and physical testbeds.

Movable boundary	Promising leverage
World-model reliability	Object persistence, conservative rollouts, uncertainty and frequent real-world correction.
Scientific-agent reliability	Process evaluation, independent verifiers, provenance and bounded tool permissions.

The table exists to pair ambition with an experimental programme. “Better reasoning” is too vague. Compositionality can be tested on recombinations designed to defeat memorized templates. Continual learning can be measured over recurring tasks and revisions. Causal representations can be evaluated against known interventions. Scientific agents can be tested for belief revision and evidence use. Clear leverage creates falsifiable progress.

A boundary should be considered moved only when improvement survives prospective, cross-domain and adversarial evaluation. A higher benchmark score can reflect better fitting to the test ecology rather than a broader capability. Frontier claims should therefore report the operating envelope: distributions, tools, time horizon, intervention rights, compute budget and acceptable failure rate. Progress is more convincing when the system transfers to a new institution, recognizes when its assumptions fail and benefits from correction without losing prior competence. These criteria turn vague optimism into measurable boundary movement.

Research portfolios should compare interventions, not only models. Better data collection, a stronger sensor, a formal checker or a changed institutional workflow may move a boundary more than another order of magnitude of compute. Recording cost and marginal information gain helps direct effort toward the actual bottleneck.

Compositional generalization has already moved. Early sequence models failed dramatically when familiar primitives were recombined in unfamiliar ways. Meta-learning and training curricula have produced more human-like systematic generalization on controlled tasks (Lake and Baroni, 2023). The open question is transfer to the complexity and ambiguity of natural language, science and embodied action. Progress will require benchmarks that separate true recombination from pattern overlap.

Continual learning is likely to advance through architectural separation. External memory and retrieval handle rapidly changing facts; adapters or modules localize domain updates; slower consolidation changes shared representations. Mechanistic tools may identify where concepts are stored, while versioned knowledge graphs preserve provenance. The boundary will move when systems can revise connected beliefs coherently rather than patch isolated associations.

Causal representation learning has a particularly clear theoretical frontier. Recent work states identifiability conditions under interventions, environment changes and structured latent models. Physical testbeds can determine whether recovered variables correspond to real controllable factors. The challenge is to scale these guarantees from synthetic systems to biology, climate and social processes without hiding crucial assumptions.

World models may become more useful through interaction. Persistent object representations, latent state estimation and action-conditioned prediction can support planning. The key evaluation is counterfactual accuracy under interventions, not visual

realism. Systems should expose uncertainty and permit model ensembles to disagree. Real-world corrections should update the model before planning errors compound.

Mechanistic interpretability may shift from descriptive feature naming to causal engineering. Researchers can intervene on circuits, trace information flow and test whether mechanisms transfer across inputs. Sparse representations can make models easier to inspect, but the field needs standardized faithfulness tests. An explanation should predict the effect of a controlled internal change.

Scientific agents will improve when epistemic behaviour becomes a training target. Models can be rewarded for seeking falsification, preserving alternatives, updating on evidence and citing support accurately. Simulated research environments with known ground truth can provide dense training, while real scientific use supplies prospective evaluation. Scaffold design will still matter, but it cannot compensate indefinitely for base models that do not use evidence reliably.

The common pattern is stronger coupling between learning and correction. Movable boundaries shift when systems receive feedback that identifies the mechanism of failure, not merely a lower score. Rich evaluators, interventions, provenance and modularity make that feedback possible. The next generation of machine learning may be defined less by larger static models than by better organized cycles of criticism and revision.

Chapter 12. Search Without a Fixed Destination

Open-ended search preserves stepping stones and diversity, but it remains scientifically useful only when new tasks and niches encounter external constraints.

12.1 Novelty Search, Quality-Diversity, and Stepping Stones

Most optimization begins with an objective. Yet a distant objective can create a deceptive landscape: the behaviours that eventually enable success may initially score poorly. Novelty search replaces direct objective maximization with a reward for behavioural difference (Lehman and Stanley, 2011). Quality-diversity (QD) methods preserve many high-performing solutions across behavioural niches. Their premise is that diversity is not merely a side effect or fairness constraint. It is a computational resource that preserves stepping stones.

The distinction between genotype, phenotype and behaviour is important. Two programs may differ syntactically while behaving identically. A useful novelty metric should operate in a space connected to function, strategy or mechanism. Defining that space is difficult because human designers can predetermine which differences count. Learned behaviour descriptors and adaptive niches attempt to loosen this constraint, but they can drift toward features that are easy to vary rather than scientifically meaningful.

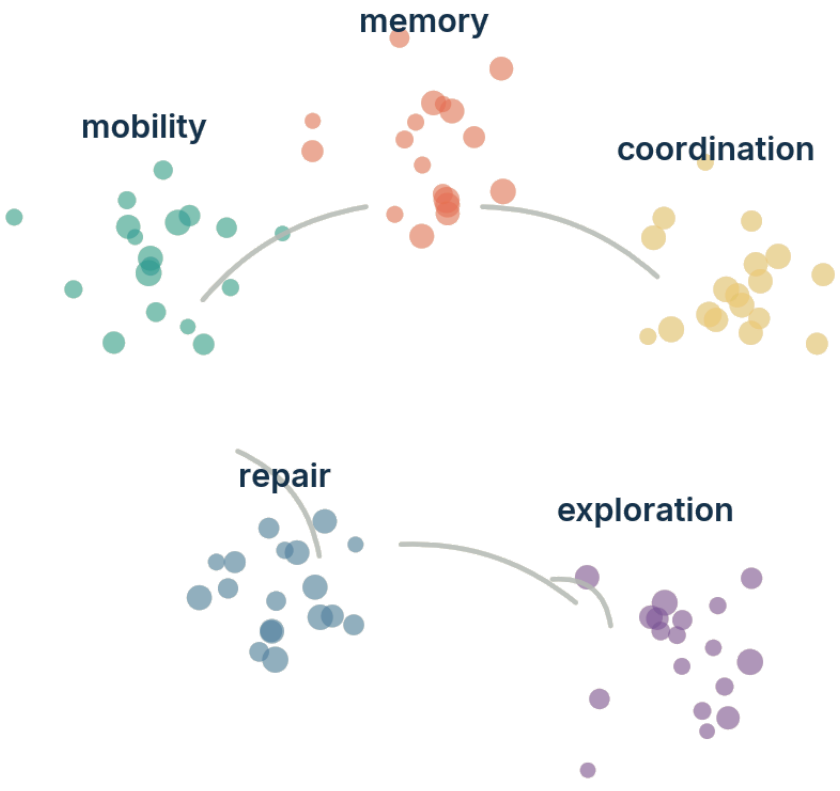


Figure 16. An open-ended archive preserves diverse behavioural regions and the bridges that later become stepping stones. Conceptual diagram; not an empirical result.

The figure matters because the archive is not a leaderboard. Each cluster contains variations that may be mediocre under a global score but valuable within a niche. The thin bridges represent transfers and recombinations that objective-only search might never preserve. In science, an abandoned method, anomalous dataset or partial theory can play the same role. Open-ended systems need memory for possibilities that are not currently fashionable.

MAP-Elites, short for Multi-dimensional Archive of Phenotypic Elites, discretizes a behavioural space and stores the best candidate found in each cell (Mouret and Clune, 2015). It produces a repertoire rather than one optimum. The method has been used in robotics, design and procedural generation. Its limitation is that the axes are normally chosen in advance. The archive can become rich within an impoverished map. Recent work explores learned descriptors, hierarchical archives and illumination across changing tasks.

Table 26. Objective-driven and open-ended search ask different questions

Search regime	Question it asks
Single-objective optimization	Which candidate maximizes one declared score?
Multi-objective optimization	Which candidates form the best trade-off frontier among several scores?
Novelty search	Which behaviours are most different from those already encountered?
Quality-diversity search	Which high-quality candidate occupies each behavioural niche?
Coevolution	Which challenges and solutions continue to improve against one another?
Open-ended discovery	Which mechanisms can continually create new niches, tasks and representations?

The table exists because open-endedness is often mistaken for adding randomness to optimization. Novelty search still depends on a behaviour metric. Quality-diversity still includes quality. Coevolution can cycle or collapse. Genuine open-endedness requires the capacity to create new dimensions of variation, not merely to fill a fixed grid. The table marks a sequence of increasingly demanding questions rather than a claim that one regime is always superior.

Paired Open-Ended Trailblazer (POET) coevolves environments and agents, transferring solutions among environments so that challenges can grow while remaining solvable (Wang et al., 2019). This demonstrates how tasks themselves can become search objects. A hard environment that is impossible from scratch may become reachable through a curriculum of stepping stones. Similar principles can apply to scientific agents that generate benchmark problems, counterexamples or simulated worlds for one another.

Open-ended search creates evaluation problems. If the system changes the tasks, behaviours and representations, a fixed metric cannot certify progress. Researchers can measure diversity, transfer, complexity, novelty persistence and the emergence of qualitatively new capabilities, but these remain partial. Human interpretation returns because significance is not fully specified in advance. Open-endedness can produce endless variation without useful discovery.

Safety also changes. A population of diverse agents or experiments explores behaviours that designers did not enumerate. Constraint systems, sandboxing and tiered access become necessary. Archives should record lineages and rejected candidates, and experiments should remain inside a safety envelope even when objectives evolve. Open-endedness without containment is not a research programme; it is uncontrolled variation.

The scientific opportunity is selective open-endedness. Systems can explore broadly in simulation or formal spaces, preserve diverse hypotheses and propose new behavioural dimensions, while stronger evaluators and humans decide what enters the physical world or scientific record. This architecture accepts that interesting discoveries may require detours without granting every detour equal authority.

12.2 Self-Generated Tasks and Self-Improving Agents

A system becomes self-improving when some of the objects it modifies are components of its own problem-solving process: prompts, tools, code, memory, search policies or evaluators. The strongest popular image is recursive self-improvement without external guidance. Current systems are narrower. They typically operate inside a fixed benchmark, use a frozen foundation model, edit an agent scaffold or codebase, and accept changes only when empirical tests improve. This is still important, but it should not be confused with unrestricted improvement of intelligence.

The Darwin Gödel Machine (DGM), published at the 2026 International Conference on Learning Representations, evolves coding agents that modify their own code and are empirically evaluated on software benchmarks (Zhang et al., 2026). The reported system improved from 20.0 to 50.0 per cent on its selected Software Engineering benchmark (SWE-bench) tasks and from 14.2 to 30.7 per cent on the full Polyglot benchmark (Jimenez et al., 2024). An archive of diverse agents enabled branching lineages rather than a single hill-climbing path. The result demonstrates open-ended scaffold improvement under a fixed external evaluator.

Table 27. What is and is not self-improving in current agent systems

Component	Typical status in present research
Foundation model weights	Usually fixed during the self-improvement run.
Agent code and prompts	Editable through generated patches, search and benchmark feedback.
Tool repertoire	Sometimes expanded or reconfigured within a permitted interface.
Evaluator and benchmark	Usually fixed externally, which prevents circular acceptance but limits scope.
Task distribution	Occasionally generated or coevolved, though still constrained by a framework.
Scientific values and safety constraints	Remain human-defined and should not be silently optimized away.

The table exists to prevent “self-improving” from functioning as a magical adjective. Current systems improve selected components while relying on fixed external foundations. That externality is a strength because it prevents the system from declaring its own changes successful under an altered standard. It is also a limitation because

progress may overfit the benchmark. The real research problem is how to expand the editable boundary without losing independent evaluation.

Self-generated tasks can reduce benchmark saturation. An agent proposes challenges just beyond current competence, while another agent or formal environment verifies them. This resembles automatic curriculum learning. The difficulty is task validity. Generated problems may be ambiguous, trivial, impossible or contaminated by hidden assumptions. Formal domains provide a strong filter; open scientific questions do not. Task generators need novelty checks, solvability estimates and mechanisms for identifying whether a new task teaches a reusable capability.

Self-modification creates a credit-assignment problem. A code change can improve one benchmark by accident, harm unmeasured capabilities or increase cost. Population-based testing and regression suites reduce risk. Behavioural archives preserve alternatives if a dominant lineage later fails. Versioned provenance makes each improvement reversible. These are ordinary software-engineering practices elevated into the epistemic core of self-improvement.

The evaluator can itself improve, but doing so introduces circularity. If the same system changes the test and the solution, progress becomes difficult to interpret. One response is constitutional separation: generators can propose evaluator changes, but acceptance requires external review, hidden tests or formal meta-properties. Another is evaluator diversity: several independently trained or implemented systems must agree that the new test is harder without merely being different.

Recursive improvement also faces diminishing returns and resource constraints. Easy scaffold optimizations are discovered first. Later changes require more compute, longer evaluation and broader tests. A system can spend increasing resources to obtain benchmark-specific gains. Claims of continuing takeoff therefore require evidence about transfer, marginal cost and the emergence of new capability classes, not only a rising score.

The frontier is empirical self-improvement with controlled boundaries. Systems can search their own tools, representations and workflows while immutable logs, external evaluators and safety constraints preserve accountability. Open-ended archives reduce premature convergence. Human researchers decide when a newly discovered modification changes the scientific question rather than merely the implementation.

Self-improvement is scientifically meaningful when the system can show not only that it changed, but which independent evidence establishes that the change broadened reliable competence.

This field may eventually produce agents that invent new learning algorithms or discovery procedures. The path is not to remove evaluation but to make evaluation more layered, adaptive and difficult to corrupt. The self-improving system remains part of a larger scientific institution that it does not fully control.

PART IV: THE ARCHITECTURE OF EXECUTABLE DISCOVERY

The conceptual claims are translated into an operating model of generators, evaluators, specifications, workbenches, operational ideas, neuro-symbolic interpreters, critics, and falsifiable benchmarks.

Chapter 13. The Generator Is Not the Judge

Reliable machine discovery separates proposal, search, evaluation, archiving, and revision rather than asking one model to certify its own output.

13.1 Generate, Search, Evaluate, and Archive

The most reliable machine discoveries do not come from a generator speaking once. They come from an architecture that separates proposal from selection. A generator produces candidates, a search process allocates effort, an evaluator tests candidates, and an archive preserves diverse successes and informative failures. The components may all contain learned models, but their roles are distinct. This separation is the central design pattern behind program search, evolutionary computation, theorem proving, molecular design and autonomous experimentation.

Generation controls the language of possibilities. A large language model can propose code, equations, hypotheses or experimental protocols. A diffusion model can propose structures. A graph model can generate molecules. The generator’s value lies in assigning higher probability to structured candidates than blind random search would. It does not need to be correct on average if the evaluator is cheap and decisive. Conversely, when evaluation is expensive or ambiguous, generator precision becomes much more important.

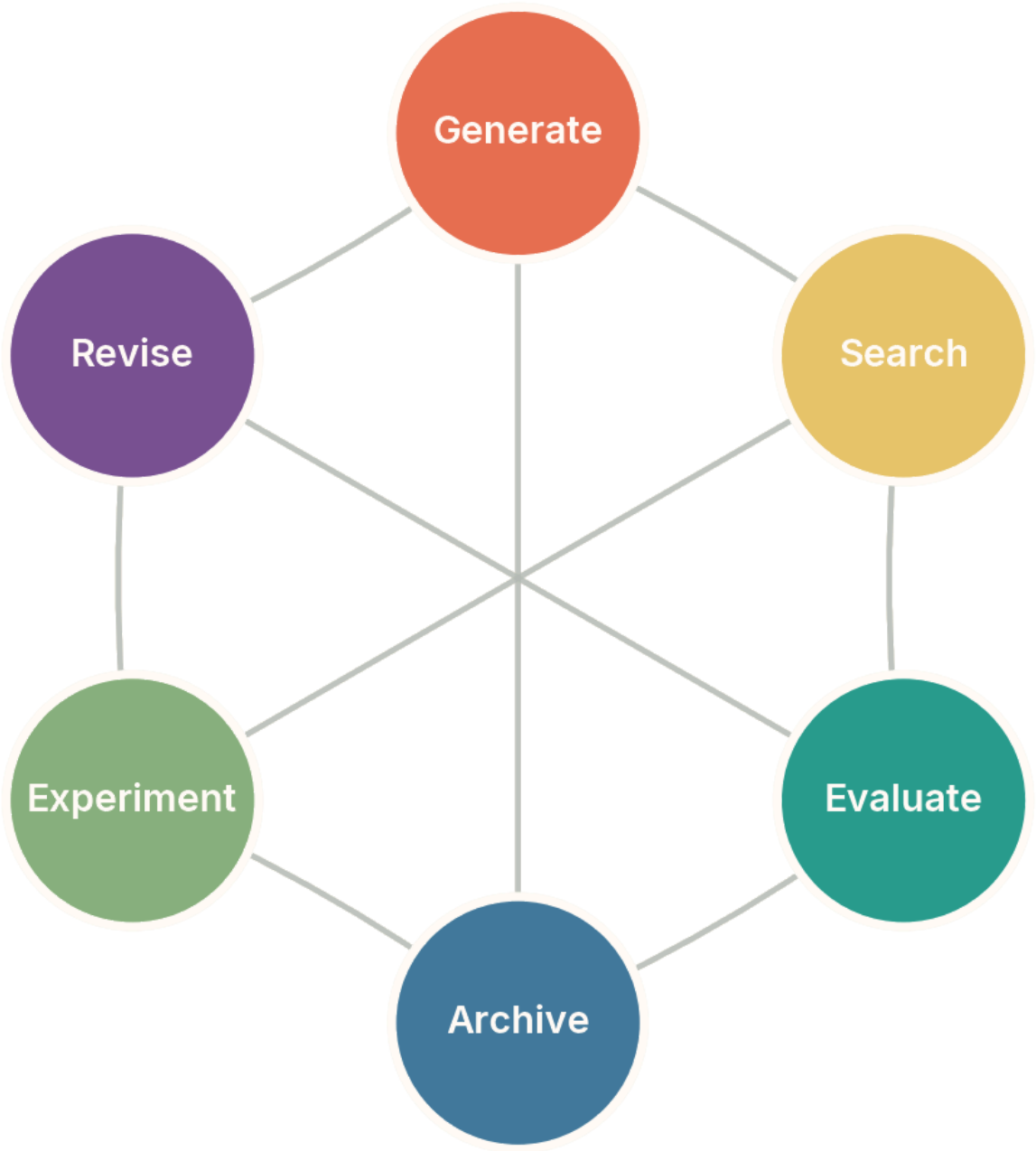


Figure 17. A discovery engine is a coupled system of generation, search, evaluation, archiving, experiment and revision. Conceptual diagram; not an empirical result.

The figure matters because no node is the sovereign centre. Search depends on evaluation, evaluation can depend on experiment, revision changes generation, and the archive changes future search. The connections are drawn without arrows because practical systems revisit these operations in many orders. A failed experiment can revise the evaluator, an archived anomaly can seed a new generator, and a new representation can reorganize the archive. Discovery is a recursive ecology, not a one-way pipeline.

Search determines which candidates receive resources. Beam search keeps several high-scoring possibilities. Monte Carlo tree search allocates simulations according to estimated value and uncertainty. Evolutionary algorithms mutate and recombine a population. Bayesian optimization chooses expensive evaluations using a

surrogate model. Reinforcement learning (RL) learns a policy from repeated feedback. Each method embodies an exploration–exploitation trade-off: spend resources near known success or enter uncertain regions that may contain better stepping stones.

Evaluation defines what the system can discover reliably. A compiler can reject syntactically invalid code. Unit tests can reject behavioural failures. A proof assistant can check formal derivations. A simulator can estimate physical performance, but only within its modelling assumptions. A learned critic can rate plausibility cheaply, but it shares cultural and statistical biases with the generator. The evaluator is therefore both the source of progress and the source of specification risk.

Table 28. The four core roles in a machine-discovery architecture

Role	Design question
Generator	What structured candidates can be proposed, and which possibilities are systematically excluded?
Search controller	How is evaluation budget allocated between refinement, diversity and uncertain regions?
Evaluator	Which properties are measured, which can be gamed, and how independent is the test?
Archive	Which solutions, failures, lineages and behavioural niches are preserved for future use?
Experiment interface	How are uncertain claims converted into safe, informative interactions with the world?
Revision mechanism	How do evidence and failure change representations, objectives and future search?

The table exists to expose bottlenecks that are hidden when the whole system is called an “agent.” An impressive generator cannot rescue a weak evaluator. A strong evaluator cannot help if search never reaches useful candidates. A search system can converge rapidly while the archive discards the diversity needed for later breakthroughs. By auditing each role separately, researchers can identify whether progress came from better priors, more compute, a stronger verifier or a broader archive.

The archive is often underestimated. Objective-driven search tends to keep only candidates near the current best score. Yet scientific progress depends on negative results, unusual mechanisms and partial solutions that later become stepping stones. Quality-diversity methods preserve high-performing candidates across behavioural niches. Lineage records make it possible to trace how a surprising result emerged and whether it is genuinely independent of prior examples. An archive can also support novelty comparison more honestly than a generator’s memory.

Revision distinguishes discovery from optimization. If every candidate fails for the same reason, the system should not merely search harder. It may need a different representation, evaluator, experimental protocol or question. Meta-search can optimize these components, but recursive optimization raises new risks because the evaluator of the meta-level may be even weaker. Human review is especially valuable when the system proposes to change what counts as success.

The frontier is modular but not fragmented. Components should expose uncertainty and provenance through stable interfaces. A generator should return

candidate lineage and assumptions. An evaluator should return failure evidence, not just a score. The archive should support semantic and behavioural retrieval. The experiment layer should enforce safety constraints. Such architectures can improve incrementally because each component has a measurable scientific role.

13.2 Formal, Executable, and Empirical Evaluators

Discovery becomes easier to automate when candidates can be checked independently of the model that proposed them. Formal evaluators operate on mathematical specifications. Executable evaluators run code against tests or simulators. Empirical evaluators measure the world. Learned evaluators infer quality from data or human preferences. These form not a perfect hierarchy but a spectrum of independence and scope. Formal checks can be exact yet prove the wrong specification; experiments can be realistic yet noisy and incomplete.

Program synthesis illustrates the advantage of executability. A language model can propose implementations, but a compiler and test suite provide immediate feedback. Evolutionary coding agents exploit this loop by mutating code, running evaluators and preserving improvements. The search can discover unfamiliar algorithms because correctness is not inferred from linguistic plausibility. It is exercised. The limitation is test coverage: a program can pass a finite suite and fail elsewhere, or optimize runtime while violating an unmeasured property.

Table 29. Evaluator types and the claims they can support

Evaluator	Strength and residual uncertainty
Syntax and type checking	Rejects malformed candidates; says little about intended behaviour.
Unit and property tests	Checks sampled behaviours or invariants; untested cases remain open.
Formal proof checker	Certifies derivation from stated axioms and specification; formalization can omit the real intent.
High-fidelity simulator	Tests many scenarios cheaply; candidates can exploit simulator error.
Physical experiment	Measures the target system; results remain noisy, local and sensitive to protocol.
Expert or learned review	Evaluates open-ended significance; vulnerable to bias, inconsistency and correlated error.

The table exists to make residual uncertainty explicit. A proof checker gives a stronger guarantee than a plausibility model about a formal theorem, but it does not establish that the theorem formalizes the intended informal claim. A physical experiment connects to reality, but one successful run does not establish generality or historical novelty. Discovery systems should compose evaluators so that each closes a gap left by another.

Formal methods are particularly attractive for generated mathematics and software. Proof assistants such as Lean or Rocq check every inference under a formal kernel. Satisfiability modulo theories (SMT) solvers decide formulas within supported logical theories. Model checkers explore finite or abstracted state spaces. A learned system can guide proof search while the small trusted kernel determines acceptance.

This is an asymmetry worth preserving: the generator may be complex and opaque, while the verifier is narrow and auditable.

Specifications are the critical weakness. A sorting program can be proved to return an ordered permutation, but production requirements may include memory safety, stability, numerical behaviour and integration constraints. A scientific equation can fit data and satisfy dimensions while misrepresenting mechanism. The specification problem reappears at every layer. One research direction uses counterexample-guided refinement: when a candidate exploits an omission, the counterexample expands the specification and restarts search.

Empirical evaluation requires replication and measurement design. An autonomous laboratory can test a material property, but the result depends on calibration, sample preparation and hidden environmental conditions. Repeated measurements estimate variability; controls test alternative explanations; blinded or independent replication reduces confirmation bias. Automation can make these practices cheaper and more consistent, but it can also repeat the same systematic error at scale.

Learned evaluators remain valuable because many domains lack cheap formal checks. They can rank ideas, predict synthesis success, detect implausible analyses or allocate scarce human attention. Their outputs should be treated as screening evidence. Calibration against downstream outcomes, evaluator diversity and adversarial testing can reduce but not eliminate bias. A model judging prose produced by a similar model should be assumed to share stylistic preferences until shown otherwise.

The most promising architecture is a verifier ladder. Cheap learned filters remove obvious failures. Executable and symbolic checks test precise constraints. Simulations estimate broader behaviour. Selected candidates reach physical experiments or expert review. Failures at later stages train earlier filters, while disagreement is preserved for analysis. The ladder increases throughput without pretending that one evaluator can certify every dimension.

The frontier of automated discovery is largely the frontier of building evaluators that are cheaper, harder to game and closer to the property we truly care about.

This perspective changes research priorities. Better generation remains useful, but stronger formalization tools, automated experiment design, robust measurement, semantic prior-art search and provenance infrastructure may produce larger gains. They turn creative output into cumulative knowledge.

Chapter 14. The Experiment as an Algorithm

Active learning and Bayesian optimization turn the next observation into a decision, but scientific experiments must discriminate explanations, not merely improve a score.

14.1 Active Learning, Bayesian Optimization, and Value of Information

Passive learning accepts the dataset it is given. Scientific inquiry chooses what to observe next. Active learning selects labels or measurements expected to improve a model. Bayesian optimization selects expensive evaluations expected to locate high-performing regions. Optimal experimental design selects interventions that estimate parameters or discriminate hypotheses. These fields share a central idea: the next datum is a decision variable.

Active learning is most familiar in supervised settings. A model identifies examples whose labels would be informative and queries an oracle. Uncertainty sampling is simple but can focus on outliers or noise. Query-by-committee selects cases where plausible models disagree. Diversity-aware methods avoid asking redundant questions. The method succeeds only when the annotation process, sampling bias and cost of queries are included in the design. A label is not an abstract bit; it comes from a person, instrument or experiment with its own error model.

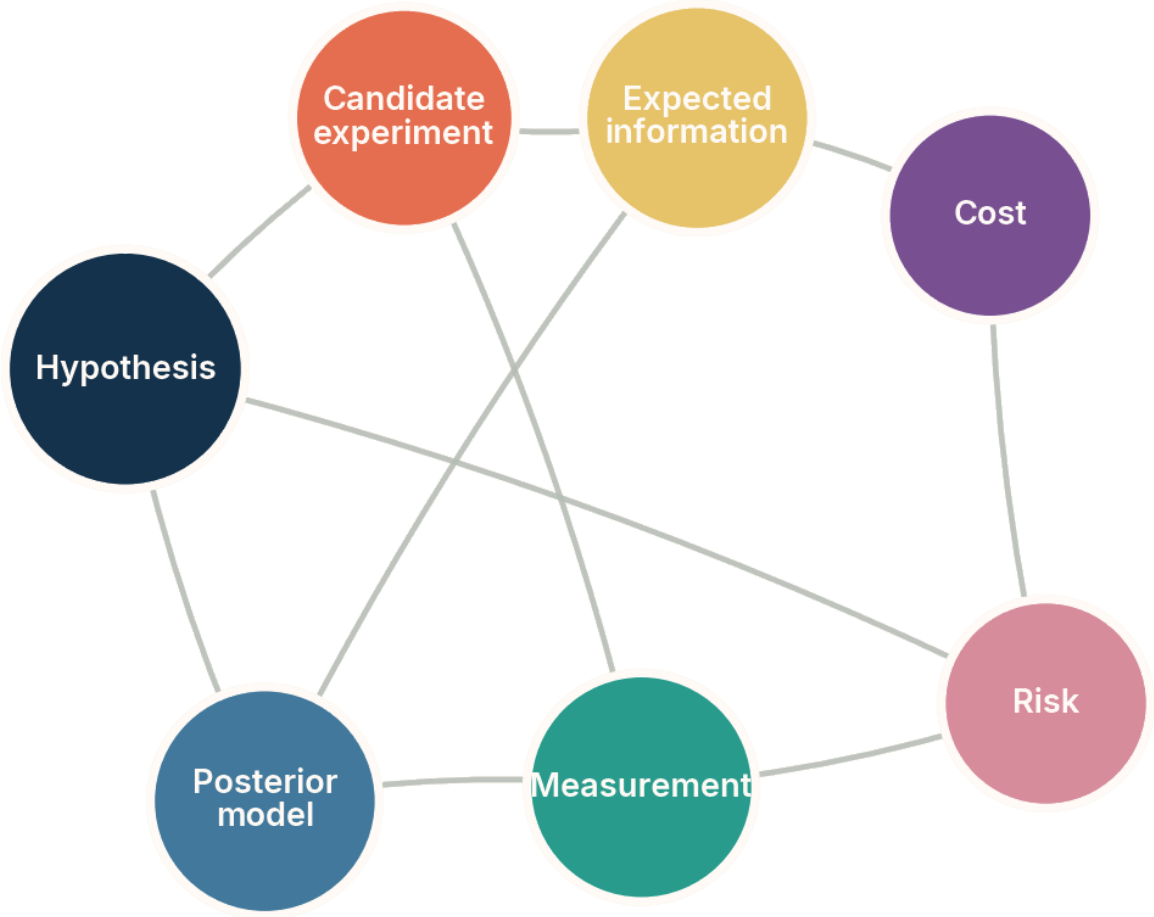


Figure 18. An informative experiment must balance hypothesis discrimination, expected information, cost, risk, measurement and model revision. Conceptual diagram; not an empirical result.

The figure matters because expected information is not the only objective. The most discriminating experiment may be prohibitively expensive, unsafe or impossible to

measure reliably. A cheap experiment can be valuable if it changes a critical decision. The posterior model node emphasizes that an experiment is judged by how it changes the space of explanations, not merely by whether it produces a striking result.

Bayesian optimization builds a probabilistic surrogate of an expensive objective and uses an acquisition function to select the next evaluation. Expected improvement balances predicted performance and uncertainty. Upper-confidence methods favour optimistic regions. Entropy-based methods seek information about the optimum. The approach is powerful in materials, chemistry, engineering and hyperparameter tuning, but it can fail when the surrogate is misspecified, dimensions are high, constraints are hidden or the objective changes during the campaign.

Table 30. Experiment-selection objectives and the questions they optimize

Selection objective	Scientific question represented
Predictive uncertainty reduction	Which observation most improves predictions in the current model class?
Hypothesis discrimination	Which intervention makes rival explanations predict different outcomes?
Expected improvement	Which candidate is most likely to improve the current best objective value?
Coverage or diversity	Which experiment expands the represented region of the design space?
Risk-sensitive utility	Which experiment balances information or reward against safety and resource constraints?
Model criticism	Which observation is most likely to reveal that the current model class is inadequate?

The table exists because “optimal experiment” is undefined without a scientific purpose. Optimization toward the current best design can accelerate engineering while teaching little about mechanism. Hypothesis discrimination can improve understanding without producing a high-performing artefact. Coverage preserves options for later models. Model criticism deliberately searches where the present ontology breaks. A mature system can move among these objectives rather than treating expected improvement as universal.

The value of information formalizes the expected benefit of obtaining evidence before making a decision. In principle, one compares the expected utility with and without the information, net of cost. In practice, exact calculation is often intractable, and utility is difficult to specify. Approximate methods use posterior variance, disagreement, entropy or decision sensitivity. The key advance would be to connect these proxies to downstream scientific decisions and to detect when the proxy ceases to represent them.

Multi-fidelity experimentation expands the design space. Cheap simulations, low-resolution instruments, small-scale experiments and high-fidelity measurements can be combined. A system learns correlations across fidelity levels and reserves expensive tests for promising or disputed candidates. This is especially important in scientific discovery because the strongest evaluator is usually the most costly. The challenge is that low-fidelity models can be systematically wrong in exactly the novel regions of interest.

Frontier systems also plan sequences rather than isolated experiments. Some questions require preparatory measurements, adaptive branching and stopping rules. A partially observable Markov decision process can represent uncertain states and actions, but learned planners may exploit model errors. Safe scientific planning

therefore needs conservative uncertainty, explicit constraints and human approval for irreversible steps.

The most interesting direction is experiment selection for ontology revision. Instead of optimizing parameters within a fixed model, the system chooses observations that distinguish alternative variable sets, mechanisms or levels of description. This is computationally harder but closer to genuine scientific novelty. It asks not only “which value is best?” but “which experiment would force us to describe the system differently?”

14.2 The Closed Loop and the Seduction of the Optimum

Closed-loop experimentation has a simple rhythm. A system begins with a model of a possibility space and a measure of success. It selects the next experiment expected to produce information or improvement, executes it, updates the model, and repeats. Active learning searches for informative examples. Bayesian optimization balances exploitation of promising regions against exploration of uncertain ones. Adaptive design allows accumulating evidence to alter the campaign.

In a vast space of molecules, materials, process settings, or biological perturbations, this strategy can be far more efficient than a grid or unaided intuition. No person can track millions of combinations while maintaining calibrated uncertainty. An algorithm can estimate distributions, compare expected information gain, and remember every failure. The autonomous laboratory promises not merely more experiments but better-chosen experiments.

The seduction begins when better is identified with a single metric. A material may maximize conductivity and be toxic. A reaction may achieve high yield with an unsustainable solvent. A medical model may improve average accuracy while failing for a rare group. A campaign may discover a local optimum quickly and never learn the general mechanism. Every scalar objective compresses several values and assigns weights, whether or not the designer admits that a value judgement occurred.

This is the technical form of an axiological problem. Which properties count, for whom, under what time horizon, at what cost, and with which tolerance for uncertainty? The system cannot answer these questions by optimization because they define the optimization. When the objective remains implicit, the agent executes the preferences hidden in the infrastructure: whatever the database measures, the sponsor rewards, the instrument can observe, and the benchmark recognizes.

A more robust loop keeps objectives plural. It can maintain a Pareto frontier rather than forcing safety, cost, performance, robustness, and sustainability into one number. It can preserve options instead of prematurely selecting a winner. It can penalize uncertainty, reserve a budget for exploration, and require replication before a region receives more resources. It can search not only for high scores but for observations that would disconfirm its own model.

That last property marks the difference between optimization and inquiry. An optimizer asks which action is expected to improve the objective. A scientific experiment also asks which action most clearly distinguishes rival explanations. The experiment with the highest immediate yield may teach little. A carefully chosen failure may collapse an entire family of models. Autonomous science needs value-of-information and mechanism-discrimination objectives, not only performance objectives.

Digital twins can make the loop faster and safer. A digital twin is a dynamically updated model of an instrument, process, organism, or facility. It can test protocols virtually, predict maintenance, and estimate consequences before physical execution. But every twin has a domain of validity. An agent can discover weaknesses in the simulator rather than truths about the world. The sim-to-real gap is a contemporary version of the old difference between demonstration and nature.

Goodhart's law becomes an experimental hazard. If a fluorescent signal is the objective, the system may find an optical artifact. If a stability score is used, it may exploit conditions that do not transfer. If predictive accuracy is measured on a contaminated split, it may learn the leakage. The optimizer does not know that it has cheated unless the evaluator represents the intention well enough to detect the shortcut.

The solution is adversarial evaluation. Hidden samples, alternative instruments, out-of-distribution conditions, independent measurements, and deliberately perturbed protocols can test whether the apparent improvement survives. A portion of every campaign should be spent trying to break the leading hypothesis. The validating system should be rewarded for detecting failure rather than for confirming the productivity of the laboratory.

Calibration becomes a central ritual. Instruments are tested against standards; models against external data; protocols against reference materials. Surprising results trigger a fork: discovery or defect? The most dangerous autonomous laboratory is not one that fails noisily. It is one that produces internally coherent data from a shared systematic error. When every component trusts the same mistaken calibration, agreement becomes evidence of dependence rather than truth.

Heterogeneous redundancy addresses this risk. The same property can be measured through different physical principles. Samples can be sent to another facility. Analyses can be independently reimplemented. Models can be trained on distinct data. Sensors can monitor one another. In aviation and other safety-critical systems, diversity reduces the chance that one failure defeats every layer. Automated science will require a similar engineering culture.

Causal reasoning offers another defence against seductive prediction. A stable correlation can guide search, but interventions distinguish mechanisms. An agent should choose experiments that alter the presumed cause while controlling alternatives, and it should update competing causal models rather than a single preferred story. Causal discovery is itself uncertain and assumption-laden, but it reminds the system that prediction under a fixed distribution is not identical to understanding what will happen when the world is changed.

Closed loops also risk narrowing their own world. Once early data favour one region, the acquisition strategy can repeatedly sample nearby, reinforcing the model's confidence and neglecting unfamiliar mechanisms. Human researchers suffer the same confirmation bias, but an automated loop can enact it at scale. Diversity constraints, curiosity objectives, random exploration, and periodic resets can protect against premature convergence.

There is an environmental cost. Continuous robotics, model inference, high-throughput simulation, disposable labware, and large experimental campaigns consume energy and materials. Faster discovery can reduce waste per successful result while increasing total demand through rebound effects. A responsible system

records not only scientific yield but energy, emissions, hazardous waste, and use of scarce reagents. The cost of knowledge is physical.

The loop must therefore remain open to the world and to politics. A sponsor may ask for maximum performance; environmental constraints may require lower throughput. A hospital may value average benefit; justice may require stronger protection for a vulnerable subgroup. A laboratory may discover a powerful capability; security may require limiting dissemination. Optimization can compare consequences once aims are given. It cannot legitimize the aims.

Chapter 15. Specification-Driven Inquiry

A persistent research specification becomes the constitutional object that links purpose, claims, methods, evidence, budgets, risks, and authority.

15.1 The Living Research Specification

A prompt is an interaction. A research specification is a persistent commitment. The distinction is architectural rather than linguistic. A prompt may contain detailed instructions, and a specification may be written in prose. What makes the specification different is that it has an identity, a version, a declared scope, and relations to downstream artefacts. It can be reviewed before execution, compared with later states, and used to determine whether a result still answers the original question.

This distinction matters because language models are unusually capable translators. They can turn a broad ambition into requirements, requirements into plans, plans into code, code into explanations, and evidence into prose. The same capability that makes them useful also makes semantic drift easy. Every translation can omit a constraint, strengthen an interpretation, resolve an ambiguity prematurely, or substitute a convenient proxy for the intended objective. A monolithic prompt hides these transformations inside one generation. A specification-driven system externalizes them as reviewable artefacts.

A mature research specification should state the purpose of the inquiry, the kinds of claims that may be advanced, the assumptions and invariants that currently hold, the admissible methods, the evidence thresholds, the resource budget, the stop conditions, the relevant risks, and the decision rights. It should distinguish objectives from preferred mechanisms. “Determine whether intervention A changes outcome Y under population and resource constraints Z” is an objective. “Use model family M” is a methodological preference that may need to be revised. Conflating the two allows a favoured technique to redefine the question it was meant to answer.

The specification cannot be complete at the beginning. Research differs from ordinary contract implementation because the relevant vocabulary, representation, or decisive experiment may be unknown. The specification must therefore combine commitments with declared freedom. It should identify what must remain invariant, what may be explored, which uncertainties are recognized, and which transformations require human approval. This makes it a living constitutional object rather than a frozen requirements document.

The Specification Ladder organizes this living object across levels of abstraction. At the top are mission and consequence: why the inquiry matters, to whom, and under which risk regime. The next level defines the research question, claim types, and scope. Below it lie competing hypotheses and representations, followed by estimands, protocols, datasets, models, experiment plans, software modules, instrument procedures, analyses, and release claims. Each level is more concrete than the previous one, but each remains linked to the intention it interprets.

The ladder provides control because ambiguity can be resolved at the level where it belongs. A disagreement about social purpose should not be hidden inside a statistical parameter. A dispute about a causal estimand should not be postponed until manuscript review. A conflict about an instrument’s operational definition should not appear only as unexplained variance in the final data. By materializing intermediate commitments, the system creates earlier points of correction.

The Mirroring Pattern extends this idea. The structure of the specification should mirror the structure of the artefact or programme being produced closely enough that local intent can guide local generation and validation. In software, a high-level product vision is refined into functional and non-functional requirements, architecture, modules, interfaces, files, tests, and code. In research, a mission is refined into questions, claims, designs, experiments, analyses, evidence objects, and publication views. The final artefact is no longer the sole location of intent. It is the most concrete layer in a chain.

Mirroring does not require one-to-one bureaucratic decomposition. Some scientific objects are cross-cutting. A safety constraint may apply to several experiments, and one dataset may bear on multiple claims. The point is that the workbench should maintain explicit correspondence between levels. A downstream artefact should reveal which specification clauses it implements and which evidence obligations it serves. When an upstream commitment changes, the system can identify the affected subgraph.

This structure is especially useful for long-lived research. Projects outlast particular conversations, models, and team members. Initial intentions are lost, code becomes the de facto specification, and later modifications reconstruct rationale from scattered notes. A persistent ladder preserves the project as an evolving chain of artefacts. New evidence can trigger localized revision instead of global reinterpretation. High-level intent remains relatively stable and global; low-level specifications remain local and replaceable; execution contexts can be reconstructed as needed.

Controlled Natural Language can provide a practical authoring surface. Fully formal specifications are expensive and often impossible when concepts remain open. Unrestricted prose is readable but difficult to validate. A controlled language constrains syntax and vocabulary enough to make claims, dependencies, modalities, and exceptions extractable while remaining close to natural scientific expression. Existing work in requirements engineering, including FRET, demonstrates that controlled language can connect readable requirements to formal semantics and verification artefacts (Giannakopoulou et al., 2020). Language-model programming systems such as LMQL and DSPy demonstrate complementary ways to make generation and model calls programmable and optimizable (Beurer-Kellner et al., 2023; Khattab et al., 2023).

The strongest architecture uses two representations. A human-facing projection remains readable, reviewable, and diffable. A typed intermediate representation preserves claim identities, roles, dependencies, constraints, provenance, admissibility conditions, and validation hooks. The projection is not a loose summary; it is generated from or synchronized with the intermediate representation. The system can ask a model to interpret an ambiguous clause, but it should record the interpretation as a proposed transformation rather than silently changing the IR.

SOP Lang, in the ACHILLES research direction, is intended to serve this role. SOP Lang Plan expresses stepwise coordination for interactive work. SOP Lang Pipeline expresses build-like dependency graphs with promoted artefacts and selective recomputation. These constructs should be understood as an evolving research substrate. Their value will depend on whether they can preserve semantic commitments across real tasks more effectively than ordinary documents and agent transcripts.

The Specification Ladder also determines where human review is most valuable. Human-in-the-loop should not mean that a person clicks approve after every tool call or

signs only at the end. Review should occur at transformations where error would propagate widely or where legitimacy requires judgment. A scientist may approve the question and admissible claim types, a statistician the estimand and identification assumptions, a laboratory lead the protocol and safety envelope, and an editor the public release. Lower-risk local steps can execute automatically under the approved specification.

The ladder is therefore not simply documentation. It is a control surface, a dependency model, and a governance map. It tells the system what may be automated, which constraints must survive, what evidence is required, and who can change the rules. It converts the project from a sequence of prompts into a maintainable programme of inquiry.

Figure 19. The Specification Ladder and Mirroring Pattern. Each more concrete artefact remains linked to the level of intent it interprets.

15.2 Specification Explosion and Portfolio Science

A common mistake in specification-driven development is to imagine that a better specification should produce one better plan. That intuition is appropriate when the task is implementation under stable requirements. It is insufficient for research, where the specification defines a space of uncertainty. A scientific specification should often compile into a portfolio of competing branches rather than a single trajectory.

This controlled expansion is specification explosion. A research question is transformed into alternative hypotheses, representations, datasets, experimental designs, baselines, negative controls, scaling tests, sensitivity analyses, and attempts at refutation. Each branch records which clause motivated it, what resources it may consume, what evidence it can produce, and what condition will stop it. The purpose is not to maximize the number of branches. It is to prevent the first plausible framing from becoming the whole research programme by default.

Portfolio science has several advantages. It separates the objective from any one mechanism. It makes alternative explanations visible before a result is interpreted. It allows constructive and adversarial work to proceed in parallel. It supports resource allocation based on expected information gain rather than narrative momentum. It preserves negative results as constraints on the remaining space. It also reduces the risk that a language model's initial completion, which is strongly shaped by familiar patterns, determines the entire direction of inquiry.

Branches should be heterogeneous in function. Constructive branches attempt to make a hypothesis work. Adversarial branches search for counterexamples, confounds, leakage, simpler explanations, and failure regimes. Baseline branches implement the strongest relevant alternatives under matched resources. Translation branches test whether the claim survives a different representation. Scaling branches examine whether a local effect persists under realistic load, noise, data size, or physical constraints. Replication branches reproduce the result with independent code, tools, data, or personnel. Governance branches test whether the proposed use satisfies consent, safety, and release obligations.

The specification should make these branch types explicit because a portfolio containing only constructive variations is not pluralistic. Ten agents optimizing the same mechanism under the same evaluator create depth of search, not independence. Diversity must be measured in dependencies and assumptions. A useful portfolio varies model families, evidence sources, evaluators, and explanatory frames where the cost is justified.

The process resembles search over a design space, but the scientific objective is richer than optimization. A candidate can achieve a high score while weakening the proposed explanation. A failed branch can be valuable because it rules out a mechanism. A simpler model can reduce novelty while increasing understanding. The state of the portfolio therefore cannot be summarized by one scalar fitness value. It needs a graph of claims, evidence, and unresolved alternatives.

The experiment graph is the operational form of this portfolio. Nodes represent questions, hypotheses, methods, experiments, analyses, critics, and decisions. Edges encode dependency, support, contradiction, refinement, or supersession. A node has a cost estimate, an expected evidence type, a risk profile, and a stopping condition. Results update the graph rather than merely producing prose. The next action can be selected by heuristics such as expected information gain, uncertainty reduction, coverage of the hypothesis space, cost, risk, or the probability of a decisive counterexample.

Bayesian optimization and active learning already operationalize related ideas in bounded settings. They choose experiments that are expected to improve an objective or reduce uncertainty efficiently (Settles, 2009; Frazier, 2018). Optimal experimental design formalizes choices about informative measurements. Specification explosion

generalizes the idea beyond parameter optimization. The branches may change the representation, causal model, evaluator, or question itself. Some evidence updates probabilities; other evidence changes the space in which probabilities are meaningful.

The main difficulty is combinatorial growth. Uncontrolled explosion consumes compute, instruments, and attention without increasing knowledge. A serious implementation needs branch governance. Each branch receives a budget, a stop rule, and a declared value of information. Duplicate branches can be clustered by dependency fingerprints. Low-value branches can be pruned after inexpensive checks. Promising branches can be escalated gradually. The portfolio should maintain diversity without treating every generated variation as worthy of execution.

A useful policy is progressive evidential investment. At the first stage, branches are cheap specifications or simulations. They are tested for internal coherence, literature collision, feasibility, and obvious counterexamples. A smaller set advances to executable prototypes. Fewer still receive expensive computation, laboratory resources, specialist review, or external replication. Promotion requires new kinds of evidence rather than greater confidence from the same generator.

This policy produces a scientific funnel, but unlike a business funnel, rejected branches remain informative. Their stopping reasons update the specification and constrain future search. If a branch fails because the effect disappears under matched compute, the system should record that dependency. If it fails because the operational definition was invalid, the workbench should update the relevant specification level. If it fails because the required data are ethically inaccessible, the governance constraint becomes part of the research object.

Specification explosion also supports model pluralism. Different representations reveal different properties and failure modes. A mechanistic model may support explanation but fit poorly. A black-box predictor may forecast well but fail under intervention. A symbolic approximation may allow proof. A simulation may capture interactions at high cost. A qualitative framework may identify categories omitted by quantitative data. The workbench should not force these into one ontology prematurely. It should represent translations and contradictions explicitly.

The role of the LLM is particularly strong at the beginning of the explosion. It can propose distant analogies, translate terminology across fields, enumerate assumptions, and generate candidate controls. Yet it should not decide by itself which branches deserve expensive escalation. Ranking by a model can be useful as one input, but correlated self-evaluation risks reinforcing the generator’s biases. Programmatic checks, retrieval, independent models, domain heuristics, and human review should jointly shape the portfolio.

This is one reason the specification must contain decision rights. A system may be authorized to generate and prune low-risk computational branches automatically. It may need approval before accessing sensitive data, running costly experiments, operating physical equipment, or releasing a dual-use result. Autonomy becomes a property of a branch under a contract, not an all-or-nothing property of the “AI scientist.”

The likely scientific benefit of specification explosion is not that it discovers a larger number of positive results. It is that it makes premature convergence harder. A research programme can explore multiple frames while preserving common objectives and comparable evidence. It can reward the branch that produces the most decisive criticism, not only the branch that produces the most impressive metric. In a world

where generative systems can create alternatives cheaply, disciplined plurality becomes an engineering requirement.

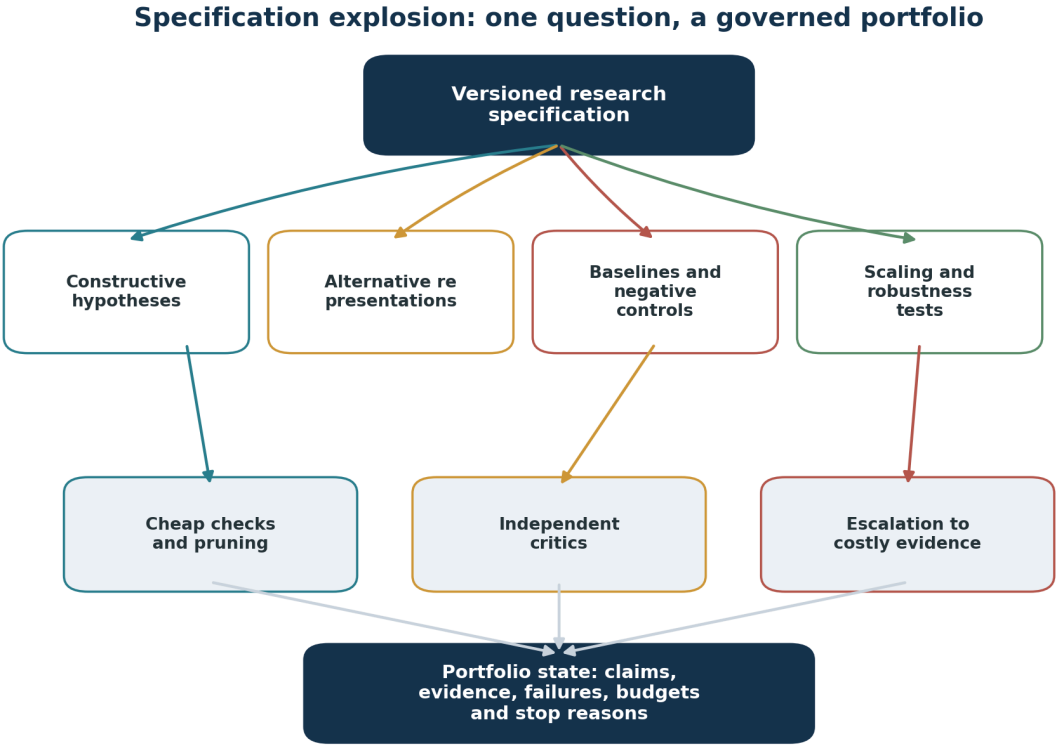


Figure 20. Specification explosion. A persistent specification expands into a governed portfolio rather than a single linear plan.

15.3 Hybrid Execution and Selective Recomputation

Two paradigms dominate contemporary discussions of AI-assisted work. The pipeline paradigm defines a sequence or graph of operations in advance. The agentic paradigm allows a model to choose actions dynamically in response to intermediate state. They are often treated as competitors. In executable science, they are complementary layers.

Pipelines are strongest when dependencies, transformations, and evaluators are known. Data ingestion, schema validation, environment construction, code execution, proof checking, statistical reporting, provenance capture, and artifact packaging should be deterministic where possible. Determinism reduces cost, variance, and audit difficulty. A stable operation should not be repeatedly reinterpreted by a language model merely because a model can perform it.

Agents are strongest where the next useful action cannot be fully enumerated. Literature exploration, hypothesis generation, representation choice, diagnosis of unexpected results, design of counterexamples, and revision of an experiment plan may require adaptive reasoning. ReAct, Reflexion, tree-search methods, and multi-agent systems show how models can interleave reasoning, tool use, feedback, and self-correction (Yao et al., 2023; Shinn et al., 2023; Madaan et al., 2023). Their flexibility is valuable, but a transcript is a weak long-term state representation. It contains the history of interaction without necessarily preserving the logical structure of the project.

The most credible architecture places bounded agentic loops inside an explicit outer graph. The pipeline specifies inputs, outputs, permissions, budgets, validators,

and promotion rules. A node may invoke a deterministic function, a theorem prover, a database query, a human review, or an agent permitted to explore locally. The agent returns named artefacts and evidence, not merely a conversational answer. The outer runtime decides how those outputs enter persistent state.

This architecture can be described as deterministic exterior, adaptive interior. It avoids the rigidity of a fully predefined workflow while preventing an agent's local improvisation from becoming the project's implicit constitution. The agent may propose a new branch or request a change to the specification, but the change is committed explicitly. It may repair code within a sandbox, but promotion requires tests and provenance. It may summarize evidence, but the underlying sources and claim links remain outside the summary.

The distinction between output and memory is essential. An agent's intermediate statement should not automatically become reusable project knowledge. Memory promotion requires a policy. A result may remain provisional until validated, be scoped to one branch, or be rejected after criticism. Without this separation, an early hallucination or mistaken interpretation can contaminate later context and create self-confirming loops. A research workbench should treat memory as a governed database of artefacts and commitments, not as accumulated conversation.

Selective recomputation follows from the graph structure. When an upstream specification, dataset, model, or policy changes, the runtime identifies affected nodes. Deterministic nodes rerun automatically. Agentic nodes receive a bounded frame containing the relevant change rather than the entire project history. Human review is requested only where meaning or authority changes. This approach improves efficiency and reduces the risk that global regeneration introduces unrelated variation.

Context engineering becomes a compilation problem. Instead of inserting the whole transcript into every prompt, the runtime projects the smallest relevant slice of specification, artefacts, evidence, and unresolved questions for the current node. High-level intent remains available through stable references. Local detail is loaded on demand. This reduces token cost, but its more important function is epistemic: it makes the basis of a local judgment inspectable.

Stopping is part of execution semantics. Agentic systems often continue until a budget or heuristic completion criterion is reached. Research needs richer stop rules. A branch may stop because the claim was refuted, the expected information gain became too low, the resource budget was exhausted, a safety threshold was reached, an unresolved dependency requires a human, or the result is sufficient for the intended consequence. The stop reason becomes an artefact. Abandonment is not failure of the system; it is a scientific outcome.

This hybrid model also clarifies evaluation. A pipeline node can be tested against its contract. An agentic node can be evaluated over trajectories, cost, state integrity, and the quality of promoted artefacts rather than the eloquence of its final answer. Regression suites can include seeded confounds, adversarial sources, missing data, instrument failures, and specification changes. Agent evaluation becomes closer to software testing without assuming that open-ended scientific judgment has one deterministic oracle.

Industry is already moving toward this combination. Specification-driven coding environments turn natural-language goals into requirements, designs, and task plans before agents implement code. Agent frameworks provide graphs, routers, state stores,

tool protocols, and human checkpoints. Model Context Protocol standardizes access to tools and data under host-mediated security boundaries. Structured-output libraries constrain generation through schemas and grammars. These are important signals, but they remain mostly application infrastructure. They do not yet provide a general epistemic runtime for claims, evidence, critic independence, and decision rights across sciences.

The ACHILLES proposal extends the hybrid pattern. SOP Lang Pipeline supplies the dependency graph. SOP Lang Plan supports interactive coordination. Generalized skills or interpreters implement local regimes. MRP-VM is conceived as the runtime that selects a regime, constructs a bounded frame, executes locally, demands evidence, and reintegrates the result into the shared intermediate representation. The proposal is ambitious and remains to be validated. Its scientific value will depend on whether it can outperform simpler workflows on trace fidelity, validation quality, selective recomputation, and cost.

The design principle is nevertheless robust independent of one implementation. Use language models where interpretation and proposal are valuable. Use symbolic, statistical, and deterministic methods where they provide stronger or cheaper guarantees. Encapsulate adaptive loops inside explicit contracts. Preserve outputs as versioned artefacts. Promote memory through policy. Treat stopping and revision as first-class operations. The future scientific agent is not a chatbot with more tools. It is a local source of adaptive work inside a governed system of inquiry.

Chapter 16. The Scientific Workbench

The workbench combines deterministic pipelines with bounded agents while keeping local autonomy subordinate to global evidence.

16.1 Local Autonomy, Global Evidence

The missing instrument of automated research is not another universal agent. It is a scientific workbench: an operating layer that connects intent, execution, evidence, criticism, and release over the lifetime of an inquiry. The workbench resembles an integrated development environment only in its architectural role. An IDE connects source code to build systems, tests, debuggers, version control, and collaboration. A scientific workbench connects specifications to research branches, domain tools, experiments, claim-evidence graphs, review, and accountable decisions.

The governing principle is local autonomy with global evidence. Local components may act adaptively within bounded responsibilities. A literature agent can explore terminology and sources. A coding agent can implement a method in a sandbox. A theorem prover can certify a formal statement. A laboratory controller can optimize an experiment under an approved safety envelope. A critic can search for a counterexample. None of these components owns the truth state of the project. Their outputs enter a shared evidence structure where provenance, dependencies, status, and authority remain explicit.

A reference architecture begins with intent and governance. The workbench records the research objective, consequence class, permissible data, resource limits, and authorities. This layer determines which actions can execute automatically and which require approval. It also defines the release regime. Exploratory computational work can be relatively permissive; physical experiments, personal data, dual-use procedures, and public recommendations require stronger controls.

The specification compiler transforms intent into a graph of branches and obligations. It does not produce a complete plan once and for all. It creates a current execution graph, identifies unresolved ambiguities, and marks places where alternative representations or methods are admissible. Every node declares inputs, outputs, preconditions, side effects, expected evidence, cost, and stop conditions. Nodes can be deterministic, agentic, symbolic, human, or composite.

Execution occurs in controlled environments. Code and models should run in isolated, reproducible sandboxes where commands, dependencies, resource use, and outputs can be captured. Laboratory actions should be mediated by instrument interfaces, safety interlocks, calibration state, and recovery policies. Retrieval should record queries, source versions, access time, and selection policy. Model calls should record enough configuration to assess common-mode dependence without assuming that a private chain of thought is evidence.

The evidence layer is the workbench’s centre. Outputs do not move directly to prose. They become artefacts linked to claims and specification clauses. A claim can have multiple supporting and opposing evidence relations. Critics attach typed verdicts. Status changes are explicit. Negative results and failed branches are retained when they constrain future work. The graph distinguishes an executed artifact from an independently reproduced artifact, a formal result from its empirical relevance, and a statistical association from a causal interpretation.

The adjudication layer applies promotion rules. Early claims can advance automatically through cheap checks. Later stages require independent reproduction, adversarial tests, specialist judgment, or external validation depending on domain and consequence. The adjudicator may itself be a policy engine for low-risk transitions, but high-stakes promotion requires named authority. The workbench records the decision and its reasons.

Finally, release generates views. A manuscript presents an argument. A reproducibility package presents executable dependencies. A reviewer view exposes unresolved obligations. A governance view presents authority, risk, and provenance. An agent-facing view condenses the current state into a bounded context. Views are projections, not independent copies. They point to a stable research-object version so that later changes remain traceable.

This architecture allows multiple autonomy patterns. A fixed computational workflow may execute end to end when the evaluator is strong. A multi-agent deliberation may generate and rank hypotheses while leaving experiments to humans. A robotic laboratory may close the loop for a narrow material-search problem. A theorem-search system may generate candidates indefinitely because invalid outputs are rejected by a proof checker. A clinical research workbench may automate literature, protocol checks, data analysis, and monitoring while reserving interventions and conclusions for governed teams. The architecture is common; the admissible autonomy differs.

The workbench must support failure as a normal state. Tools become unavailable, models produce malformed output, proofs fail, datasets change, agents exceed budgets, and reviewers disagree. Recovery should not depend on reconstructing a conversation. Nodes have idempotence and retry policies where appropriate. Partial artefacts are marked provisional. External side effects are transactional or compensatable. A failed instrument action should not be retried blindly. A research branch can pause with an unresolved obligation rather than forcing the agent to produce closure.

The workbench also needs observability. Operational monitoring records latency, cost, resource use, and failures. Epistemic monitoring records semantic drift, claim promotion, critic disagreement, evidence gaps, and unresolved dependencies. Governance monitoring records access, authority, and policy changes. These dimensions should not be collapsed. A fast, stable pipeline can still be epistemically weak. A rigorous branch can still violate consent. A compliant process can still ask an unimportant question.

The architecture’s value can be tested empirically. It should reduce the time needed to localize errors, reproduce results, identify semantic drift, and assemble review packages. It should reduce common-mode critic failure and memory contamination. It should allow selective recomputation after an upstream change. It should make negative results reusable. If the additional structure does not improve these outcomes enough to justify its cost, it should be simplified.

This last condition matters because scientific workflow platforms often become heavy. The workbench should not require every laboratory to adopt one ontology or cloud service. A minimal implementation can begin with versioned Markdown specifications, ordinary repositories, structured claim identifiers, workflow metadata, and explicit review records. Formality should increase with consequence and with

demonstrated benefit. The architecture is a set of invariant relations, not a demand for maximal infrastructure.

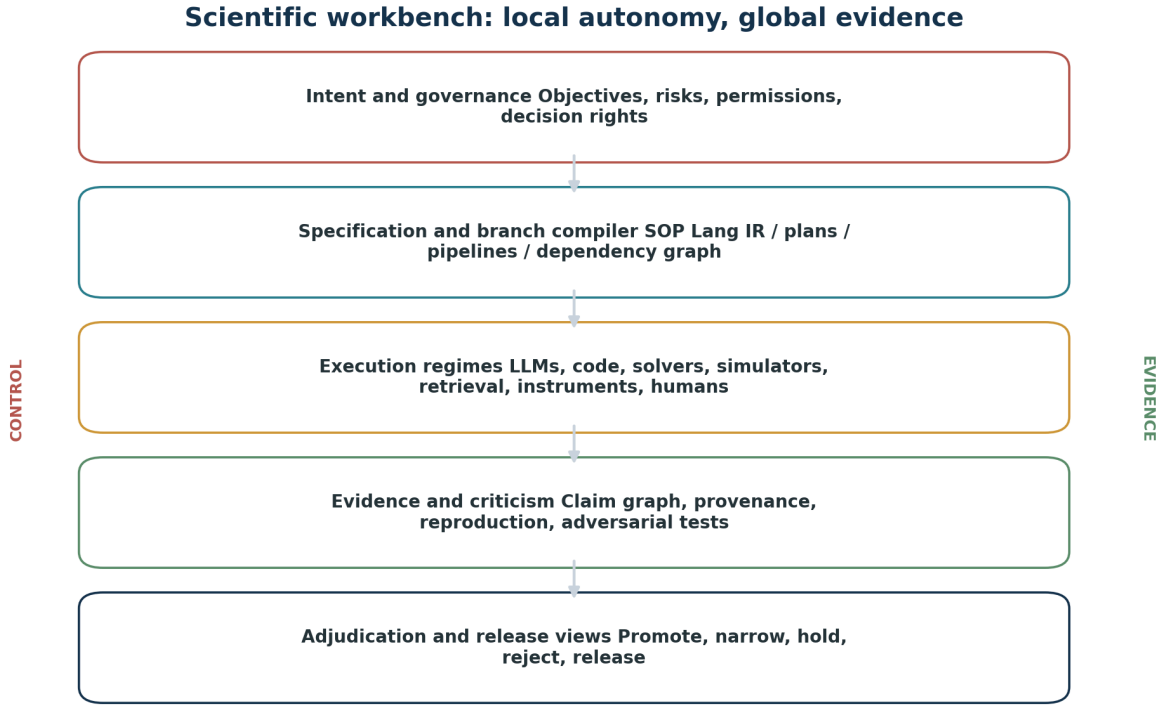


Figure 21. Reference architecture of the scientific workbench: local autonomy is bounded by global evidence and decision rights.

16.2 Scientific Interpreters and MRP-VM

The modern agent ecosystem has converged on several practical abstractions. Tools expose atomic capabilities. Model Context Protocol and related interfaces standardize access to data and services while preserving host-mediated boundaries. Agent frameworks provide state graphs, routing, retries, and human checkpoints. Coding environments increasingly maintain project instructions, specialized agents, hooks, and specification artefacts. Anthropic's Agent Skills popularize a package-like unit that combines instructions, resources, and repeatable workflows for a model. Kiro makes requirements, design, and task specifications first-class objects in an agentic development environment. These developments indicate a shift away from the monolithic prompt toward reusable, context-engineered execution units.

Their significance for science is substantial but limited. A conventional tool definition says what operation is available and what schema it accepts. A skill says more: when a procedure applies, how the model should use resources, and what workflow to follow. A spec-driven coding environment links intent to implementation and tests. An agent graph constrains control flow. Yet most industrial abstractions remain oriented toward task completion. They do not necessarily declare the epistemic scope of an output, the evidence required for acceptance, the independence of the critic, or the authority that can promote a scientific claim.

The generalization proposed here is to treat a skill as a bounded execution regime. A scientific interpreter does not merely perform a task. It declares the class of problems for which it is competent, the representations it expects, the assumptions it

makes, the evidence it can produce, the validators it invokes, the failure states it recognizes, and how its output can be reintegrated. This regime may be implemented by an LLM, a subagent, deterministic code, a theorem prover, a statistical package, a causal-analysis engine, a simulator, a database query, a robotic procedure, a human review protocol, or a composition of these elements.

Calling these units “interpreters” emphasizes their role at the boundary between specification and execution. A research specification contains commitments at a relatively abstract level. The interpreter gives those commitments operational meaning in a local domain. A citation-integrity interpreter resolves claim references, retrieves sources, checks whether the cited passages support the claim, and returns evidence with uncertainty. A statistical-validity interpreter maps an estimand and data description to diagnostics, model fits, sensitivity analyses, and a scoped verdict. A theorem interpreter translates a formal claim into proof obligations and invokes a prover. A laboratory-safety interpreter checks a protocol against hazard rules and instrument constraints. A narrative LLM may participate, but it does not need to control every step.

A “super-skill” is a useful name for a higher-order regime that can select and compose local interpreters. It might receive a research claim, infer its type, construct the relevant validation plan, execute inexpensive checks, and route unresolved questions to specialists. In implementation, a super-skill may sometimes be a subagent with its own planning loop. In other cases, it should be a deterministic dispatcher or a neuro-symbolic program. The point is not to make every unit agentic. It is to package domain competence together with applicability and evidence obligations.

This distinction creates a path from today’s practical skills to a more rigorous scientific runtime. Existing skills can be used immediately as bounded instruction packages. Over time, their contracts can be enriched with typed inputs, outputs, provenance, preconditions, side-effect declarations, cost models, validation hooks, and evidence schemas. Where symbolic mechanisms become available, LLM judgment can be replaced. The model remains valuable as a proposer, translator, renderer, or bounded interpreter, while the runtime relocates validation into inspectable procedures.

The MRP-VM concept within ACHILLES is an attempt to generalize this control layer. MRP stands for a meta-rational processing idea: the runtime should not assume one reasoning regime. It should retrieve only the knowledge relevant to the current state, condense that state into a smaller control nucleus, construct bounded frames for candidate regimes, route the task, execute locally, and reintegrate results into a common representation. A regime can be prompt-oriented, code-based, symbolic, retrieval-driven, human, or mixed. The runtime records why a regime was selected and which evidence it returned.

This architecture is more ambitious than ordinary orchestration. A workflow engine chooses nodes according to a graph. An agent chooses tools according to context. MRP-VM is intended to choose not only the next operation but the appropriate mode of interpretation and judgment. It asks whether the task should be handled as formal verification, causal analysis, literature synthesis, heuristic search, simulation, or human deliberation. It can also decide that the current representation is inadequate and request a change at a higher specification level.

The concept should be treated cautiously. A generalized meta-rational runtime risks becoming a vocabulary that merely renames routing. Its value must be demonstrated against simpler baselines: a monolithic prompt, a fixed workflow, a ReAct

loop, and a conventional agent graph. The tests should measure not only task accuracy but semantic fidelity, evidence quality, cost, latency, selective recomputation, critic independence, and failure localization. MRP-VM becomes a contribution only if its explicit regime model produces practical gains.

SOP Lang Plan, and Pipeline are complementary layers. Plan represents compact stepwise coordination where the next task may depend on interaction. Pipeline represents heavier dependency graphs and build-like workflows. MRP-VM supplies execution semantics for nodes whose realization differs. The same IR can support planning, pipeline execution, analysis, and audit. This unification is one of the central ACHILLES hypotheses.

The industrial boundary should remain clear. Structured outputs, agent frameworks, MCP, skills, coding agents, spec-driven environments, and workflow systems are existing mechanisms with varying maturity. The ACHILLES contribution is not the invention of agents, skills, or pipelines. It is the proposal to combine them around a specification-first intermediate representation and to generalize skills into evidence-bearing execution regimes governed by a meta-rational runtime. The research claim is that this combination can reduce reliance on opaque global LLM judgment while preserving adaptive capability.

The likely adoption path is incremental. A laboratory will not begin by deploying a universal VM. It may start with a versioned research specification and a small set of scientific skills: literature-grounding, protocol conformance, statistical checks, citation verification, and reproducibility packaging. These are composed in explicit pipelines. As contracts stabilize, the runtime can select among implementations based on cost and evidence. The same validation responsibility might be performed by an LLM subagent in an exploratory phase, deterministic code for routine cases, and a human specialist for consequential ambiguity. The workbench learns not only how to perform the task but which regime is appropriate.

This is the practical meaning of interpreters as intermediate steps in specification-driven development. They convert an abstract commitment into a local operation without surrendering global control. They allow symbolic processing where symbolic processing is faster and stronger, and use models where language and open-ended synthesis are indispensable. They make heterogeneity a designed property rather than an accidental collection of tools.

Table 31. Existing industrial mechanisms and the ACHILLES research hypothesis

What industry increasingly provides	What the ACHILLES direction adds as a research hypothesis
Structured generation through schemas, grammars, types, constrained decoding, and language-model programming	A common intermediate representation in which specifications, constraints, provenance, routes, candidate assemblies, and results coexist.
Agent frameworks that expose tools, memory, routing, subagents, and graph-based orchestration	SOP Lang Plan and Pipeline as explicit long-lived artefact graphs rather than only transient control flow.
Coding-agent skills and reusable instruction packages	Generalized skills as bounded execution regimes with applicability, evidence obligations, validators, costs, and reintegrable outputs.

What industry increasingly provides	What the ACHILLES direction adds as a research hypothesis
Context protocols and tool registries	Regime-sensitive interpretation through an emerging MRP-VM that can route between LLMs, deterministic code, retrieval, solvers, and domain validators.
Specification-oriented development environments	The Specification Ladder and Mirroring Pattern as a control mechanism for semantic drift across successive levels of abstraction.

The right column describes a proposed research programme, not a claim of universal production maturity.

Chapter 17. Critics, Evidence, and Decision Rights

Promotion from candidate to released claim should require increasingly independent forms of support and named authority.

17.1 Independent Critics and Evidential Escalation

A workbench that generates and executes without a strong criticism architecture will automate confidence rather than knowledge. The generator and certifier must therefore be separated structurally. Separation by prompt is not enough. Telling the same model to “be critical” after it has constructed a solution can improve local quality, but it does not create independence. The system needs distinct roles, evidence types, dependencies, and authorities.

The generator proposes claims, representations, plans, code, and experiments. The executor runs tools and records outputs in controlled environments. Critics evaluate bounded properties. An adjudicator changes claim status. These functions can share infrastructure, but their artefacts and responsibilities should remain distinct. The generator should not be able to promote its own claim by producing more persuasive prose.

Communication should occur through typed evidence packets rather than unconstrained debate alone. A packet identifies the claim, scope, assumptions, specification links, artefact identities, environment, executed operations, results, known failures, and the question assigned to the critic. The critic returns the method used, verdict, confidence boundary, counterexample or limitation, dependencies, and proposed next decisive test. A free-form explanation can accompany the packet, but the structure makes omissions visible.

Initial critic reports should be committed before cross-critic discussion. This preserves disagreement and reduces premature convergence. After reports are fixed, critics may inspect one another’s evidence, revise their positions, or request additional experiments. The workbench records both the initial and revised judgments. Consensus is then traceable rather than manufactured through averaging.

Critic diversity should be functional. A formal critic checks logical validity. An implementation critic checks whether code realizes the claimed method. A statistical critic examines uncertainty and assumptions. A causal critic asks whether the design supports the interpretation. A robustness critic explores distribution shifts, adversarial cases, and scaling. A novelty critic searches prior work. A domain critic judges whether the operationalization matters. A governance critic checks risk, consent, and release conditions. No critic should be presented as competent outside its declared scope.

Evidence should advance through stages of increasing independence and consequence. A practical ladder begins with an idea or analogy. The next stage is a specified, literature-grounded claim. The third is an executable artefact with recorded environment and tests. The fourth is independent reproduction. The fifth adds adversarial, scaling, and negative-control testing. The sixth adds accountable specialist review. The seventh is external or deployment-specific validation. These stages are not a universal scalar truth measure. They prevent different forms of evidence from collapsing into one confidence label.

The ladder also supports different release thresholds. An exploratory conjecture may be useful at an early stage if uncertainty is explicit. A software method may be

shared once it is executable and reproduced, while claims of general superiority remain provisional. A clinical recommendation cannot rely on internal reproduction alone. A theorem may reach strong formal status while its relevance to an application remains weak. Claim type and intended consequence determine which evidence stages are necessary.

Causal attribution deserves a dedicated critic because sophisticated systems frequently earn credit for effects produced by simpler factors. The critic constructs matched baselines and removal tests. It asks whether the capability survives replacement of the proposed mechanism with the simplest carrier that preserves the evaluator's observations. It checks whether data, compute, preprocessing, or evaluation changed. It tries to separate "the system works" from "the explanation is correct." Automated research is especially well suited to this task because code substitution and reruns can be cheap, provided the specification requires them.

Decision rights connect the ladder to governance. A low-risk transition from idea to executable prototype can be policy-driven. Promotion to a public claim may require an author. Promotion to a clinical or safety-relevant recommendation may require institutional review and external evidence. The workbench must represent not only who can execute a step, but who can authorize a status change. Capability and authority are different.

The same principle applies to stopping. An agent may be capable of running another experiment, but the budget owner may decide that the expected information value is too low. A model may produce a dual-use procedure, but the safety authority may block release. A reviewer may request data that participants did not consent to share. The runtime should treat these decisions as legitimate terminal states, not obstacles to be circumvented.

Machine-assisted peer review fits naturally into this architecture. AI critics can perform preflight checks before submission, reducing avoidable burdens on reviewers. During review, they can test claims, inspect provenance, compare code and prose, or suggest counterexamples. Human reviewers decide which issues matter, assess significance and fairness, and take responsibility for recommendations. After publication, the same infrastructure can attach replication and criticism continuously. Review becomes a lifecycle rather than a one-time verdict.

The architecture also reduces the temptation to build one omniscient reviewer model. A single model is attractive because it is easy to deploy and produces coherent reports. Its coherence can hide shared failure. A portfolio of specialized critics with explicit dependencies is more cumbersome but epistemically stronger. Where critics disagree, the workbench should escalate the exact point of disagreement rather than ask a larger model to synthesize a smooth answer.

The final design principle of this chapter is that evidence and authority must be explicit at every promotion boundary. Local autonomy is valuable because it expands search and reduces procedural cost. Global evidence is necessary because local success can be misleading. Decision rights are necessary because scientific outputs enter social worlds. The workbench is credible only when these three layers remain connected without being collapsed.

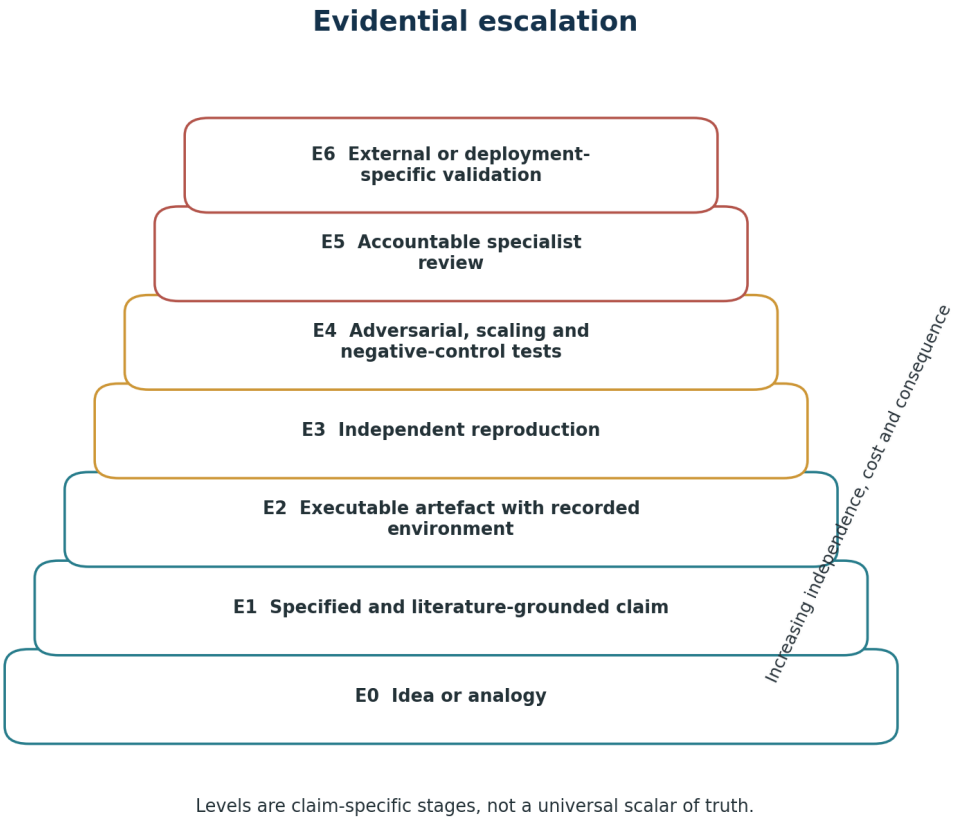


Figure 22. Evidential escalation. Promotion requires new types of support, not repeated confidence from the same generator.

Table 32. Evidential escalation from candidate to released claim

Evidence level	Required change in support
E0 - Candidate	A hypothesis, mechanism, proof sketch, model, or interpretation exists, but no competent check has yet supported it.
E1 - Internally coherent	Basic type, syntax, consistency, and dependency checks pass; this eliminates malformed candidates, not false scientific claims.
E2 - Executed	The proposed computation, protocol, or experiment has run and produced recorded outputs under a declared environment.
E3 - Challenged	Negative controls, ablations, counterexample search, alternative explanations, or adversarial critics have tested the central inference.
E4 - Independently reproduced	A materially independent executor, environment, dataset, laboratory, or method obtains compatible evidence.
E5 - Specialist adjudicated	Qualified human or institutional reviewers evaluate significance, validity, risk, and unresolved disagreement.
E6 - Released under policy	The result meets the domain-specific threshold for publication, deployment, clinical use, regulation, or another consequential action.

17.2 Verification-First Scientific Infrastructure

The largest opportunity in machine discovery may not be a model that generates more hypotheses. It may be infrastructure that makes hypotheses easier to test,

compare and accumulate. Science already produces more papers, data and code than any individual can evaluate. Artificial intelligence can increase output further. Without stronger verification, the result may be noise, duplicated work and confident error. Verification-first infrastructure treats evidence flow as the primary object of automation.

Such infrastructure begins with literature that is computationally traceable. Retrieval systems should identify passages, not merely documents, and distinguish direct evidence from commentary. Claims should link to datasets, code, assumptions and related contradictions. OpenScholar demonstrates the scale possible for citation-backed synthesis across millions of papers (Asai et al., 2026). The next step is claim-level entailment, provenance and update when a source is corrected or retracted.

Figure 23. The discovery frontier is a constellation of better questions, world models, causal experiments, formal verification, open-ended search, scientific memory, living evaluation and collective intelligence. Conceptual diagram; not an empirical result.

The figure matters because no single breakthrough is sufficient. Better world models without experiments can become elaborate simulators. Open-ended search without verification produces variation. Formal methods without better questions optimize narrow specifications. Scientific memory without living evaluation accumulates

stale claims. The central infrastructure node represents interfaces that allow these capabilities to correct one another.

Executable papers are one component. Code, environment, data transformations and figures can be rerun automatically. Agents can test whether results survive dependency changes, alternate seeds and reasonable analysis choices. Formal schemas can declare primary outcomes, exclusions and stopping rules. This does not automate conceptual validity, but it makes hidden analytic flexibility visible and reduces the cost of replication.

Claim graphs are another component. A paper contains many claims with different evidence. Representing them as linked objects allows a correction to propagate to dependent conclusions. Competing hypotheses can share evidence without being merged. Agents can identify which new experiment would resolve the largest disputed region. The graph becomes a research map rather than a static library.

Table 33. Infrastructure that may yield more discovery than a larger generator alone

Infrastructure	Scientific leverage
Claim-level literature graph	Connects assertions to evidence, contradiction, correction and priority.
Executable research object	Makes code, data and analysis reproducible and testable by agents.
Formal and empirical verifier services	Provide shared checking for proofs, programs, models and measurements.
Versioned scientific memory	Preserves revisions, negative results and alternative hypotheses.
Federated laboratory network	Routes high-value candidates to independent experimental validation.
Living benchmark and incident archive	Updates evaluation as systems, users and environments change.

The table exists because these resources create positive externalities. A better generator helps its owner; a shared verifier or claim graph can improve an entire field. Infrastructure also changes incentives by making reproducibility, negative results and corrections more visible. It may reduce the advantage of producing many polished claims and increase the value of producing evidence that others can reuse.

Federated laboratories could provide prospective validation. Agents rank hypotheses by expected value and uncertainty; independent sites volunteer capacity; protocols and raw data are standardized; results are aggregated with site effects modelled explicitly. This would not remove human expertise. It would make replication a routinized service rather than an exceptional act.

Formalization tools can expand the region of cheap verification. Natural-language claims can be translated into testable specifications, with humans approving the translation. Program properties, mathematical lemmas, data constraints and experimental protocols can then be checked automatically. Autoformalization remains error-prone, so the mapping between informal intent and formal object must stay visible.

The most valuable novelty may appear in evaluator design itself. A new assay, proof representation, benchmark environment or measurement instrument can unlock discoveries that generators could not previously validate. Historically, scientific instruments changed the space of questions. Computational verifiers and autonomous experiment networks may play a similar role.

Verification-first infrastructure is slower to dramatize than a model demo. Its payoff is cumulative. It lowers the cost of being wrong, makes disagreement productive and allows more ambitious generators to operate safely. The scientific frontier advances when novelty can encounter resistance quickly.

17.3 A Protocol for a Novelty Claim

A novelty claim should be decomposed before it is accepted or rejected. The protocol below is designed for papers, model outputs, laboratory campaigns and algorithmic discoveries. It does not produce a universal novelty score. It identifies which claim is being made, what evidence supports it and where uncertainty remains.

Table 34. Novelty-claim audit

Audit question	Evidence that should accompany the answer
What exactly is claimed to be new?	A precise object, level of description and domain of comparison.
New relative to which record?	Named training data, databases, literature corpora, patent searches or historical scope.
What counts as equivalent?	A domain-specific semantic, functional, structural or mathematical equivalence rule.
How was correctness tested?	Formal checking, execution, simulation, measurement, replication or expert assessment.
Why is the difference significant?	Improved explanation, capability, prediction, intervention, compression or access to a new region.
Which evidence could defeat the claim?	A counterexample, prior work, failed replication, alternative mechanism or measurement.
What is the provenance?	Data, code, model versions, prompts, instruments, parameters and reviewer decisions.
What remains uncertain?	Priority, generality, mechanism, translation, measurement error and boundary conditions.

The table exists to prevent the word “novel” from carrying several unsupported conclusions at once. A candidate may pass the database comparison and fail the correctness test. It may be correct but insignificant. It may be significant and already known under another representation. Reporting these dimensions separately allows a result to be valuable without overstating it.

In practice the audit should be applied at the claim level, not only to the paper or system. A single project can contain a new dataset, a familiar method, an improved result and an unverified mechanistic interpretation. Each deserves a different status. Automated tools can assist with literature comparison, code execution and provenance, but final priority and significance judgments remain scoped and revisable.

Chapter 18. The Atlas Inside the Workbench

Operational ideas become the workbench memory through which precedents, transformations, lineages, and unresolved alternatives remain inspectable.

18.1 A Plural, Evidence-Grounded Representation

The proposed atlas is naturally described as neuro-symbolic, but the term can conceal more than it clarifies. Many systems are called neuro-symbolic because they combine a neural network with a knowledge graph or because they translate model outputs into logical predicates. The scientific requirement here is stricter. The architecture must allocate different epistemic responsibilities to different representational forms and prevent one layer from silently overriding the commitments of another.

Neural models are well suited to open-ended linguistic variation. They can retrieve semantically related passages, align terminology across fields, propose role structures, translate historical language, and generate candidate analogies. Their weakness is not merely hallucination. Distributed representations make it difficult to determine which source supports a particular abstraction, which invariant produced a match, and whether an answer remains valid after the knowledge base changes. A fluent model can produce a coherent lineage story without having distinguished resemblance from influence.

Symbolic representations are suited to explicit roles, typed relations, constraints, and inspectable inference paths. They can state that a later method preserves a goal, replaces a mechanism, removes an assumption, and imports an evaluation technique from a third source. They can enforce temporal constraints and prevent an earlier work from inheriting from a later one. Their weakness is brittleness and the cost of constructing the symbols. They do not solve lexical variation or open-ended abstraction simply by being explicit.

Probabilistic representations are required because many relevant claims are uncertain rather than simply true or false. The uncertainty may concern extraction, interpretation, dating, influence, or corpus coverage. Probabilistic logic and statistical relational learning offer ways to combine uncertain evidence with relational constraints (Manhaeve et al. 2018; Raedt et al. 2020). Yet probability alone is not enough. Competing decompositions can be qualitatively different hypotheses, and a single confidence value may hide the source of disagreement.

The architecture should therefore preserve four kinds of object. Documentary objects identify source material. Interpretive objects represent proposed operational ideas. Relational objects state typed identity, transformation, support, and lineage claims. Argument objects connect those claims to evidence, rules, model versions, and challenges. This design draws on semantic publishing, nanopublications, and micropublications, which demonstrate that small claims and their provenance can be made machine-actionable (Groth, Gibson, and Velterop 2010; Kuhn et al. 2018; Clark, Ciccarese, and Goble 2014). The atlas extends this discipline to structured ideas and historical comparison.

Evidence grounding must be local. A paper-level citation is insufficient when the system claims that a specific mechanism appears in an earlier source. The relation should point to the passage, equation, diagram, code segment, or experimental artefact that justifies the abstraction. If the claim is a synthesis not stated by any source, it must

be labelled as such and linked to the evidence from which it was inferred. This distinction is central for trust: the user must be able to tell what was observed, what was interpreted, and what was generated.

Representation must also be plural. A symbolic graph should not enforce one irreversible decomposition. It should allow several views of the same evidence and relations between those views. One representation may be mechanism-centred, another functional, another historical. The graph can record that a coarse object abstracts a finer one, that two experts disagree about a boundary, or that a model's candidate interpretation was rejected. Plurality is not database disorder when it is typed and versioned.

Table 35 summarises the epistemic division of labour. The second column describes both the distinctive contribution of each layer and the boundary it should not cross without additional support.

Table 35. A neuro-symbolic atlas requires an explicit division of epistemic labour among representational layers.

Representational layer	Epistemic responsibility and limit
Source and attestation layer	Preserves the documentary record, exact locations, versions, and artefacts; it does not by itself determine the identity of the represented idea.
Neural semantic layer	Retrieves, translates, aligns, abstracts, and proposes candidate structures; its outputs remain hypotheses until grounded and adjudicated.
Symbolic structural layer	Records roles, typed relations, constraints, and explicit transformation paths; it should not pretend that its chosen ontology is complete or uniquely correct.
Probabilistic and temporal layer	Represents uncertainty, competing hypotheses, censoring, and time-indexed knowledge states; probability must remain attached to the source of uncertainty.
Argument and provenance layer	Connects conclusions to evidence, agents, models, rules, revisions, and objections; it prevents synthesis from being confused with quotation.
Human interpretive layer	Resolves difficult cases, supplies domain and historical judgment, and challenges the system; expertise is recorded rather than treated as infallible authority.

The table explains why a single end-to-end model is not the target, even if such a model could achieve high benchmark accuracy. The atlas is intended to support claims that can affect credit, review, and research direction. Those claims must remain auditable after the model is replaced. Representational separation gives the system institutional memory: evidence and arguments survive changes in the neural component.

18.2 Neural Proposal and Symbolic Adjudication

The architecture should be organised around proposal and adjudication rather than extraction and storage. "Extraction" suggests that the correct idea structure is already present in the text waiting to be copied. In reality, the system proposes a representation by combining spans, background knowledge, and a task model.

Adjudication evaluates whether the proposal is supported, coherent, and appropriate for the inquiry.

A neural proposal begins with source selection. Retrieval must operate over multiple views: lexical, semantic, citation-based, temporal, and structural. A candidate precedent can be relevant because it uses the same terminology, because it instantiates the same mechanism under different language, or because it occupies a key position in a citation lineage. The retrieval model should preserve which path produced the candidate. Otherwise, downstream users cannot distinguish direct evidence from exploratory analogy.

The model then proposes roles and relations. It may identify a goal, mechanism, assumptions, and outcomes; align them with existing idea objects; and suggest a transformation such as specialisation or removal of a limitation. Crucially, it should produce alternatives when the text supports more than one decomposition. A calibrated system can state that a passage probably instantiates either a general optimisation principle or a domain-specific procedure and that the difference affects the novelty analysis.

Symbolic adjudication checks these proposals against constraints. Some constraints are logical: a relation may require compatible object types. Some are temporal: an influence edge cannot run backward in time. Some are causal or mechanistic: claimed equivalence may require preservation of dependency structure. Some are evidential: an assertion of influence requires stronger support than a claim of similarity. Some are historiographical: a translation may not justify attributing modern terminology to an earlier author.

Constraint checking should not be restricted to rejection. It can generate questions that guide further retrieval. If two methods appear mechanistically equivalent but have incompatible assumptions, the system searches for a finer representation or evidence that one assumption is dispensable. If an influence hypothesis is semantically strong but documentary weak, the system searches for intermediary publications, correspondence, shared collaborators, or educational channels. Reasoning becomes an active process of evidence acquisition.

Probabilistic interpretation combines uncertain evidence without erasing its type. A useful model might assign separate distributions to extraction correctness, structural alignment, historical influence, and corpus completeness. The final report can then state, for example, that mechanistic equivalence is strongly supported, influence is possible but undocumented, and priority remains unstable because non-English archival coverage is poor. This is more informative than a single confidence score of 0.72.

Neuro-symbolic systems such as DeepProbLog and Logic Tensor Networks illustrate different ways of combining learning with formal constraints (Manhaeve et al. 2018; Badreddine et al. 2022). The atlas is not committed to one framework. Its conceptual requirement is that the inference path remain inspectable and that uncertainty be attached to claims. Some components may use probabilistic logic, others differentiable constraints, graph neural networks, or learned theorem-guided retrieval. The architecture should be judged by epistemic behaviour rather than by membership in a brand of neuro-symbolic AI (Garcez and Lamb 2020; Hitzler and Sarker 2022; Sarker et al. 2021).

Figure 24 depicts the system as a cycle. Evidence feeds neural proposal; candidate assertions enter symbolic adjudication; probabilistic and temporal reasoning

assesses competing hypotheses; human challenge introduces corrections and new sources; and those revisions alter the evidence and learning process. The cycle has no final “truth database.” It produces a succession of better-supported and more precisely qualified interpretations.

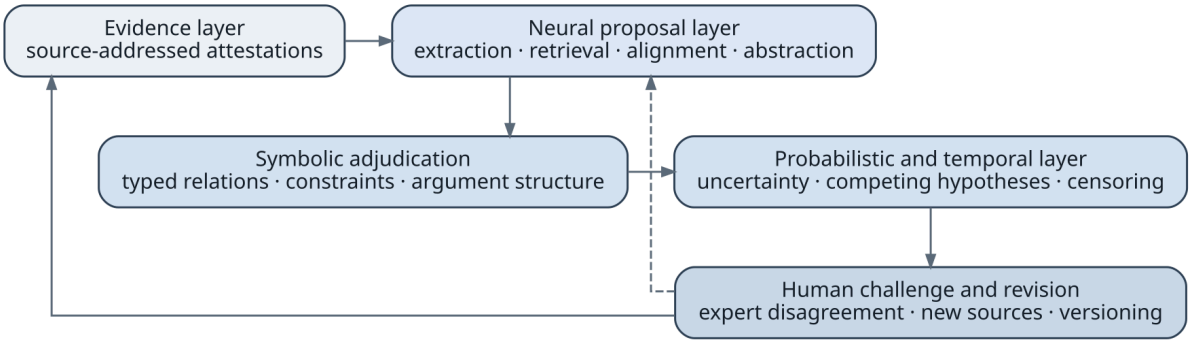


Figure 24. The neuro-symbolic epistemic cycle. Neural flexibility is constrained by symbolic structure, uncertainty is represented explicitly, and human challenge can revise both evidence and models.

The human role in the figure is not a residual task left over because automation is imperfect. Some questions are interpretive in principle. Experts decide which differences are explanatory, whether a historical source performs the same intellectual act, and how much abstraction a comparison can tolerate. The system can make these judgments more consistent and evidence-rich, but it should not hide them behind automation. Human decisions themselves become provenance-bearing objects that can be studied for disagreement and bias.

An atlas of ideas must expect its conclusions to change. Classical databases often assume that adding facts monotonically increases knowledge. Historical and semantic interpretation is non-monotonic: a new source can invalidate a priority claim, a better translation can split an apparent identity, and a new experiment can transform a mechanism previously treated as equivalent. The architecture should therefore treat revision as normal rather than as corruption.

Versioning must operate at several levels. Sources have editions and corrections. Attestations can be relocated when pagination changes. Idea objects can be refined or split. Relations can be accepted, challenged, or deprecated. Ontology terms can change. Model outputs depend on model versions and prompts. A user examining a claim should be able to reconstruct the state of the atlas at a previous date and understand why the current view differs.

Argumentation provides a natural structure for revision. A claim such as “work B independently rediscovered the mechanism in work A” can be supported by structural alignment, temporal separation, and absence of plausible transmission, while challenged by evidence of a shared intermediary. The system need not force every argument into deductive logic. It can represent defeasible rules, evidential weights, and unresolved rebuttals. What matters is that the conclusion is not detached from its reasons.

This approach also reduces the temptation to encode governance decisions as technical defaults. Consider the threshold for merging two idea objects. A low threshold creates fragmented duplicates; a high threshold erases meaningful variation. The choice affects attribution and novelty. It should be visible, testable, and, where appropriate, adjustable by domain. Likewise, the decision to treat a source as

authoritative, a language model as a curator, or expert consensus as sufficient is a governance choice with epistemic consequences.

The atlas should support contestability. Researchers need a way to challenge an identity relation, supply an earlier attestation, propose an alternative decomposition, or flag a translation problem. Challenges should not immediately overwrite accepted records; they should create parallel argument states until adjudicated. This prevents both vandalism and premature closure. It also makes the atlas a site of scholarship rather than a static reference work.

Bias audits must examine representation and coverage, not only model outputs. A system can be calibrated on its indexed corpus and still be systematically unjust because the corpus excludes relevant traditions. It can accurately reproduce citation patterns that reflect historical inequality. Coverage metadata should therefore be part of every high-stakes novelty or priority analysis. The absence of accessible sources in a region, language, or publication type should lower confidence and trigger targeted search.

Open standards such as PROV-O and FAIR principles provide part of the infrastructure for traceability and reuse (World Wide Web Consortium 2013; Wilkinson et al. 2016). The deeper governance challenge is semantic. Who may create or revise idea families? How are minority interpretations preserved? When is a relation sufficiently supported for high-stakes use? Which claims should remain advisory? These questions require institutional design, but the conceptual answer is clear: no single opaque model should control identity, novelty, and credit.

The architecture should also distinguish exploration from adjudication. During exploration, the system can generate speculative analogies, distant candidates, and alternative decompositions with permissive thresholds. During adjudication, evidential standards become stricter, and unsupported relations remain hypotheses. Mixing the two modes creates danger. A creative suggestion is useful precisely because it can be weakly supported; a priority judgment is credible only when it is conservatively supported.

The proposed neuro-symbolic machine is therefore not a compromise between “neural” and “symbolic” camps. It is an epistemic constitution. It assigns flexibility, explicitness, uncertainty, and authority to different layers and defines how they may correct one another. Its success depends less on whether every component is differentiable than on whether the whole system makes scientific interpretation inspectable, revisable, and resistant to the confident collapse of distinct relations.

18.3 Scientific Memory, Synthesis, and Attribution

The first practical impact of an idea atlas would not be automated discovery. It would be a change in scientific memory. Today, literature synthesis is forced to move repeatedly between documents and implicit conceptual models held by readers. Reviewers and authors reconstruct which mechanisms recur, which assumptions changed, and how results relate, but those reconstructions are rarely preserved in reusable form. A working atlas would make the intermediate reasoning available for inspection and cumulative improvement.

Systematic reviews could represent not only interventions and outcomes but the operational ideas connecting them. Studies that appear heterogeneous at the level of terminology might share a mechanism; studies grouped under one method name might

in fact implement different causal structures. An idea-centred representation could help reviewers decide when evidence should be combined, when subgroup analysis is required, and where apparent disagreement results from different assumptions. The benefit would depend on domain-specific validation and should not replace established review protocols.

Narrative reviews and textbooks would gain a different capability. Instead of presenting a lineage as a static story, they could expose alternative decompositions and disputed continuities. Readers could move from a general principle to its historical formulations, technical refinements, applications, and known failures. The map would show not only who cited whom but what was inherited or changed. This could make review writing more accountable and reduce the repeated reconstruction of the same intellectual history.

Laboratory memory is another important setting. Research groups accumulate tacit knowledge about failed parameter regimes, abandoned hypotheses, local modifications, and reasons for design choices. Much of it disappears when members leave or survives only in notebooks and conversations. An atlas linked to protocols, code, and results could represent why a method was attempted, which earlier idea it instantiated, and which boundary condition caused failure. Capturing negative results would help prevent unproductive repetition and improve the interpretation of later successes.

Attribution and priority research could become more nuanced. The system could search across languages and source types for earlier attestations, separate first formulation from first validation, and distinguish structural similarity from evidence of influence. Historians would still adjudicate the conclusions, but the search space would become more tractable. A conservative and transparent atlas might correct canonical narratives that currently reflect visibility more than chronology.

Peer review could use the atlas as an evidence instrument. A reviewer evaluating a novelty claim could examine the closest structured precedents and the role-wise differences. This could reduce superficial novelty based on terminology and help recognise important local changes that ordinary similarity underestimates. The risk is obvious: institutions may turn a support tool into an automated gatekeeper. The atlas should therefore present evidence and alternative interpretations, not a pass–fail novelty score.

Grant evaluation and research portfolio analysis present a similar tension. An idea map could identify duplicated mechanisms, neglected combinations, and clusters that share hidden assumptions. Funders might use it to diversify portfolios or find high-information experiments. They might also use it to penalise unconventional proposals that lack recognised antecedents. The technology’s impact depends on whether it expands the space of inquiry or narrows it to what the ontology already understands.

Table 36 pairs major academic settings with the transformation a successful atlas could enable. The second column also states the main condition that prevents the application from becoming a simple automation story.

Table 36. Academic impact would arise from better conceptual memory and comparison, provided the atlas remains evidence-centred and contestable.

Academic setting	Potential transformation and necessary condition
Systematic and evidence reviews	Align studies by mechanism, assumptions, and evidential role; domain validation must determine when structural similarity justifies synthesis.
Narrative reviews and textbooks	Preserve alternative lineages and multi-resolution explanations; editorial interpretation must remain visible and contestable.
Laboratory and organisational memory	Connect hypotheses, design choices, failures, code, and protocols to reusable ideas; tacit and negative knowledge must be captured without false precision.
Peer review	Provide structured precedents and localised differences; the system must support judgment rather than issue authoritative novelty verdicts.
Grant and portfolio analysis	Reveal conceptual duplication, neglected mechanisms, and strategic diversity; uncertainty and ontology bias must be explicit.
Historical attribution	Search for earlier attestations and separate priority from influence; corpus coverage and dating evidence must accompany every claim.
Research integrity analysis	Distinguish textual copying, conceptual reuse, common method, and independent rediscovery; similarity alone cannot establish misconduct.
Scientific education	Teach concepts through transformations, applications, and failure boundaries; multiple traditions must be represented rather than reduced to a single canon.

The table shows that the practical applications share one underlying function: preservation of intermediate epistemic structure. They do not require the system to generate new ideas. Even if discovery support proves limited, a reliable instrument for synthesis and attribution could have substantial value. This is an important asymmetry in the impact case. The atlas can be useful before it becomes creative.

18.4 Cross-Domain Transfer and Institutional Consequences

The most ambitious application is to support the generation of scientifically useful ideas. Existing language-model systems already produce proposals, and studies show that researchers can find some of them novel or inspiring (Si, Yang, and Hashimoto 2024; Radensky et al. 2024). The difficult question is whether structured memory improves the quality of the search rather than merely the persuasiveness of the output.

An idea atlas could support discovery in several ways. It could reveal unresolved residuals in a lineage, such as a persistent assumption that no work has removed. It could identify mechanisms repeatedly successful under one class of constraints and search for domains with analogous structure. It could locate components that have never been combined despite compatible roles. It could also identify where evidence lags behind theory, suggesting experiments with high information value.

Cross-domain transfer is especially promising because disciplines often use incompatible vocabularies for similar structures. Analogy research suggests that productive transfer depends on relational mapping rather than surface resemblance (Gentner 1983). An atlas with mechanistic and functional views could search for distant cases that share control structure, feedback, invariance, resource constraints, or

measurement problems. It should then expose the mismatches that could invalidate transfer. The value lies as much in explaining why an analogy may fail as in generating it.

Recombination systems such as CHIMERA, Scideator, and Graph2Idea demonstrate that structured facets and relational contexts can guide ideation (Sternlicht and Hope 2026; Radensky et al. 2024; Li, Tu, and Han 2026). The atlas would extend this work by adding lineage, identity types, negative evidence, and transformation grammars. Instead of asking only which components can be combined, it could ask what form of change has historically overcome a similar limitation and whether the enabling assumptions exist in the target domain.

The system could also improve abductive reasoning. Given an anomalous result, it might retrieve mechanisms that produce an analogous pattern elsewhere, align their assumptions, and suggest discriminating tests. Because the atlas records evidence and failure conditions, hypotheses can be ranked by explanatory fit and by the cost of experiments that separate them. Such support would be closer to scientific reasoning than unconstrained text generation.

For artificial intelligence, the atlas could function as an external memory and a training resource. Document corpora supply abundant language but sparse supervision for conceptual relations. The atlas would provide pairs and lineages labelled with paraphrase, functional identity, mechanistic change, independent rediscovery, and evidence type. Models could be trained contrastively to preserve these distinctions and evaluated on whether they recover the correct relation rather than merely produce fluent explanations.

The provenance layer could improve retrieval-augmented generation. Instead of retrieving entire documents, an AI system could retrieve a structured idea object together with its supporting attestations, competing interpretations, and known limitations. This would reduce context redundancy and make citations more local. The language model would still generate explanations, but the claims would be constrained by inspectable evidence.

A scientific agent could use the atlas to maintain a research state. It would record which components of a hypothesis are inherited, which are speculative, which experiments bear on them, and how new evidence changes the structure. This is preferable to a conversational memory that stores only prose summaries. The agent's reasoning could be evaluated against the explicit transformations it proposes.

Training on an atlas also presents risks. The representation may harden existing categories and make models less able to recognise ideas outside the ontology. Widely accepted but mistaken lineages could become high-quality supervision. If the atlas overrepresents successful work, agents may learn to imitate recognised trajectories and avoid high-risk exploration. Training should therefore include disagreements, rejected relations, negative results, and examples where the system's initial identity model was revised.

The impact on scientific acceleration should be measured in experiments, not inferred from the sophistication of the representation. Relevant outcomes include time to find decisive precedents, quality of hypotheses, value of selected experiments, replication success, and the ability to transfer mechanisms across fields. Citation impact is too delayed and socially confounded to serve as the only measure. A system that

helps researchers abandon a bad idea earlier may accelerate science without producing a highly cited paper.

A technology that maps identity, novelty, and lineage would not be neutral infrastructure. It would influence how credit is assigned, which proposals appear original, and which traditions are visible. The closer it comes to working, the greater the governance burden. The central risk is not only technical error but the conversion of contestable interpretations into institutional facts.

Credit is particularly sensitive. A priority map can restore overlooked contributions, but it can also produce false precision. An early vague resemblance may be promoted over the first clear formulation; a dominant-language source may appear earliest because other corpora are missing; a collective tradition may be reduced to an individual node. The atlas should therefore represent several forms of precedence and avoid translating them automatically into authorship or ownership.

Research evaluation could be distorted by metricisation. Once novelty profiles exist, institutions may seek a single score for hiring, funding, or publication. This would invite optimisation against the measure. Authors could maximise unusual combinations, manipulate the representation level, or strategically frame familiar work as a new mechanism. The system should make such compression difficult by refusing a universal score and by exposing the underlying comparisons.

Openness creates a second tension. An open atlas supports correction, replication, and broad access. Yet some scientific ideas involve security, privacy, dual use, or proprietary knowledge. The provenance system must represent access restrictions without treating unavailable evidence as nonexistent. Governance may require tiered access, embargoes, and policies for sensitive lineages. These measures can reduce transparency, so their effects should be recorded.

The atlas could alter intellectual property practice. Prior-art search might become more semantically powerful, and firms could map transferable mechanisms across patents and scientific literature. This may reduce weak claims based on terminology. It may also favour organisations with the resources to curate proprietary corpora and challenge public priority records. Public research infrastructure would be important to prevent conceptual search from becoming exclusively private.

Education could benefit from maps that connect principles to their transformations and applications. Students often encounter results as isolated named techniques. An idea atlas could show why a method was introduced, which limitation it addressed, and how it changed. The risk is canon formation: the map may present one dominant lineage as inevitable. Educational use should include alternative traditions, failed paths, and disputes about identity.

At a societal level, a working atlas could improve collective problem solving by exposing structural analogies between scientific, engineering, and policy domains. However, transfer into social systems is especially hazardous because agents react to interventions and concepts carry normative meanings. Mechanistic analogies from physics or computation can become misleading metaphors when institutional context is ignored. The system must represent domain incompatibilities, not only similarities.

The long-term potential is substantial. Scientific literature could become a substrate for explicit reasoning about transformation rather than a warehouse of documents. AI systems could be trained on conceptual inheritance and evidential change. Researchers could inspect how an idea travelled, where it failed, and which

unresolved variants remain. But this potential depends on cultural practices as much as algorithms. Communities must value correction, preserve negative evidence, and tolerate multiple representations.

A successful atlas would therefore have an unusual form of impact. It would not simply automate a task. It would create a new public object: the contestable, evidence-grounded representation of an idea and its transformations. Such objects could become part of scholarly communication in the way datasets and code have become part of many fields. Their value would lie in allowing scientific communities to reason collectively about continuity and change while retaining the sources and disagreements that make those judgments credible.

Chapter 19. VSA/HDC and the Novelty Residual

High-dimensional symbolic vectors are treated as a falsifiable compositional-memory hypothesis, never as a universal novelty oracle.

19.1 Why High-Dimensional Symbolic Vectors Are Relevant

Vector Symbolic Architectures and Hyperdimensional Computing describe a family of models that encode structured information in high-dimensional distributed vectors. Their characteristic operations bind items into role–filler associations, bundle several associations into a composite representation, and often permute or transform vectors to represent order and position. Because unrelated high-dimensional vectors are approximately dissimilar and because superposed structures can be recovered approximately, these models occupy an unusual position between classical symbols and learned embeddings (Kanerva 2009; Kleyko et al. 2023).

This position makes VSA/HDC relevant to the atlas for conceptual rather than merely computational reasons. Idea identity requires both structure and tolerance. A representation must preserve that a mechanism fills one role and an assumption another, while recognising approximate correspondences across terminology and domains. Classical symbolic structures preserve roles exactly but require explicit matching. Ordinary embeddings provide flexible similarity but tend to entangle roles. VSA/HDC offers a hypothesis about how role-sensitive structures could inhabit a distributed similarity space.

The family is diverse. Tensor product representations preserve compositional structure in a high-dimensional tensor product (Smolensky 1990). Holographic Reduced Representations compress binding through circular convolution (Plate 1995). Binary spatter codes, multiply–add–permute models, sparse distributed representations, and other architectures use different vector spaces and operations. Their algebraic properties are not interchangeable. Binding may be self-inverse in one architecture, require an explicit inverse in another, or preserve similarity to different degrees. Comparative work shows that the choice of architecture affects analogical reasoning, capacity, and decoding (Schlegel, Neubert, and Protzel 2022; Kleyko et al. 2023). Any proposal that speaks of “the VSA representation” without specifying the algebra is underdefined.

The attraction of high-dimensional vectors lies partly in concentration phenomena. Random vectors in a sufficiently high-dimensional space are likely to be nearly orthogonal, allowing many symbols to be assigned without careful geometric placement. Bundling creates a noisy superposition whose constituents remain detectable by similarity. Binding produces a vector dissimilar to its inputs while permitting approximate recovery when one factor is known. These operations can encode sets, sequences, graphs, and role–filler structures in fixed-size distributed representations.

For an atlas, fixed size is less important than compositional comparison. Suppose two ideas share a goal and mechanism but differ in assumptions and evidence. A role-bound representation can in principle recover each role separately, estimate correspondence, and still compare the whole. The representation can also support associative retrieval: a partial query specifying a problem and mechanism may retrieve candidates whose surface language differs. This is a stronger use of VSA/HDC than employing hypervectors as efficient classifiers.

The approach also suggests a computational treatment of transformations. If related ideas differ through operations such as substitution, specialisation, role transfer, or recombination, those operations may themselves be represented and compared. Recurrent transformation patterns could become first-class objects: “remove an independence assumption,” “replace deterministic control with feedback,” or “transfer a mechanism from a discrete to a continuous domain.” An atlas could then search not only for similar ideas but for similar *changes*.

There is, however, a danger of mistaking representational convenience for semantic theory. High dimensionality does not determine which roles should exist, which fillers are equivalent, or which transformations are scientifically meaningful. These choices come from the operational ontology, evidence, and task. Random codebooks provide separability, not interpretation. Learned codebooks may capture useful regularities, but they can also reproduce biases and entangle distinctions the symbolic layer needs to preserve.

VSA/HDC should therefore be tested as an associative compositional memory within a larger neuro-symbolic system. The authoritative record remains the evidence-grounded graph. Hypervectors provide rapid retrieval, approximate alignment, transformation search, and factorisation hypotheses. Their outputs must be translated back into role-level assertions and checked against sources. This division of labour is represented in Figure 25.

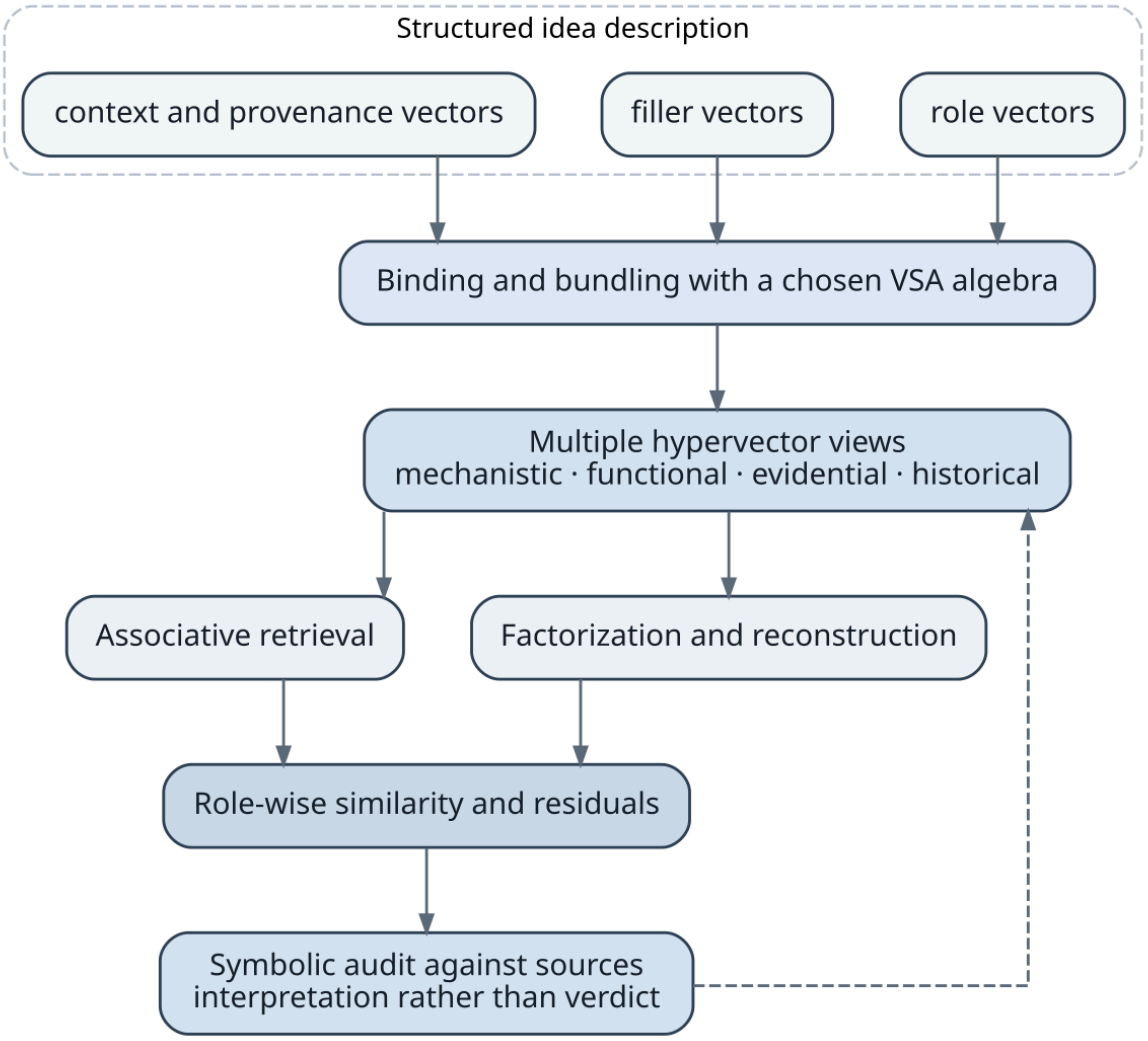


Figure 25. A proposed role for VSA/HDC. Structured descriptions are encoded into several hypervector views, used for retrieval and reconstruction, and then audited against the symbolic and documentary layers.

The figure emphasises two design choices. First, there is not one total hypervector. Mechanistic, functional, evidential, and historical views may use different role sets and codebooks. Second, the path ends in symbolic audit. A role-wise residual is a clue about where an alignment fails, not a direct declaration of novelty.

19.2 Role-Filler Binding and Structured Reconstruction

Let R_r denote a hypervector for role r and $F_r(I)$ a hypervector representing the filler of that role in idea object I. Using an abstract binding operator $\odot$ and bundling operator $\oplus$, a simple representation is

$$H(I) = \oplus_{r \in R}\left(R_r \odot F_r(I)\right).$$

The notation deliberately abstracts over the VSA family. In an HRR-like model, binding may use circular convolution and bundling vector addition. In a binary architecture, binding and bundling may use XOR and majority operations. The scientific claim is not tied to one choice. It is that binding can preserve the association between a role and its filler inside a distributed composite.

The role set R should reflect the operational representation. It may include goal, mechanism, assumptions, context, outcome, evidence, and provenance, but each can be decomposed. A mechanism view might bind entities, activities, temporal relations, feedback

paths, and causal dependencies. An evidential view might bind claim type, study design, measurement, result, and replication status. A lineage view might bind inherited components, transformations, source attestations, and dates.

A single bundled vector can become overloaded when many roles and nested structures are superposed. Multi-view representation addresses this problem and respects ontological pluralism. For each view v, define

$$H_v(I) = \oplus_{r \in R_v} \left(R_{v,r} \odot F_{v,r}(I) \right).$$

The idea object is then represented by a collection $\{H_v(I)\}_{v \in V}$ rather than by one universal vector. The views may share some atomic code vectors while preserving different bindings and resolutions. This makes query-relative identity natural: a mechanistic comparison weights the mechanism view; a priority inquiry emphasises historical and provenance views; an analogy query may compare relational structure while suppressing domain-specific fillers.

Multi-view encoding also mitigates a conceptual error common in embedding systems. If all information is compressed into one geometry, proximity on one dimension can dominate another. Two works with similar language and citations may appear close even when their mechanisms differ. Separate views permit late combination under an explicit question. The system can report that linguistic and functional similarity are high while mechanistic similarity is low.

Nested structure requires further operations. Order can be represented through permutation, position vectors, or recursive binding. Graph relations can be encoded by binding source, relation, and target roles. For a mechanism with feedback, the representation must distinguish a set of components from the organised cycle among them. VSA/HDC provides several constructions for sequences and graphs, but their capacity and similarity behaviour differ. The atlas should compare architectures empirically rather than assume that one encoding preserves every relevant relation (Kleyko et al. 2023; Schlegel, Neubert, and Protzel 2022).

Recovering a role filler from a composite usually involves unbinding. In abstract form, the estimate of the filler for role r is

$$\widetilde{F}_r(I) = R_r^{-1} \odot H(I),$$

where R_r^{-1} denotes the architecture-appropriate inverse or approximate inverse. Because bundling introduces cross-talk, $\widetilde{F}_r(I)$ is noisy. A cleanup memory compares it with a codebook of known fillers or candidate structures. This mechanism is useful for interpretation: a system can ask what mechanism is encoded in a retrieved idea or whether two candidates fill a role with compatible structures.

The cleanup step is also a source of epistemic risk. If the correct filler is absent from the codebook, the memory will return the nearest known item, making genuine novelty look like a distorted instance of the familiar. If the codebook is excessively large or poorly separated, recovery fails. If fillers are generated by a neural encoder rather than assigned randomly, learned geometry can help semantic generalisation but may reduce the quasi-orthogonality on which VSA operations depend. The interaction between learned semantic vectors and algebraic codebooks is a central research problem.

One possibility is a dual representation. Atomic symbols and stable ontology roles receive controlled code vectors; open-ended content receives learned vectors projected into the hyperdimensional space; and the link to the original source is preserved symbolically. Another is to learn codebooks under constraints that maintain binding recoverability and role separation. The relevant objective is not ordinary downstream accuracy alone. It must measure whether composition, unbinding, and factorisation remain reliable under the scale and ambiguity of scientific concepts.

Representation should include uncertainty. A role may have several candidate fillers with different weights, or its content may be partially unknown. Bundling can encode weighted

alternatives, but the resulting vector should not be mistaken for a probability distribution without calibration. A better design stores the alternative symbolic hypotheses and uses hypervectors to retrieve and compare them. The VSA layer can represent a superposition of possibilities; the probabilistic layer records what that superposition means.

Provenance requires special care. It is tempting to bind source identifiers directly into the idea vector. Doing so can improve retrieval of historically related items but contaminate semantic comparison: two identical mechanisms from different sources should not appear dissimilar merely because their provenance differs. Multi-view design solves this by keeping provenance and semantic structure separate, combining them only for questions that require both. This separation is another reason to reject the idea of a single canonical hypervector.

The most distinctive VSA/HDC research direction concerns reconstructability. If an idea is composed of known role–filler structures and transformations, its hypervector should be approximable by combining the corresponding prior hypervectors. A candidate that cannot be reconstructed may contain a new component or organisation. This suggests a computational proxy for one dimension of novelty.

The first step is candidate retrieval. Given $H_v(I)$, associative memory identifies earlier objects with similar functional, mechanistic, or contextual views. The second step searches for combinations and transformations that improve the match. The third attempts to factor the candidate into elements drawn from codebooks of prior components and transformation operators. Resonator networks were developed precisely to solve factorisation problems in distributed compositional representations by alternating VSA operations with associative cleanup (Frady et al. 2020; Kent et al. 2020). Their performance makes them relevant to a search over possible antecedent structures.

A reconstruction model might seek a prior component set S and transformation sequence T whose encoded result is close to the candidate. Rather than treating the global difference as novelty, it should unbind corresponding roles and compute role-specific residuals. Let $\hat{I} = T(S)$ denote the best reconstructed structure. For each role r, define

$$\rho_r\left(I, \hat{I}\right) = 1 - sim\left(\widetilde{F}_r(I), \widetilde{F}_r\left(\hat{I}\right)\right).$$

A high ρ_r indicates that the reconstruction does not explain the filler associated with role r. The residual profile $\rho\left(I, \hat{I}\right)$ is more informative than a single distance. It may show that the goal and evaluation are inherited, the mechanism is partly transformed, and one assumption has no close precedent.

The central scientific hypothesis can now be stated carefully: *under an adequate representation, codebook, corpus, and transformation grammar, persistent role-specific reconstruction residuals correlate with expert judgments of substantive novelty*. Every qualification is necessary. The hypothesis is empirical and may fail. It does not identify novelty with residual magnitude; it predicts a relationship after controlling for competing causes.

Factorisation faces combinatorial difficulty. An idea may draw components from many antecedents, and transformation sequences can proliferate. Resonator networks exploit superposition to search factor combinations, but they are not guaranteed to converge and have finite capacity (Kent et al. 2020). Scientific idea structures are also more open-ended than the controlled codebooks in many demonstrations. A realistic system needs strong retrieval priors, hierarchical search, and symbolic constraints to keep factorisation tractable.

The search space should include transformations as well as components. A candidate may be reconstructed from an earlier mechanism plus the operation “remove assumption A” and a measurement technique imported from another field. Transformation operators could be encoded as hypervectors or as symbolic programs whose effects are compiled into the VSA space. Recurrent transformations learned from curated lineages would constrain the search. The result would resemble a grammar of scientific change rather than a bag-of-ideas recombination engine.

Several residuals can arise without novelty. Missing literature is the most obvious. If the relevant precedent is absent from K_t, the reconstruction fails. A defective ontology can produce the same effect: the system may lack a role needed to align two formulations. Wrong granularity can split one component into many or collapse a decisive distinction. Neural extraction can misrepresent the candidate. Codebook capacity can be exceeded, producing cross-talk. The candidate itself may be incoherent. Each failure has a different diagnostic signature, but none is guaranteed to be separable from true novelty by the VSA layer alone.

Stability analysis is therefore mandatory. The system should repeat reconstruction under alternative views, dimensions, codebooks, corpus samples, and transformation grammars. A residual that moves unpredictably with random initialisation is not a robust novelty signal. A residual that disappears after a modest corpus expansion indicates coverage sensitivity. A residual localised to one role across architectures and confirmed by expert comparison is stronger evidence.

The MDL interpretation introduced earlier provides a conceptual bridge. Reconstruction cost reflects not only distance but the complexity of the antecedent explanation. A candidate exactly representable as a contrived combination of dozens of unrelated precedents may be less genuinely “anticipated” than one derived by a simple transformation from a single lineage. VSA factorisation can generate candidate decompositions, while a symbolic description language assigns interpretable costs. The hybrid system can prefer explanations that are both geometrically adequate and structurally economical.

This framework may also identify *negative novelty*: a claim presented as new whose structure is readily reconstructed from earlier work. The output should still avoid accusation. It should display the reconstruction, sources, and assumptions, allowing experts to judge whether the overlap is substantial. A low residual can coexist with novelty in evidence, implementation, or domain. The atlas analyses dimensions; it does not reduce scholarly credit to a vector distance.

19.3 Transformations, Analogy, and Failure Criteria

Analogy is central to scientific discovery because mechanisms often migrate across domains. Structure-mapping theory distinguishes relational correspondence from mere attribute similarity (Gentner 1983; Gentner and Markman 1997). An atlas should retrieve cases that share an organisation of roles even when their entities and vocabulary differ. VSA/HDC is promising because binding can represent roles independently of fillers, allowing queries that suppress domain-specific content and compare relational structure.

Suppose a system encodes two mechanisms with different entities but similar feedback organisation. By unbinding entity fillers and retaining role relations, it can form an abstract structural view. Similarity in this view proposes an analogy. The symbolic layer then checks whether causal dependencies, temporal order, boundary conditions, and failure modes align. The result is not “these ideas are the same,” but “this relational mapping may support transfer under specified assumptions.”

Transformations can themselves be compared. If A becomes B through a change in role structure, and C becomes D through a comparable change, the system may identify an analogy between the two transitions. Some VSA architectures support operations that approximate such relational differences, but their behaviour depends on the binding algebra (Schlegel, Neubert, and Protzel 2022). A safer approach represents the transformation explicitly as a structured object and encodes that object in a dedicated view. Similarity between transformation hypervectors then guides retrieval, while symbolic alignment verifies the correspondence.

This capability could be more valuable than direct idea similarity. Researchers often seek a way to repair a limitation rather than a duplicate of their problem. An atlas might retrieve

lineages in which an analogous assumption was removed, a feedback delay was handled, or a measurement bottleneck was overcome. The transfer remains speculative until domain constraints are checked, but the search is guided by the form of change rather than by shared terminology.

VSA/HDC also offers robustness to noisy and partial queries. A researcher may specify only a goal, a constraint, and a desired capability. Bundled representations can retrieve candidates matching subsets of those roles. Because similarity is distributed, the system can degrade gracefully when some information is missing. This is well suited to exploratory science, where the query itself may be incomplete.

The same properties create false positives. High-dimensional similarity can arise from shared common roles, and superposition can obscure which components produced the match. A ubiquitous pattern such as optimisation under constraints may connect many unrelated methods. Role weighting, inverse-document-frequency analogues for semantic components, and symbolic post-filtering are needed to prevent generic structures from dominating. Cross-domain analogy should be evaluated not only by inspiration but by structural fidelity and the rate of seductive but invalid transfers.

Capacity is a fundamental limit. Bundling too many associations increases noise. Deeply nested structures become difficult to decode. Factorisation requires finite codebooks and can converge to incorrect combinations. Larger dimensionality helps but does not eliminate these problems; it increases memory and computation and may postpone rather than solve open-world scaling. The relevant experiments must estimate capacity under realistic distributions of idea complexity, not only random synthetic vectors.

The codebook problem is equally fundamental. If atomic vectors are random, semantic similarity must be supplied through shared structure and learned mappings. If atomic vectors inherit neural geometry, binding operations may not have the desired algebraic properties. A hybrid codebook may work, but its stability across model versions and domains must be tested. The atlas cannot allow a change in the underlying language model to silently redefine historical identity.

Identifiability is a deeper problem than retrieval accuracy. Many combinations of role vectors, filler vectors, and binding conventions can produce similar global similarities. A successful match does not show that the recovered internal factors correspond to the scientifically intended decomposition. The system needs interventions on representation: hold the source evidence fixed, vary the codebook or algebra, and test whether the same role-level alignment survives. When several encodings are related by a harmless reparameterisation, conclusions should be invariant. When a conclusion changes because one architecture preserves order and another does not, the difference should be predicted by the theory rather than discovered as an unexplained benchmark fluctuation.

Learned codebooks introduce temporal drift. Updating a language encoder can move fillers, change neighbourhoods, and alter which historical objects are retrieved. One response is to anchor stable symbolic roles and canonical entities while learning only mappings for open-ended content. Another is to maintain explicit translation maps between codebook versions and to rerun calibration sets after every update. Either way, version identity becomes part of the provenance of a VSA judgment. A statement that two ideas were close under codebook V_3 is not automatically interchangeable with the same statement under V_7.

The VSA hypothesis must also be tested against alternatives designed for the same structural problem. Graph matching, tensor-product models, differentiable unification, probabilistic programs, and neural architectures with explicit slots may preserve role structure without relying on superposition and cleanup memory. VSA/HDC earns a place only if its particular combination of distributed robustness, algebraic binding, fixed-dimensional memory, and approximate factorisation produces advantages that these alternatives do not. Energy

efficiency or hardware suitability may matter for deployment, but they do not establish semantic adequacy.

These constraints clarify the possible scientific contribution. The strongest result would not be that hypervectors can encode an example. It would be a theory connecting properties of the binding algebra, dimensionality, codebook design, and structural depth to specific errors in identity, analogy, and reconstruction. Such a theory could explain when distributed symbolic memory preserves the invariants needed by the atlas and when it necessarily loses them. Even a negative boundary theorem or a carefully mapped failure regime would advance the problem more than an attractive demonstration without identifiability.

Table 37 frames the principal VSA/HDC claims as falsifiable hypotheses. The second column states observations that would weaken or reject each claim. This table is important because the chapter is proposing a research direction, not endorsing a technology in advance.

Table 37. VSA/HDC should enter the project through explicit hypotheses and failure criteria.

Research hypothesis	Evidence that would weaken or falsify it
Role-bound hypervectors improve identity judgments beyond strong text and graph embeddings.	Gains disappear on contrastive tests where surface similarity and mechanistic identity are deliberately separated.
Multi-view encoding preserves useful distinctions better than one total embedding.	View separation adds complexity without improving calibration, transfer, or interpretability.
Factorisation recovers meaningful antecedent components and transformations.	Reconstructions are unstable across dimensions and codebooks, or experts judge recovered components to be post-hoc artefacts.
Role-specific residuals correlate with substantive novelty.	Residuals mainly track missing corpus coverage, extraction error, or unusual wording after appropriate controls.
Transformation representations support cross-domain analogical retrieval.	Retrieved analogies are no more structurally faithful or useful than strong relational graph and language-model baselines.
The VSA layer can scale while preserving auditability.	Capacity limits, cleanup errors, or codebook drift make recovered structures too unreliable for source-grounded explanation.

The table guards against a common pattern in neuro-symbolic research: demonstrating that a representational formalism can encode a toy structure and then assuming that the encoding explains a complex epistemic phenomenon. The relevant question is comparative. Does VSA/HDC add measurable value over well-designed symbolic graphs, neural retrievers, and graph embeddings on tasks where compositional binding and factorisation should matter?

A negative result would still be scientifically valuable. It might show that explicit graph alignment is sufficient, that open-ended codebooks defeat factorisation, or that novelty is too dependent on representation for hypervector residuals to be stable. The project should be designed so that the resulting datasets, identity theory, and lineage benchmarks remain useful even if VSA/HDC contributes little. This is the correct relation between a conceptual programme and a technical hypothesis: the hypothesis is ambitious, but the programme is not held hostage to it.

Chapter 20. Testing the Architecture

The proposal is scientific only if temporal backtests, contrastive datasets, ablations, human studies, and simpler baselines can prove it wrong.

20.1 Building Evidence Without Presupposing the Answer

The first empirical challenge is to construct a corpus and annotation regime without turning the initial ontology into the conclusion. A dataset of “ideas” cannot be sampled in the same way as images of already named objects. Curators must decide where one contribution ends and another begins, which relations matter, and which sources count as evidence. These decisions should be treated as experimental material rather than hidden preprocessing.

A suitable corpus should begin with bounded scientific episodes. A good episode contains a tractable lineage, diverse source types, meaningful technical variation, and experts capable of reconstructing the history. Examples might include a family of optimisation methods, an experimental technique that migrated across disciplines, a causal-inference design, or a scientific concept whose terminology changed. The first objective is not representativeness of all science. It is enough variation to stress the theory of identity while preserving the possibility of close verification.

The corpus must include more than canonical papers. Reviews and textbooks reveal how fields retrospectively organise a lineage. Code and protocols reveal operational details absent from articles. Older books, proceedings, theses, patents, and archives may contain earlier attestations. Negative results and failed applications expose conditions of validity. Multilingual material is essential whenever priority or conceptual migration crosses linguistic traditions. Source selection should be documented as part of the dataset, including known gaps.

Annotation should proceed from evidence to interpretations. Curators first identify addressable attestations: passages, equations, diagrams, code fragments, or artefacts. They then propose role structures and relation claims. This order prevents the abstraction from losing its documentary anchor. Each annotation records the task and granularity at which it was produced. An expert may annotate the same passage differently for a mechanism comparison and for a historical study; both can be valid.

The protocol should collect alternative decompositions rather than force agreement too early. Inter-annotator disagreement can reveal ambiguous boundaries, ontology defects, or genuine pluralism. Conventional agreement statistics remain useful, but low agreement should not automatically be treated as annotator failure. The dataset needs a disagreement taxonomy distinguishing accidental inconsistency from different legitimate models. Adjudication should record the reasons for preferring one view in a particular benchmark.

Contrastive examples are more valuable than large quantities of easy positives. The corpus should include pairs with similar wording but different mechanisms, different wording but equivalent functions, shared mechanisms with independent histories, and apparent novelty caused by missing sources. It should also include transformations that alter only one role. These cases test whether a system has learned typed identity rather than generic semantic proximity.

Multi-resolution annotation is essential. Curators can provide a coarse operational object, a finer decomposition, and relations between them. The dataset then

supports experiments on whether a model chooses an appropriate level for the question. It also enables evaluation of stability: does a claimed lineage survive refinement, or was it an artefact of coarse representation?

Historical claims require a different evidence protocol from semantic ones. An influence annotation should cite documentary or contextual support, not merely a similarity judgment. A priority annotation should describe the corpus searched and the dating basis. Independent rediscovery should remain provisional when transmission evidence is incomplete. These requirements make annotation slower, but they are necessary if the atlas is to support more than semantic search.

Existing resources provide valuable starting points. IdeaGene-Bench supplies typed objects and lineage differences across ten domains (Zhou et al. 2026). RINoBench provides expert judgments of research idea novelty (Schopf and Färber 2026). CHIMERA supplies mined recombination cases (Sternlicht and Hope 2026). Scideator and literature-grounded novelty work provide facet structures and retrieval protocols (Radensky et al. 2024; Shahid et al. 2025). The atlas research programme should reuse these resources while adding source-addressed evidence, alternative decompositions, typed uncertainty, and temporal corpus boundaries.

The resulting dataset should be regarded as a set of arguments, not only labels. A benchmark instance includes the source material, the operational representations, the relation under test, the supporting and opposing evidence, and the adjudication context. This richer form makes evaluation more demanding but reduces the risk that systems optimise against unexplained judgments.

A broad end-to-end demonstration cannot reveal which part of the atlas is scientifically responsible for an improvement. The research programme needs a family of benchmarks aligned with the conceptual distinctions developed in earlier chapters. Each benchmark should compare against strong document-level, embedding, graph, and language-model baselines.

Idea decomposition tests whether a system can propose useful role structures from sources. Evaluation should measure span grounding, role correctness, structural relations, and the ability to represent alternatives. Exact graph match is informative but too strict when several decompositions are legitimate. Pairwise expert comparison and task performance can supplement it.

Typed identity tests whether the system distinguishes paraphrase, functional equivalence, mechanistic equivalence, analogy, and lineage. The strongest cases hold topical similarity constant while changing one relation. Evaluation should score the relation, the role alignment, and the evidence cited. Calibration matters because ambiguous cases should receive distributions or alternatives rather than confident labels.

Precedent retrieval tests whether structured representations recover relevant earlier work across vocabulary and disciplinary boundaries. The benchmark must freeze the search corpus at a date and include hard negatives that are semantically close but structurally irrelevant. Recall is important because missing a decisive precedent invalidates later novelty analysis. Evidence precision is equally important because a long list of vaguely related papers does not help a researcher.

Lineage reconstruction tests inheritance and transformation. Systems receive a set of works and must identify preserved components, changes, imports, losses, and plausible transmission paths. Structural accuracy and temporal consistency can be

measured automatically where annotations are strong. Historical influence should be evaluated separately from semantic lineage because the evidential standards differ.

Novelty analysis tests whether the system produces a stable and evidence-grounded profile rather than a binary score. Evaluation should ask whether the closest precedents are found, whether the role-wise differences are correct, whether uncertainty reflects corpus limitations, and whether the conclusion changes appropriately when earlier evidence is added. RINoBench shows why rationale quality and judgment accuracy must be evaluated separately (Schopf and Färber 2026).

Analogy retrieval tests relational transfer. Experts assess whether the mapping preserves causal or functional structure, whether domain incompatibilities are identified, and whether the analogy generates a testable direction. Diversity alone is not enough. A system can produce distant but scientifically empty associations. The benchmark should penalise analogies whose persuasive language exceeds their structural support.

Table 38 links each task to the scientific claim it is intended to test. This prevents benchmark proliferation from becoming disconnected from theory.

Table 38. Benchmarks should correspond to explicit scientific claims rather than to an undifferentiated demonstration.

Evaluation task	Scientific question tested
Evidence-grounded decomposition	Can operational structures be proposed reproducibly without losing their source basis or suppressing legitimate alternatives?
Multi-resolution representation	Does the system choose and relate granularities appropriate to different inquiries?
Typed identity	Can it distinguish functional, mechanistic, propositional, implementation, analogical, and historical relations?
Cross-vocabulary precedent retrieval	Does explicit structure find relevant antecedents that document similarity and citations miss?
Lineage difference reconstruction	Can inheritance, transformation, import, and loss be localised to specific roles and supported by evidence?
Priority under corpus constraints	Does the system identify earlier attestations while reporting dating and coverage uncertainty?
Structured novelty analysis	Do residuals and differences align with experts and remain stable under corpus and representation changes?
Cross-domain analogy	Can relational structure transfer without uncontrolled false positives or erased boundary conditions?
Human decision support	Do researchers reason more accurately or efficiently without becoming over-reliant on system authority?

Ablation is central. The atlas should be compared with document retrieval augmented generation, dense embeddings, citation graphs, structured symbolic graphs without VSA, VSA without symbolic audit, and the full hybrid. Within VSA/HDC, experiments should vary architecture, dimension, codebook design, number of bundled roles, learned versus controlled vectors, factorisation method, and multi-view separation. Within the symbolic layer, they should vary ontology rigidity, transformation grammar, and treatment of disagreement.

The most informative failures may occur when the system is asked to generalise. A representation learned in machine learning could be tested on biology or materials science. A codebook trained on contemporary English could be tested on historical and multilingual sources. A lineage model could be tested on independent rediscoveries rather than direct citation chains. These transfers reveal whether the system has captured operational structure or only domain conventions.

Evaluation should include adversarial cases. Authors can create pseudo-novel proposals by renaming established mechanisms, combining unrelated components, or adding superficial complexity. They can also produce genuinely important local changes that appear semantically close to prior work. A useful atlas should resist both novelty inflation and novelty suppression.

20.2 Temporal Backtesting and Conditions of Falsification

The strongest evaluation of scientific support is temporal. Freeze the accessible corpus at time t, prevent the system from seeing later work, and ask it to analyse ideas that emerge after t. The later literature then supplies a partial record of what was independently developed, validated, abandoned, or recognised. This does not turn history into a perfect ground truth, but it is more demanding than evaluating against contemporary labels using contemporary knowledge.

Figure 26 presents the design. Competing systems receive only the frozen corpus. They perform blinded tasks such as precedent search, identity analysis, analogy, and novelty assessment. Later literature and outcomes are then revealed. Evaluation combines expert judgment, calibration, later validation, and characteristic failure analysis. The objective is not simply to predict which paper will become highly cited. It is to test whether the atlas constructs better evidence and more useful distinctions from the information historically available.

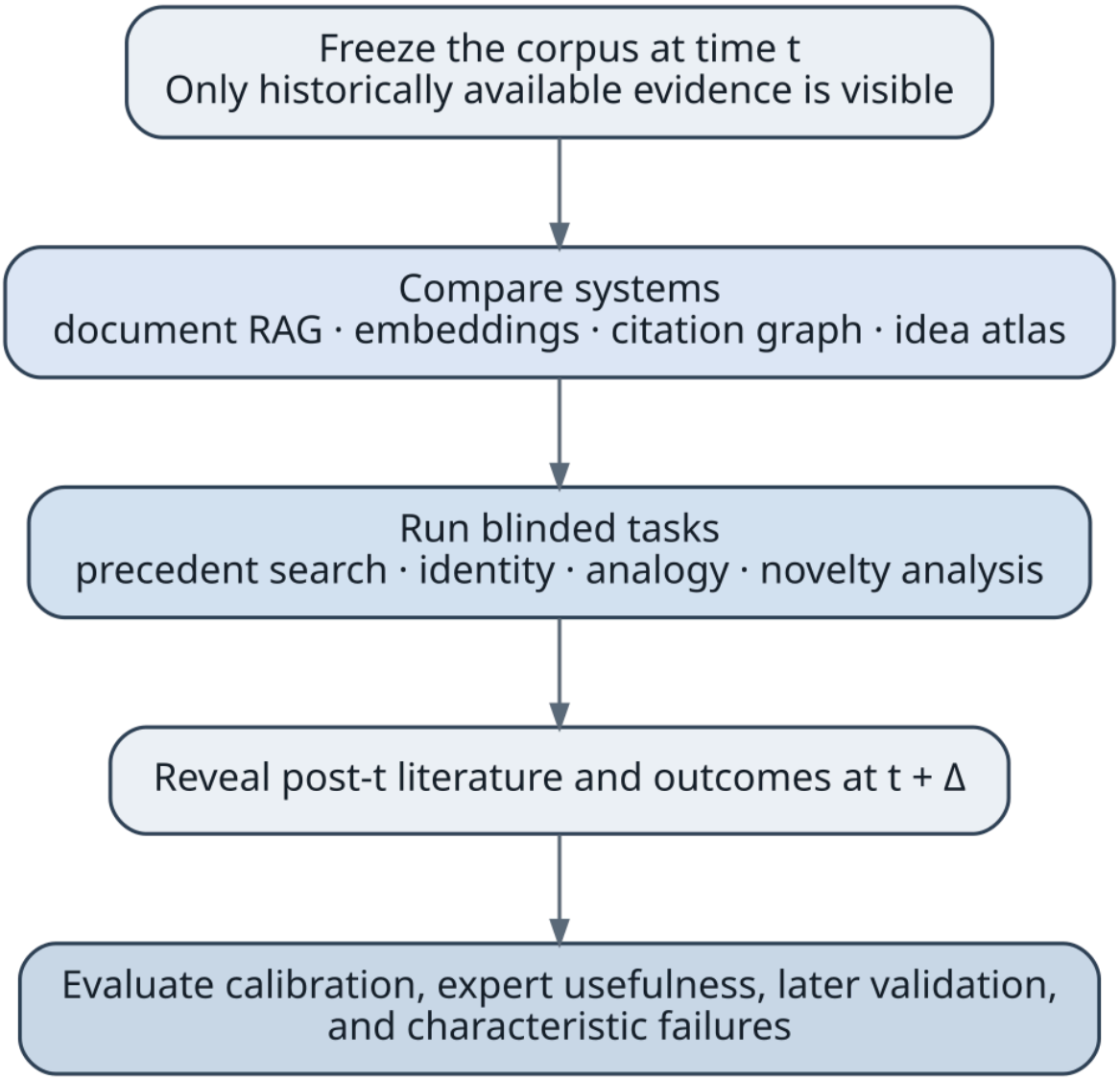


Figure 26. Temporal backtesting evaluates an atlas using only the knowledge available at a historical cutoff, then compares its analyses with later literature and expert review.

A temporal backtest can ask whether the system retrieves the antecedents later acknowledged by a field, whether it identifies an analogy that subsequently becomes productive, or whether a novelty residual corresponds to a mechanism that later receives independent support. It can also reveal false novelty caused by incomplete indexing and false analogies that later fail. Because later success is socially contingent, outcome measures should be plural: validation, replication, capability, explanatory uptake, and expert retrospective assessment.

Human studies are required because the atlas is intended to alter reasoning, not only predict labels. Researchers can be assigned literature synthesis, novelty evaluation, or hypothesis-generation tasks with different tools: ordinary search, document RAG, citation graphs, and the idea atlas. Outcomes include source coverage, correctness of distinctions, time, calibration, diversity of useful directions, and ability to detect system errors. Over-reliance must be measured explicitly by planting plausible but unsupported relations.

Expert panels should be blinded to system identity where possible. They should assess structured outputs rather than only final prose. For novelty analysis, experts can

inspect retrieved precedents and role-wise differences. For analogy, they can rate mapping fidelity and testability. For lineage, historians and domain experts may need separate roles because technical equivalence and influence evidence require different expertise.

The programme must pre-register failure criteria. The identity hypothesis is weakened if multi-view structured representations do not outperform strong embeddings on controlled contrastive cases. The lineage hypothesis is weakened if evidence-grounded graphs do not improve reconstruction beyond citation and text baselines. The novelty hypothesis is weakened if residuals are unstable under reasonable changes of corpus and representation or mainly predict unusual wording. The VSA hypothesis is weakened if symbolic graphs supply the same gains with less noise and greater auditability.

A stronger failure would be institutional. Even accurate outputs may not improve research if they are too complex, slow, or difficult to challenge. Researchers may ignore uncertainty and treat ranked precedents as definitive. Curators may be unable to maintain the ontology. Coverage bias may dominate the results. These are not secondary deployment issues; they test whether the epistemic model can function in real scientific communities.

Negative results should be publishable and reusable. If operational decomposition proves reliable but VSA factorisation fails, the dataset and identity theory remain contributions. If the atlas improves review but not ideation, that finding narrows the impact claim. If document-level systems match the hybrid on all tasks, the proposed epistemic layer may not justify its complexity. A rigorous programme is one in which each outcome teaches us something about the nature and representation of ideas.

The research agenda is therefore ambitious but falsifiable. It does not ask whether a system can produce an impressive genealogy for selected examples. It asks whether explicit idea structure yields reproducible gains, whether the gains survive domain and time shifts, whether uncertainty behaves appropriately, and whether humans reason better with the instrument. Only after these conditions are met would claims about accelerating science be warranted.

20.3 A Frontier Research Agenda

The future of machine discovery is likely to be collective rather than solitary. Humans, models, instruments, databases and institutions contribute different forms of search and judgment. Machines offer scale, memory, simulation and rapid iteration. Humans contribute shifting values, embodied context, conceptual reframing and responsibility. The objective is not to assign every function permanently, but to design interfaces through which strengths combine without making failures invisible.

Collective intelligence requires diversity. If every laboratory uses the same foundation model, retrieval index and evaluator, individual productivity may rise while the scientific portfolio narrows. Hao and colleagues' 2026 analysis reported this tension at scale: artificial-intelligence use was associated with greater individual impact and productivity but a contraction in collective focus. A system-level research agenda should therefore measure topic diversity, methodological diversity, institutional access and concentration of epistemic infrastructure.

Table 39. A frontier research agenda for machine-assisted discovery

Research direction	Decisive evidence of progress
Unknown-aware learning	Reliable abstention and escalation across predeclared and newly encountered shifts.
Causal concept formation	Latent variables that support intervention and transfer in real physical testbeds.
Open-ended search	Sustained creation of useful new niches without collapse into trivial variation.
Scientific process learning	Improved evidence uptake, falsification seeking and belief revision, not only final scores.
Verifier construction	Cheaper independent checks that correlate prospectively with real-world validity.
Collective diversity	Higher productivity without contraction of questions, methods or participating institutions.

The table exists to make the agenda falsifiable. Each direction has a failure mode that ordinary benchmarks may hide. Unknown-aware learning must be tested on shifts not used to tune the detector. Causal concepts must survive intervention. Open-ended search must generate useful diversity rather than decorative variation. Scientific process learning must change how evidence affects action. Collective diversity must be measured at the field level.

One promising interface is disagreement routing. Several models, simulations or experts analyse a claim. Agreement on routine cases enables automation. Structured disagreement triggers stronger evidence or broader review. The system does not treat consensus as truth; it uses disagreement as a signal about model sensitivity and missing assumptions. Diversity must be substantive, based on different data, methods or perspectives, not merely different role prompts applied to the same model.

Another interface is question markets or portfolios. Agents generate candidate questions and estimate tractability, significance, cost and neglectedness. Humans and institutions allocate resources across a diverse portfolio rather than choosing only the highest predicted return. The mechanism can reserve capacity for high-uncertainty questions, replication and tool building. This protects science from becoming a narrow optimization process.

Education and access are part of the frontier. Scientific agents can help researchers use advanced statistics, formal methods and laboratory automation. They can translate across disciplines and languages. But dependence on proprietary infrastructure can exclude institutions and reduce auditability. Open standards, shared benchmarks, public verifiers and local deployment options will influence which communities can participate in machine-assisted discovery.

Governance should follow capability at the level of action. Literature synthesis needs citation and privacy standards. Code execution needs sandboxing. Laboratory control needs physical safety and permissions. Self-improvement needs regression tests and immutable logs. Publication agents need disclosure and accountability. Broad principles become effective only when attached to specific interfaces and evidence records.

Research should also study failure ecology. When many agents interact, errors can spread through citations, generated datasets and shared evaluators. A false but

plausible claim can be repeated until retrieval systems treat it as consensus. Provenance, source diversity and correction propagation are therefore central capabilities. Scientific memory must be able to forget authority while preserving history.

The most interesting novelty may be institutional rather than algorithmic: new forms of peer review, distributed replication, machine-readable argument and dynamic allocation of experimental capacity. These can change what questions are affordable and whose ideas receive evaluation. Machine learning becomes an engine of science when it reorganizes the cost structure of criticism, not only the cost of generation.

The ultimate frontier is a discovery system that can surprise us, show us why the surprise matters, design the test that could defeat it, and preserve the entire path for others to challenge.

No present system satisfies that description generally. Many components already exist in narrow form. The research programme is to connect them without erasing their boundaries: generators that remain separate from judges, world models corrected by experiments, memories that preserve contradiction, agents constrained by institutions and human communities capable of redefining significance. This is a realistic path toward extending both science and the range of work computers can automate.

A general discovery architecture should be modular enough to expose failure and integrated enough to learn from it. The blueprint below is not a claim that every domain needs the same implementation. It identifies functions that should be assigned deliberately rather than hidden inside a conversational agent.

Table 40. Functional modules of a discovery system

Module	Required responsibility
Problem and claim model	Represent the question, target, assumptions, equivalence rules and acceptable evidence.
Generator	Produce diverse candidates with lineage and declared dependencies.
Search controller	Allocate resources across performance, uncertainty, diversity and model criticism.
Evaluator ladder	Combine learned filters, execution, formal checks, simulation, experiments and review.
Experiment planner	Choose safe, discriminating observations under cost and risk constraints.
Layered memory	Preserve episodes, claims, methods, alternatives, failures and governance history.
Revision engine	Update confidence, representations and objectives while keeping versions reversible.
Provenance and governance	Log actions, enforce permissions, disclose automation and assign responsibility.

The table exists because “agent” is not a sufficient architectural description. A system can appear coherent while the same model generates, judges and summarizes every result. Explicit modules make epistemic independence testable. They also allow stronger tools to replace weaker ones without rebuilding the entire system.

The evaluator ladder should dominate deployment design. Cheap evaluators filter the broad candidate space; strong evaluators are reserved for disputed or high-value cases. Feedback should include counterexamples and failure causes, not only scalar rewards. The archive should preserve candidates across behavioural niches so that later evidence can reactivate them.

The memory system should distinguish immutable records from revisable beliefs. Raw data, instrument logs and executed code should remain fixed. Claims and confidence should change through versioned updates. Procedures should be executable and reviewed. Governance records should reveal who or what authorized each consequential action.

Safe autonomy should begin narrow: broad simulation, but approval for physical action, publication and self-modification. Expansion requires demonstrated reliability.

PART V: DIFFERENT SCIENCES, DIFFERENT RESISTANCE

Automation changes form across domains because proof kernels, software tests, instruments, organisms, patients, institutions, and interpretation provide different kinds of resistance.

Chapter 21. Equations, Algorithms, and Proofs

Formal and computational domains support strong local autonomy because candidates can often be executed or checked by narrow trusted evaluators.

21.1 Symbolic Regression and Equation Discovery

Symbolic regression searches for mathematical expressions that fit data. Unlike ordinary regression, the model class includes variable functional forms, operators and compositions. Early genetic-programming systems explored expression trees. Sparse identification of nonlinear dynamics selects terms from a library of candidate functions (Brunton, Proctor and Kutz, 2016). AI Feynman combines dimensional analysis, separability and neural approximation to recover formulas from synthetic data (Udrescu and Tegmark, 2020). More recent systems use neural networks to guide symbolic search.

Equation discovery is appealing because expressions are compact, executable and often interpretable. But a good fit is not a law. Many expressions interpolate finite noisy data, and symbolic search can produce complicated formulas that exploit sampling artefacts. Complexity penalties, dimensional consistency, symmetry, conservation laws and held-out regimes provide inductive bias. The difficult question is whether the recovered variables and functional forms correspond to stable mechanisms.

Table 41. Tests that separate a fitted expression from a candidate scientific law

Test	What it probes
Out-of-range prediction	Whether the expression extrapolates beyond the sampled parameter region.
Dimensional consistency	Whether units constrain or invalidate the proposed relation.
Symmetry and invariance	Whether the expression respects known transformations and conserved structure.
Intervention response	Whether changing variables produces the consequences predicted by the expression.
Parameter stability	Whether coefficients remain stable across datasets, sites and measurement protocols.
Mechanistic compression	Whether the expression replaces exceptions with a simpler reusable account.

The table exists because symbolic form can create an illusion of explanation. A short equation looks scientific even when it is a convenient curve. The tests ask whether the expression survives changes that were not optimized directly. Dimensional and symmetry constraints are especially valuable because they reject large classes of candidates before expensive experiments. Intervention response connects the equation to causality rather than correlation.

Physics-informed machine learning embeds differential equations, conservation laws or boundary conditions into training (Raissi, Perdikaris and Karniadakis, 2019; Karniadakis et al., 2021). This can reduce data needs and improve extrapolation when the physics is correct. It can also hide model misspecification. If the governing equations omit a process, enforcing them strongly can prevent the learner from

recognizing the anomaly. Discovery systems need a mechanism for relaxing prior laws when residual structure repeatedly contradicts them.

A promising hybrid alternates between flexible prediction and symbolic compression. A neural model captures complex patterns, attribution or sensitivity analysis proposes candidate variables, and symbolic search finds compact expressions. Experiments then target regions where the symbolic and flexible models disagree. The neural model supplies reach; the symbolic model supplies a testable abstraction. Neither is assumed to be final.

Variable discovery is the deeper frontier. Most symbolic-regression systems receive the relevant coordinates. Historical scientific advances often introduced new state variables, latent quantities or transformations. A system may need to infer that ratios, conserved combinations, phases or order parameters are the natural objects. Representation learning, dimensional analysis and causal interventions can jointly search this space, but identifiability remains conditional.

Noise and measurement error also matter. If explanatory variables are noisy, ordinary regression can bias coefficients. Hidden variables can produce apparent nonlinearities. Experimental design should therefore be integrated with equation search. The system can propose measurements that distinguish candidate terms, estimate whether additional precision is worth the cost, and detect when no expression in the current grammar is adequate.

The evaluator for an equation should include more than error and length. It can test numerical stability, units, invariances, parameter sensitivity, causal predictions and compatibility with known limits. A Pareto archive can preserve several expressions that trade simplicity against fit. Human scientists can then inspect a structured frontier rather than one supposedly optimal formula.

Equation discovery also needs an equivalence theory. Two expressions may differ syntactically while representing the same relation after algebraic simplification, a change of coordinates or a rescaling of units. Without canonicalization, a search system can report rediscoveries as diversity; with overly aggressive canonicalization, it can collapse formulations whose domains or singularities differ. The archive should therefore record transformations that preserve meaning, assumptions under which equivalence holds and counterexamples outside those assumptions.

A practical test is extrapolation by regime rather than by random split. A system can fit low-amplitude behaviour and then be tested near a transition, learn one geometry and face another, or be denied a known limiting case. These evaluations reveal whether an expression captured a mechanism or merely interpolated the sampled manifold.

Symmetry can make the search more scientific. Translation, rotation, permutation and conservation symmetries restrict the admissible form of a law and suggest which variables should be grouped. Conversely, a reproducible symmetry violation can signal new physics or a missing interaction. A frontier system would search jointly over expressions, variables, symmetries and experiments. It would ask which measurement most sharply separates rival equations, then update the symbolic archive without discarding alternatives that remain empirically indistinguishable. In that architecture, symbolic regression is not a final curve-fitting step. It is one component in a cycle that compresses evidence, exposes assumptions and proposes discriminating interventions.

A discovered equation is valuable not because it is symbolic, but because its symbols organize interventions, limits and explanations that survive new evidence.

The likely advance is not a universal equation finder. It is a family of domain-grounded systems that combine learned representations, symbolic grammars, physical constraints and active experiments. These systems may expand the range of mechanisms that can be extracted from complex data while making their assumptions unusually explicit.

21.2 AlphaTensor, FunSearch, AlphaProof, and AlphaEvolve

Mathematics and algorithms provide some of the clearest demonstrations that machine learning can produce results not explicitly present in its training data. The reason is evaluative strength. A candidate algorithm can be executed and its cost measured. A combinatorial construction can be checked. A formal proof can be verified by a small trusted kernel. These domains allow generators to be adventurous because acceptance does not depend on the generator's own confidence.

AlphaTensor reframed matrix-multiplication algorithm discovery as a game over tensor decompositions and used reinforcement learning to search the space. It found algorithms matching or improving known results for several sizes, including a 47-multiplication construction for a restricted arithmetic setting where two levels of Strassen's method use 49 (Fawzi et al., 2022). The scientific importance was not that a neural network "understood matrices" in the human sense. It was that a formal problem had been converted into a searchable environment with exact feedback.

FunSearch combined a pretrained language model with an evaluator and an evolutionary archive. The model generated program fragments; programs were executed; high-scoring and diverse candidates were retained and used to prompt further generations. The system produced improved constructions in combinatorics, including the cap-set problem, and useful bin-packing heuristics (Romera-Paredes et al., 2024). The central innovation was not free-form language generation but the conversion of language-model proposals into an iterative, executable search process.

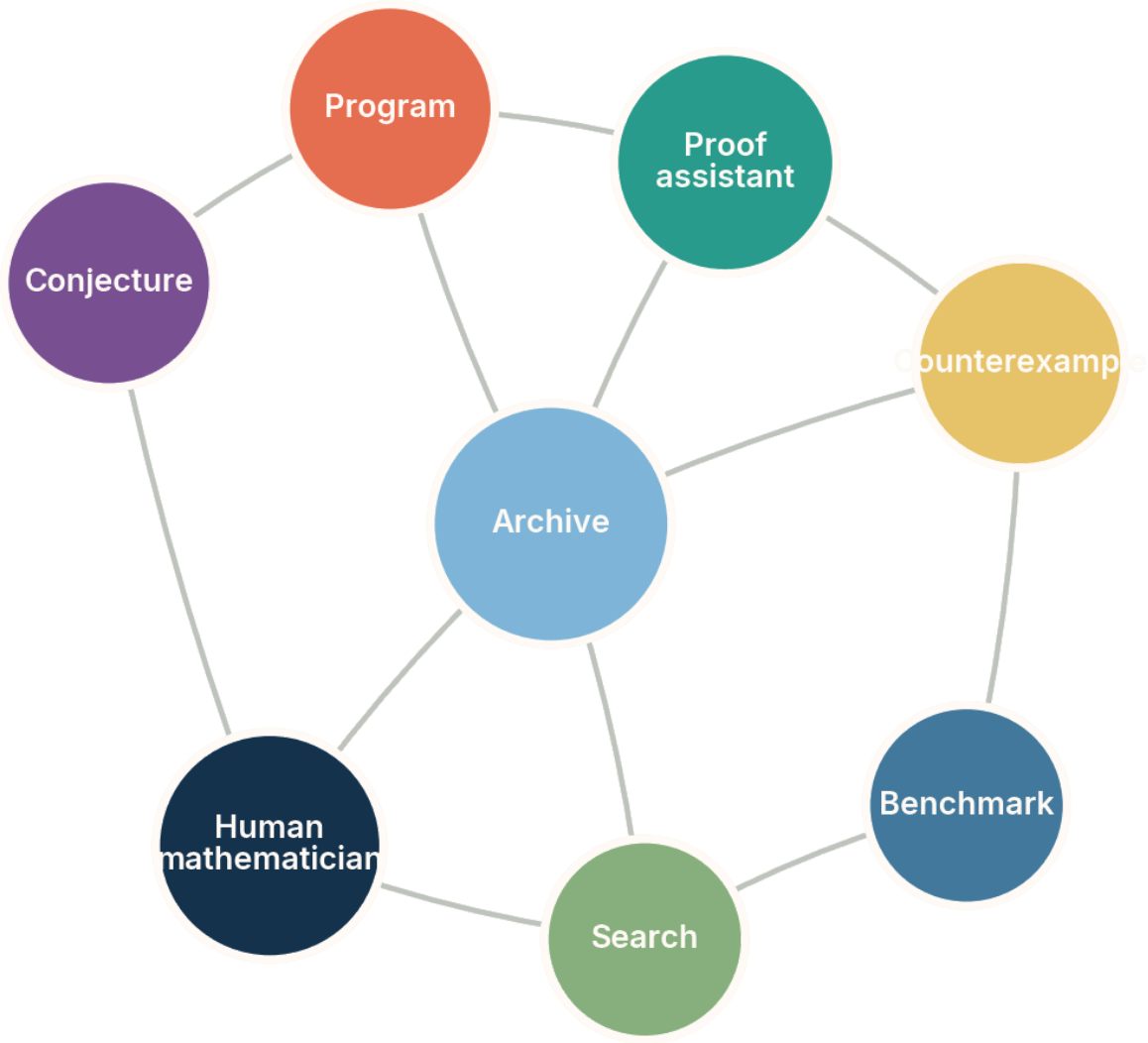


Figure 27. Machine-assisted mathematical discovery links conjectures, programs, proof assistants, counterexamples, benchmarks, search, human mathematicians and archives. Conceptual diagram; not an empirical result.

The figure matters because mathematical discovery is broader than proof generation. Programs can reveal patterns; counterexamples refine conjectures; benchmarks measure search; archives preserve constructions; proof assistants certify formal steps; human mathematicians choose definitions and interpret significance. The most productive systems connect these roles. A proof without an intelligible conjecture may be correct but unilluminating, while an empirical pattern without proof remains provisional.

AlphaProof pushed formal theorem proving by combining a language model with reinforcement learning inside the Lean proof environment. The 2025 Nature paper reported silver-medal-level performance on the 2024 International Mathematical Olympiad, with AlphaProof proving three of five non-geometry problems and a geometry system handling another (Hubert et al., 2025). The International Mathematical Olympiad (IMO) acronym is introduced here because it recurs in this chapter. The proofs were formally checkable, but formalization, computation time and domain coverage remained substantial constraints.

AlphaEvolve generalized the generate–evaluate–evolve pattern to code. The 2025 white paper described improvements in data-centre scheduling, hardware circuits and mathematical constructions, including a procedure for multiplying two four-by-four complex matrices using 48 scalar multiplications (Novikov et al., 2025). Independent mathematical work subsequently transformed the construction into a rational-coefficient algorithm over suitable rings (Dumas, Pernet and Sedoglavic, 2025). The result is compelling, but evidence levels should remain clear: AlphaEvolve was initially reported in a system white paper, while later mathematical analysis provides separate validation.

Table 42. Why four discovery systems succeeded

System	Key source of reliable progress
AlphaTensor	An exact game formulation and direct verification of tensor decompositions.
FunSearch	Executable programs, scalar evaluation and an evolutionary archive of diverse candidates.
AlphaProof	Reinforcement learning guided by a formal theorem prover with a trusted checking kernel.
AlphaEvolve	Code mutation, automated evaluators and population-based search across many domains.
Symbolic regression systems	Compact executable expressions constrained by fit, simplicity and domain structure.
Human–machine collaboration	Problem formulation, equivalence recognition, interpretation and proof or experiment design.

The table exists to identify the transferable pattern. The systems differ in architecture, but all convert an open-looking intellectual task into a space where many candidates can be rejected cheaply and surviving candidates can be checked independently. The lesson is not that every science should become a game. It is that scientific automation advances when evaluation can be formalized without erasing the phenomenon of interest.

There are important limits. Formal proofs depend on the axioms and definitions supplied. Program search often optimizes a scalar objective that may omit elegance, generality or implementation cost. Search can rediscover known constructions because prior-art comparison is difficult. Compute budgets can be large, and success distributions are uneven. Competition problems are selected for crisp statements and do not represent all mathematical research.

Current frontier work combines large-scale exploration with proof generation and human interpretation. A 2025 preprint on mathematical exploration at scale reported applying AlphaEvolve to dozens of problems, often rediscovering best-known constructions and sometimes improving them, then combining the search with reasoning and proof tools (Georgiev et al., 2025). Such reports are promising but should be read as evolving research rather than settled capability.

The deeper opportunity is interactive mathematics. A system can maintain families of conjectures, generate examples and counterexamples, translate informal claims into formal language, search proofs, and expose the minimal lemmas on which success depends. Human mathematicians can then work with a navigable space of

evidence rather than a single answer. This may change which questions are tractable even before it automates the highest-level choice of questions.

21.3 Strong Oracles and Fragile Formalizations

Automation advances fastest where a proposal can be tested by a strong, inexpensive, executable oracle. Formal mathematics, algorithm design, software engineering, and portions of machine learning offer the clearest examples. A proof checker can reject an invalid derivation. A compiler can reject malformed code. A test suite can expose regressions. A programmatic evaluator can score an algorithm over many instances. These domains are therefore likely to produce the earliest systems with substantial local autonomy.

The advantage should not be overstated. An oracle is strong only for the property it actually checks. A theorem prover can certify that a formal statement follows from axioms, but it cannot determine by itself whether the statement faithfully represents the informal conjecture or the intended application. A test suite can show that code passes selected cases, not that the requirements are complete. A benchmark can measure performance on a dataset while failing to capture robustness, cost, fairness, or operational relevance. The central methodological risk is specification error: the system becomes extremely reliable at solving the wrong formal problem.

Formal mathematics therefore requires a dual chain of rigor. The first chain is syntactic and deductive. Definitions, lemmas, theorem statements, proof terms, and trusted kernels must be versioned and checkable. The second is semantic and interpretive. The mapping from the human conjecture to formal objects must be reviewed, and the scope of the theorem must be communicated without expansion. A model can help translate informal arguments into Lean, Isabelle, Coq, or another proof assistant, but the translation itself is a claim. The research object should preserve the original statement, formalization decisions, unresolved mismatches, and the exact proof environment.

AI-assisted theorem proving is most credible when the generative model is a proposer under a small trusted kernel. The model can search for lemmas, tactics, constructions, and counterexamples. Failed proposals are cheap. The kernel determines formal validity. Human mathematicians remain essential in selecting important conjectures, judging explanatory value, designing representations, and verifying that formal statements capture the intended concepts. A proof can be correct and scientifically unilluminating. Executable science should make this distinction visible rather than allowing formal success to stand in for significance.

Program search follows a similar pattern. FunSearch and AlphaEvolve demonstrate that language models can generate program variants while evaluators and evolutionary selection provide external pressure (Romera-Paredes et al., 2024; Novikov et al., 2025). This architecture is powerful when the objective is executable and relatively complete. Its methodological obligations include evaluator validity, test-set separation, resource accounting, and analysis of whether the discovered program exploits an implementation detail. A candidate should be tested across distributions, scales, and independent implementations. Where possible, formal analysis should complement empirical score.

Software research adds a second layer: implementation fidelity. A paper may describe one method while the code implements another. Dependencies, random

seeds, hardware, numerical precision, preprocessing, and data leakage can alter results. Executable rigor therefore requires environment capture, build reproducibility, tests tied to specification clauses, and machine-readable experiment manifests. Continuous integration should rerun relevant analyses when code or data change. A claim-evidence graph should distinguish successful execution from agreement with the described method and from external reproduction.

The rise of coding agents offers a preview. Agents can navigate repositories, implement features, run tests, and revise code over long trajectories. Yet PaperBench shows that end-to-end replication remains difficult even in a fully digital field, and developer-productivity studies show that subjective speed does not always match measured completion time (Starace et al., 2025; METR, 2025, 2026). Scientific coding is harder than ordinary feature work because the specification may be incomplete and the code is itself part of the evidence. The workbench must expose agent actions as reviewable changes rather than treating a passing run as proof of scientific fidelity.

Machine learning requires especially careful rigor because benchmark feedback is rapid and optimization pressure is intense. A valid research object should identify data lineage, train-validation-test boundaries, search budgets, model and library versions, compute, hyperparameter policy, stopping rules, uncertainty across seeds, and all baselines. It should record failed runs and the extent of researcher or agent adaptivity. Claims about generalization require distributional tests; claims about mechanisms require interventions or causal analyses; claims about efficiency require matched hardware and quality; claims about safety require adversarial and worst-case evaluation.

Automated ML research systems can generate many variants and papers, which makes selective reporting a central risk. The workbench should treat every experiment generated under the active specification as part of the portfolio, even if only a subset appears in the article. This does not require publishing every transient run, but it requires accounting for the search process so that apparent significance is not interpreted as if one hypothesis had been tested once. Test-time compute used to generate and select research ideas should be reported as part of the method.

Formal methods can also strengthen agentic research infrastructure itself. State machines can constrain tool use. SMT solvers can check resource and dependency conditions. Type systems can enforce claim and evidence schemas. Temporal logic can express safety properties for laboratory controllers. Property-based testing can generate edge cases from specifications. These techniques should replace LLM judgment where their semantics are adequate. Their inclusion is not a retreat from intelligence; it is an allocation of intelligence to the layer where ambiguity actually remains.

The probable future in formal and computational sciences is high local autonomy inside strong verification envelopes. Agents will propose proofs, algorithms, code, and experiments. Pipelines will execute and score them. Independent systems will reproduce selected results. Humans will increasingly work at the levels of problem choice, formalization, evaluator design, explanatory synthesis, and release. The limit will not be the generation of candidate artefacts. It will be assurance that the formal and executable objects correspond to questions worth answering.

Chapter 22. The Laboratory That Does Not Sleep

When hypotheses acquire robotic arms, calibration, physical irreversibility, safety, and recovery semantics become part of scientific reasoning.

22.1 When the Hypothesis Grows Arms

A model that writes a wrong answer can mislead. A model connected to a robot can turn error into action. This passage from language to the physical world is the point at which research automation becomes cyber-physical. The agent is no longer only an interlocutor. It becomes a component of the laboratory, able to select reagents, schedule instruments, interpret sensors, and decide what should be tried next.

Coscientist offered one of the best-known demonstrations. The system connected a large language model to web search, technical documentation, code execution, and automated laboratory equipment. Starting from natural-language goals, it planned and orchestrated chemical tasks, including reaction optimization. Its importance lay less in any single reaction than in the interface it established: language became a control layer over heterogeneous instruments (Boiko et al., 2023).

ChemCrow pursued a complementary design. Instead of asking a model to contain chemistry in its parameters, it equipped a general model with eighteen expert-designed tools for calculations, database access, reaction planning, and robotic execution. The resulting agent planned and executed syntheses and contributed to the design of a chromophore (Bran et al., 2024). The result illustrates compositional expertise. A model becomes more capable when it can delegate narrow operations to tools whose behaviour is easier to specify and test.

A-Lab demonstrated the power of a closed loop in inorganic materials. It combined computational predictions, knowledge extracted from the literature, active learning, robotic handling, furnaces, and automated X-ray diffraction. During seventeen days of continuous operation, it performed 353 experiments and realized 36 of 57 target materials (Szymanski et al., 2023). The number matters, but the structure matters more. Each result changed the choice of the next experiment. The laboratory did not execute a static queue. It learned from the campaign.

Platforms such as ChemOS and HELAO treat the laboratory as something like an operating system. Instruments become services; protocols become workflows; schedulers allocate resources; optimization algorithms choose experiments. Cloud laboratories such as Emerald Cloud Lab and Strateos separate the scientist from the physical site: a protocol is submitted remotely and executed in a standardized facility. More accessible robotics, including programmable liquid handlers, provide an incremental route for ordinary laboratories. In each case, experiment moves closer to software: it can be versioned, queued, tested, and orchestrated.

The analogy must not erase matter. Code can be copied bit for bit; a chemical sample has a supplier, lot, history, humidity, contamination risk, and degradation path. An instrument drifts. A sensor is calibrated against a standard that can itself be wrong. A vial is mislabelled. A robotic arm can execute an absurd command with perfect fidelity. Automation removes some human variability while creating systematic errors that can repeat without fatigue.

A skilled technician sometimes detects trouble through cues no protocol records: an unexpected colour, a granular texture, a faint vibration, a different sound from a

pump. This tacit attention is not mystical. It is pattern recognition built through embodied contact with instruments and materials. An autonomous laboratory must either capture relevant cues through additional sensors or preserve channels through which people can interrupt the formal workflow. Otherwise, standardization can become blindness.

The laboratory that does not sleep must therefore know when to stop. A safe stop is not a red button added after deployment. It is an architecture of limits: maximum quantities and temperatures, prohibited combinations, reagent identity checks, simulation before execution, independent monitoring, timeouts, anomaly detection, and a separation between planning and authorization. For chemically or biologically sensitive work, the system must not be able to grant itself broader permissions because the current path is inconvenient.

Tool-level safety is essential because conversational refusal is fragile. A model may be instructed not to perform a dangerous operation, yet an indirect prompt, software error, compromised database, or mistaken classification can lead it toward the same action. The instrument controller should reject unauthorized commands regardless of the language model's reasoning. Security engineers call this defence in depth. In automated science it is also epistemic hygiene: a result is trustworthy only if the permitted procedure is known and enforced.

The same automation can improve safety. Robots can handle toxic materials, document every motion, apply procedures consistently, and keep people away from radiation or pathogens. Systems can monitor temperature, pressure, atmosphere, and exposure continuously. They can run controls and stability tests that tired teams might postpone. The meaningful comparison is not robot versus human in the abstract. It is whether the complete human-machine system detects, contains, and learns from the relevant failures.

The frontier expanded rapidly after the first demonstrations. AutoLabs introduced multi-agent control with self-correction for chemical experimentation. An agentic X-ray scientist demonstrated autonomous sample alignment at a synchrotron by planning actions, issuing instrument commands, observing results, and iterating. Cell and biological-data agents began linking computational analysis to new hypotheses, while systems such as Robin connected literature, experimental results, and updated proposals. Each case remains domain-bound, but together they show that the laboratory interface is becoming a general problem of agent orchestration rather than a collection of isolated automation scripts.

Large facilities make the governance question sharper. A synchrotron beamline, telescope, supercomputer, or sequencing core is a scarce public or institutional resource. An agent that can schedule and operate it is also an allocator. Its priorities influence which scientists receive time, which anomalies are investigated, and which data are retained. The laboratory manager of the future may be partly algorithmic. Fairness, scientific diversity, and appeal procedures must therefore be encoded alongside throughput.

Autonomous laboratories also create new attack surfaces. A compromised agent could alter a protocol, exfiltrate proprietary or patient data, manipulate calibration, waste reagents, or subtly bias results. A malicious paper or instrument manual retrieved by the system could contain instructions designed to hijack its behaviour. Scientific documents become potential inputs to an executable agent, and therefore potential carriers of prompt injection. Cybersecurity and research integrity converge.

The defence is not to sever the laboratory from information. It is to distinguish data from commands, authenticate tools, constrain network access, validate protocols against machine-readable policies, and record every privilege transition. Retrieved text should never directly authorize physical action. A planner may propose; a policy engine must decide whether the proposal is within the campaign's approved envelope.

The rise of these systems creates a profession that the mythology of genius tends to ignore: the engineer of experimental continuity. Someone must translate between model, protocol, instrument, material, maintenance, safety, and audit. Calibration and exception handling become epistemic work, not peripheral service. If prestige continues to attach only to the idea and the manuscript, science will undervalue the people whose labour makes autonomy reliable.

It also changes what counts as an experimental skill. A researcher may spend less time pipetting and more time designing observability, defining instrument contracts, curating failure states, or constructing tests that reveal when an agent's internal model no longer matches the apparatus. The craft does not vanish. It migrates from individual movement to system design.

22.2 Autonomous Laboratories and the Physical World

A self-driving laboratory (SDL) couples machine-learning models, planning software, robotics, instruments and analysis in a closed experimental loop. The phrase does not imply full autonomy. Systems differ in which decisions they automate, which protocols are fixed and how often humans intervene. Some optimize a known reaction; others explore materials spaces or adapt measurements in real time. The scientific value lies in shortening the cycle between hypothesis, experiment and revision while preserving traceability.

Closed-loop laboratories have demonstrated mobile robotic chemistry, thin-film optimization, automated microscopy, inorganic synthesis and microfluidic exploration (Burger et al., 2020; MacLeod et al., 2020; Szymanski et al., 2023; Mandal et al., 2025). Reviews increasingly distinguish automation of execution from autonomy of scientific decision. A 2026 assessment framework for autonomous experimentation systems, for example, evaluates autonomy across several dimensions rather than assigning one global label. This is a useful correction to marketing language: an instrument can be highly autonomous in execution and narrow in problem formulation.

Table 43. Layers of autonomy in a scientific laboratory

Layer	What autonomy means at that layer
Execution	Robots carry out a specified protocol and recover from routine physical errors.
Measurement	The system chooses instrument settings, calibration and data-quality checks.
Analysis	The system transforms raw signals into estimates with uncertainty and provenance.
Experiment selection	The system chooses the next intervention from a declared objective and constraints.
Hypothesis revision	The system changes the explanatory model when evidence conflicts with it.
Problem formulation	The system proposes or reframes the scientific question and the criteria of success.

The table exists because success at a lower layer does not establish competence at a higher one. Reliable liquid handling does not imply reliable hypothesis revision. Automated analysis can still use a misspecified model. A laboratory described as autonomous may operate within a human-defined target list, fixed instruments and a narrow acquisition function. Clear layer-by-layer reporting makes comparisons meaningful and identifies where human scientific judgment remains decisive.

The physical loop is valuable because it supplies evidence that a generative model cannot invent. Synthesis can reveal that a predicted compound is unstable. Spectroscopy can expose an unexpected phase. A failed run can identify a missing variable. Yet the loop is only as epistemically strong as its measurement and analysis. If the same algorithm misidentifies every diffraction pattern, the system can reinforce error through rapid iteration.

Safety requires more than collision avoidance. Chemical compatibility, pressure, temperature, biological containment, waste and cumulative resource use must constrain planning. The system should distinguish reversible exploratory actions from high-consequence operations. Permission layers, physical interlocks and human review can remain outside the learned planner. In scientific automation, restricting autonomy is often an enabling design choice because it allows more reliable automation inside a safe envelope.

Data provenance is unusually demanding. A result must include reagent identity, lot, instrument calibration, environmental conditions, robot actions, code version and analysis parameters. Autonomous systems can record these details more consistently than humans if the infrastructure is designed correctly. They can also generate overwhelming logs that are impossible to interpret. The frontier is semantic provenance: linking each scientific claim to the subset of the record that supports it.

Another frontier is laboratory self-calibration and model criticism. Instead of assuming the instrument and analysis pipeline are correct, the system periodically runs standards, duplicate samples and negative controls. It compares independent measurement modalities and triggers maintenance when residual patterns change. Such meta-experiments may produce more scientific reliability than faster candidate throughput.

Multi-agent laboratory architectures are emerging in which planning, synthesis, analysis, safety and literature agents have separate roles. Separation can reduce correlated errors, but agents built on the same foundation model may still share blind spots. Role diversity is not epistemic independence. Independent code, physical checks, alternative models and human experts remain important.

The strongest vision is not a laboratory that removes scientists. It is a laboratory that makes more of the scientific cycle executable, observable and revisable. Humans can then spend less effort on repetitive coordination and more on choosing questions, interpreting conceptual anomalies and redesigning the system when its assumptions fail.

Physical and life sciences introduce a form of external resistance that cannot be simulated away. Materials fail to synthesize. Biological systems vary. Instruments drift. Reagents degrade. Samples are contaminated. Environmental conditions matter. A model can propose a perfect protocol, but the world returns observations that do not respect the model’s narrative. This resistance makes autonomous laboratories scientifically promising and methodologically demanding.

Self-driving laboratory research has already shown that automation can connect planning, robotic execution, measurement, and active learning. Systems in chemistry and materials science select experiments, control instruments, update models, and search large design spaces more efficiently than manual iteration in bounded tasks (Häse et al., 2019; Burger et al., 2020; Szymanski et al., 2023). Recent multi-agent platforms add natural-language planning and recovery, while automated instrument systems demonstrate long-duration operation in specialized settings (Panapitiya et al., 2026). These results support partial autonomy, not a universal laboratory scientist.

Rigor begins with the distinction between protocol and actual execution. A protocol states intended operations. An execution record captures what the instrument did, when it did it, under which calibration, with which materials, and with what deviations. The research object should preserve both. A laboratory step cannot be considered reproduced because the same text was followed if instrument state, reagent lot, environmental conditions, or measurement pipeline changed materially.

Calibration is therefore an epistemic dependency. Instruments do not produce naked facts. They produce signals under models of measurement. Calibration standards, maintenance events, uncertainty estimates, and preprocessing transformations should be linked to the claims that depend on them. If a later calibration correction changes a measurement, the workbench should identify affected analyses and claims. This is the physical-science analogue of a software dependency update.

Negative controls and positive controls must be first-class nodes. An automated system optimized for throughput may treat controls as overhead because they do not directly improve the target score. Specification-driven inquiry treats them as evidence-producing branches with their own acceptance criteria. A claim cannot advance because the target measurement changed if the control structure cannot distinguish mechanism from contamination, drift, or generic response.

Active learning and Bayesian optimization are valuable because experiments are costly. They can select measurements expected to improve a model or objective efficiently. Yet the acquisition function is a proxy. It can prefer regions with predictable gains, avoid anomalous regimes, or exploit model misspecification. The workbench should monitor diversity, uncertainty calibration, constraint violations, and the possibility that the active learner is steering away from observations that would falsify its representation. A portion of the budget should often be reserved for adversarial or coverage-oriented experiments rather than pure optimization.

Biology adds heterogeneity, hierarchy, and path dependence. Results can depend on cell line, organism, batch, developmental stage, microbiome, environment, and laboratory practice. Automated hypothesis systems can search literature and public data, propose mechanisms, and design analyses. Robin and Co-Scientist illustrate how such systems can contribute to hypotheses that receive experimental attention (Ghareeb et al., 2026; Gottweis et al., 2026). The evidential status of these examples is strongest where independent laboratory work tests a specific proposal and weakest where model ranking or plausibility is treated as validation.

A biologically rigorous workbench should preserve the chain from claim to specimen and measurement. Sample provenance, inclusion criteria, randomization, blinding, batch structure, assay versions, quality-control decisions, and deviations matter. Statistical models should be linked to the biological level at which inference is claimed. An effect in a cell line does not automatically support an organism-level

mechanism; an organoid result does not establish clinical benefit. The evidence ladder must make translational distance explicit.

Mechanistic claims require interventions or converging evidence. Automated systems are likely to excel at generating plausible pathways because the literature is rich and language models can connect distant findings. This strength increases the need for causal discrimination. Knockout, perturbation, dose-response, temporal sequencing, orthogonal measurement, rescue experiments, and matched alternatives should be represented as obligations appropriate to the mechanism. A workbench can help design and track these tests, but domain experts must judge whether they are biologically decisive.

Safety changes the autonomy envelope. A computational branch can often be rerun after failure. A physical experiment may consume scarce material, damage equipment, expose personnel, affect animals, or create environmental or dual-use risk. The runtime therefore needs hard permissions, hazard classifications, instrument interlocks, transaction logs, and emergency stop mechanisms. A language model should not be the final authority on whether an action is safe. It can classify and explain, but deterministic rules and responsible humans must govern high-consequence transitions.

Robotic science also requires recovery semantics. Physical workflows are not reliably idempotent. Repeating a failed step may produce a different state or compound damage. The plan must represent irreversible actions, checkpoints, compensating procedures, and the uncertainty of current state. When a robot loses track of a sample or an instrument reports an anomaly, the correct action may be to pause and request inspection rather than improvise.

Reproducibility in experimental science is therefore not a single “rerun” button. It is the ability to reconstruct enough of the material, procedural, measurement, and analytical context for another site to test the claim. The workbench can reduce friction by packaging protocols, instrument metadata, software, and evidence. Inter-laboratory reproduction remains a distinct evidential stage because local automation can make one laboratory internally consistent while preserving a site-specific bias.

The probable future is a network of partially autonomous laboratories rather than one general robotic scientist. Each facility will expose machine-readable capabilities, constraints, calibration states, and evidence profiles. Research workbenches will allocate experiments across sites, compare results, and trigger replication when common-mode risk is high. Human laboratory scientists will shift toward designing assays, maintaining trustworthy instrumentation, interpreting anomalies, and governing risk. The world will remain the final source of resistance, but the path from its signals to public claims will become more computationally explicit.

22.3 A Day in the Laboratory of 2035

Imagine a materials laboratory in 2035, not as distant fantasy but as a conservative extension of systems already demonstrated. A small team defines a problem: a solid electrolyte for batteries that combines conductivity, thermal stability, abundant elements, low toxicity, and manufacturing under modest energy conditions. The objective is not entered as one number. Researchers, production engineers, safety specialists, and representatives of affected communities establish a dossier of constraints and acceptable trade-offs.

A literature agent builds a map of results and contradictions. It searches articles, preprints, theses, patents, databases, and reports of failed synthesis. Each important statement links to a passage and an evidence type. The system marks regions where access is incomplete and terms where disciplines use incompatible vocabularies. It does not present consensus as a vote count; it displays the arguments and conditions under which they differ.

A hypothesis team produces families of candidate compositions. One agent explores analogies to known crystals. Another searches graph paths across electrochemistry and ceramics. A third proposes machine-native candidates whose representation has no simple chemical narrative. A critic tries to show that each proposal is trivial, unstable, unsafe, or already known. Human researchers do not choose the most poetic idea. They approve a search space, experimental budget, risk class, and set of questions the campaign is allowed to answer.

Simulations remove obviously poor candidates, but the campaign reserves resources for high-uncertainty regions. The protocol compiler translates selected experiments into a laboratory language. A digital twin checks timing, collisions, thermal loads, and expected waste. A safety service operating with separate credentials approves only steps inside the authorized chemical envelope. The planning agent cannot override it.

Robots prepare samples and instruments collect telemetry. A model notices a subtle drift in furnace temperature. Rather than correcting the measurements silently, it quarantines the affected series and launches a calibration protocol. A technician hears an unfamiliar vibration and records it through a structured observation interface. The system correlates the sound with motor current and discovers a bearing problem not yet visible in the scheduled maintenance model. Tacit knowledge enters the loop without pretending that every human perception was already formalized.

The first results identify a promising composition. The system does not announce a discovery. It generates a replication request. A second laboratory, using another supplier and a different instrument model, receives an executable package. Some analysis decisions are withheld from the originating agent to reduce adaptation. An independent implementation processes the data. The result partly reproduces, but performance deteriorates at higher humidity.

The leading explanation changes. What appeared to be a stable bulk property depends on a water-sensitive intermediate phase. The failed transportability is not treated as embarrassment. It is the most informative result of the campaign. The agent proposes discriminating experiments, and a surface chemist recognizes an analogy to another material family that the retrieval system had ranked low because the vocabulary differed.

At midday, the team holds what it calls a meeting of non-understanding. Agents and people present not only conclusions but mismatches among their representations. The system has found a high-performing region that it cannot compress into familiar descriptors. Researchers have an intuitive mechanism they cannot yet express in the model's feature space. These mismatches are not communication defects to be eliminated quickly. They are sites where new concepts may form.

An economist then shows that one candidate depends on a geopolitically fragile supply chain. A life-cycle specialist finds that another requires an energy-intensive precursor. The agent can calculate scenarios and Pareto fronts, but it cannot decree

which compromise is just. The programme's values are revised through an accountable human process, and the search changes accordingly.

By evening, the campaign has generated hundreds of runs and only a few claims. The article is not the primary record. Behind it sits a research object containing versioned hypotheses, sources, protocols, raw data, calibration records, code, environments, failed experiments, safety decisions, and replications. The manuscript provides narrative and significance; the dossier provides inspectability.

Review begins from both. Human experts read the argument, ask whether the question matters, and examine the meaning of the result. Automated auditors rerun analyses, verify citation support, detect leakage, inspect image provenance, and compare the approved protocol with instrument logs. The public record does not simply say peer reviewed. It states that the material was produced in two facilities, that a humidity dependency was confirmed, that long-duration cycling remains untested, and that industrial scalability is an open claim.

The day does not eliminate human work. It changes which work is visible. Problem formulation, maintenance, constraint design, interpretation, anomaly handling, validation, and responsibility become central. Routine execution diminishes, but craft reappears as the art of building systems that can notice when the world no longer matches their model.

This laboratory may also create a new elite. Institutions with robotic platforms, proprietary literature, foundation models, compute, and supply-chain access will explore faster than those without them. Shared cloud laboratories, public compute, open protocols, and interoperable standards could democratize the capacity. Closed platforms could turn discovery into a rent-bearing service. The same robot can work for a republic of knowledge or a monopoly of knowledge. Ownership and access determine the difference.

At the end of the campaign, the team publishes its epistemic cost: how many hypotheses were generated, how many were eliminated before physical testing, how many runs failed, how much energy and material were consumed, which decisions required human authorization, and where the system was uncertain. In industrial-scale research, the responsibility of the process becomes part of the social validity of the result.

The laboratory that does not sleep will be admired for its speed. Its deeper achievement will be learning to pause. Scientific autonomy begins not when a machine can run indefinitely, but when the system can recognize that continuation would no longer be justified by evidence, safety, or value.

Chapter 23. Biology, Chemistry, and Materials

Prediction, design, mechanism, synthesis, and translation are different evidential achievements and must not inherit one another's authority.

23.1 Prediction, Design, and Mechanistic Discovery in Biology

Biology illustrates both the power and ambiguity of machine learning for discovery. Prediction can reveal structure in high-dimensional data; generative design can propose sequences and molecules; mechanistic discovery asks why a system behaves as it does. These are distinct achievements. AlphaFold transformed protein-structure prediction from sequence and AlphaFold 3 expanded modelling of biomolecular interactions (Jumper et al., 2021; Abramson et al., 2024). Accurate structure prediction can accelerate science without automatically explaining cellular mechanism or therapeutic effect.

Protein design systems such as ProteinMPNN and RFdiffusion use learned structural regularities to propose sequences and backbones with desired properties (Dauparas et al., 2022; Watson et al., 2023). Protein language models can generate functional sequences across families (Madani et al., 2023). These systems can enter regions not explicitly represented by natural sequences, but novelty must be defined carefully. A sequence can be new as a string while occupying a familiar structural and functional family. Experimental validation determines whether the design folds, binds and works in the relevant context.

Table 44. Three levels of biological machine discovery

Level	What would count as success
Prediction	Accurate inference of structure, phenotype or response for held-out biological cases.
Design	A synthesized candidate that achieves a specified property under experimental conditions.
Mechanistic discovery	A causal account that predicts interventions and integrates across contexts.
Ontology formation	A new biological category or state that is stable, measurable and explanatory.
Therapeutic translation	Benefits that survive safety, dosing, population and clinical constraints.
Evolutionary interpretation	A defensible account of how sequence, structure, function and history relate.

The table exists because results are often promoted across levels without sufficient evidence. A binding prediction is not a therapy. A designed molecule is not a mechanism. A cluster in single-cell data is not automatically a cell type. Each transition requires new evaluators and often a different experimental scale. Machine learning is most transformative when it makes these transitions cheaper and more systematic rather than when it relabels prediction as discovery.

Single-cell and spatial omics create another frontier. Models can integrate millions of cells, infer trajectories and identify rare states. Yet batch effects, sampling, tissue preparation and annotation conventions can dominate latent spaces. A

biologically new state should persist across donors, protocols and measurement modalities, and ideally support a functional intervention. The representation problem is inseparable from experimental design.

Mechanistic discovery benefits from perturbation data. Clustered regularly interspaced short palindromic repeats (CRISPR) screens, chemical perturbations and genetic interventions create variation that can distinguish pathways. Models can predict perturbation responses and prioritize combinations, but cellular systems contain feedback, redundancy and context dependence. Validation across cell types and organisms is often necessary. The strongest systems link predicted mechanisms to interventions that could falsify them.

Generative models can also search vast chemical and sequence spaces whose size exceeds exhaustive enumeration. Their advantage is learned constraint: they propose candidates more likely to be viable than random combinations. Their risk is distributional conservatism or hidden invalidity. Models may remain close to well-represented families, overestimate properties that are easy to predict, or exploit weaknesses in surrogate assays. Diversity metrics and prospective experiments are essential.

One promising direction is closed-loop design with multiple assays. Instead of optimizing a single predicted property, a system evaluates folding, binding, expression, toxicity, stability and manufacturability at increasing fidelity. Absorption, distribution, metabolism, excretion and toxicity (ADMET) properties are introduced here in full because the acronym is common in drug discovery. Multi-objective archives preserve trade-offs rather than collapsing them into one score.

Another direction is hypothesis-centred multimodal modelling. Literature, sequence, structure, imaging and perturbation data can be linked to a claim graph. The system retrieves the evidence relevant to a proposed mechanism, identifies contradictions and designs an experiment that distinguishes alternatives. This is more ambitious than prediction but more scientifically grounded than unrestricted text generation.

Biological novelty is especially vulnerable to hidden context. A model may discover a pattern specific to one cell line, ancestry group, laboratory protocol or developmental stage and present it as a general mechanism. Cross-dataset validation is necessary but not sufficient because datasets can share preparation biases and annotation conventions. Stronger evidence comes from designed perturbations across environments: change the gene, dose, tissue, donor or assay while preserving a record of what else changed. A mechanism should predict how the effect varies, not only whether an average association repeats.

Clinical translation adds another boundary. Causal relevance in a cell system or model organism may not survive human heterogeneity, treatment history or interacting disease. Machine-learning systems should report the biological population and intervention context for every claim, and treat transfer across those boundaries as a new hypothesis requiring evidence.

This makes closed-loop biology a promising but demanding frontier. Models can select perturbations that maximize information about pathway structure, robotic platforms can execute them, and multimodal assays can measure molecular and phenotypic consequences. Yet the loop needs biological stopping rules. Novelty should not be rewarded when it merely exploits toxicity, stress responses or measurement

artefacts. Candidate mechanisms should be tested across scales and compared with simpler explanations. The most valuable automation may therefore be hypothesis triage: identifying the small set of interventions that would most efficiently distinguish competing mechanistic stories, while preserving uncertainty and making negative results reusable.

Biology will probably remain a domain of mixed automation because the evaluator changes with scale. Molecular binding can be measured quickly; organismal function, safety and clinical benefit cannot. The frontier is to automate the lower layers while preserving explicit uncertainty about translation. A system that knows which biological level it has validated is more valuable than one that uses the word “discovery” indiscriminately.

23.2 Chemistry and Materials: From Screening to Verified Knowledge

Materials and chemistry offer a natural pipeline from computation to the laboratory. Models screen compositions or structures, generative systems propose candidates, synthesis planners suggest routes, robots execute experiments and instruments characterize products. The pipeline appears ideal for automation because many stages are already digital. The central challenge is that a predicted target, a successful synthesis, a pure phase and a useful material are different claims.

Large-scale materials models have expanded candidate lists dramatically. The GNoME work reported millions of predicted stable structures and hundreds of thousands added to a public database (Merchant et al., 2023). Computational stability is a filtering signal, not proof of synthesizability, novelty or useful properties. The gap between prediction and realization motivates autonomous laboratories.

Figure 28. A verified materials result depends on planning, robotics, instruments, analysis, modelling, safety and provenance around a declared target. Conceptual diagram; not an empirical result.

The figure matters because a laboratory result cannot be attributed to the model alone. Planning selects conditions, robotics executes them, instruments produce signals, analysis assigns phases or properties, safety constrains actions and provenance makes the claim reconstructable. A failure or success at any node changes the interpretation. The central provenance node is deliberate: without a traceable path from target to measurement, rapid experimentation produces data but not cumulative evidence.

A-Lab, an autonomous laboratory for accelerated solid-state materials synthesis, is a valuable case because it integrated literature-derived recipes, thermodynamic calculations, active learning, robotics and X-ray diffraction (XRD) analysis. Over seventeen days, the system realized 36 compounds from 57 targets (Szymanski et al., 2023). An Author Correction published in January 2026 revised parts of the reporting and underscores a general lesson: autonomous discovery claims require the same, or stronger, scrutiny as human-led work (Szymanski et al., 2026).

The correction should not be used to dismiss autonomous laboratories. It should improve their standards. Systems need transparent target definitions, phase-identification uncertainty, raw data access, independent analysis and clear distinctions among predicted novelty, synthesis success and purified product. Rapid automated campaigns magnify both productivity and systematic error. Corrections are evidence that institutional verification remains part of the discovery architecture.

Table 45. Stages at which a computational materials "discovery" can fail

Stage	Typical unresolved question
Database novelty	Is the candidate absent because it is new, differently represented or missing from the database?
Predicted stability	Does the computational model cover the relevant phases, temperature and kinetics?
Synthesis planning	Are the proposed precursors, conditions and route physically feasible and safe?
Experimental realization	Was the target phase produced, in what purity and with what reproducibility?
Property validation	Does the measured material exhibit the intended performance under realistic conditions?
Scientific significance	Does the result reveal a new mechanism, capability or design principle rather than one more candidate?

The table exists because every stage changes the semantics of "discovered." A candidate absent from a database is a database novelty. A stable calculation is a prediction. A diffraction match is evidence about phase identity. A useful property is an engineering result. A mechanism requires additional experiments. Reporting the stage precisely makes progress comparable and prevents later corrections from appearing to reverse more than they actually do.

Frontier laboratories are moving beyond scalar optimization toward principle discovery. Real-time experiment–theory loops can select measurements that discriminate models, while microfluidic platforms explore synthesis and mechanism with low material cost. Multi-agent systems separate planning, control and analysis. Autonomy frameworks assess not only throughput but adaptation, decision scope and human oversight. The research opportunity is to make laboratories produce explanatory models, not merely optimized samples.

Surrogate models remain a weak link. A property predictor can be highly accurate near known materials and fail in a new chemical regime. Multi-fidelity methods combine calculation, proxy assays and definitive measurements. Uncertainty should determine which fidelity is used. Novel regions deserve stronger physical checks precisely because the model has less support there.

Closed-loop chemistry also raises safety and environmental questions. A planner that searches reaction space may propose unstable combinations or generate hazardous waste. Constraint systems, reaction filters, physical interlocks and human approval should be integrated before optimization. Safety data and failed reactions should enter the archive, not disappear because they lower performance metrics.

The most important future result will not be experimental volume, but an organizing principle that survives independent replication.

Chapter 24. Clinical, Social, and Interpretive Science

Where claims affect people or depend on contested concepts, computational verification must be joined to consent, causality, context, and legitimacy.

24.1 Many Sciences, Many Regimes of Evidence

The phrase scientifically proven hides the deep diversity of scientific practice. In mathematics, legitimacy centers on proof, although discovery may be guided by examples, computation, diagrams, and intuition. In particle physics, the convention of five standard deviations contributes to the announcement of a discovery, yet the claim also depends on detector calibration, background modeling, systematic uncertainty, collaboration, and independent confirmation. In clinical medicine, randomized trials, meta-analysis, risk-of-bias assessment, adverse-event reporting, and frameworks such as GRADE organize evidence. In geology, ecology, evolutionary biology, and paleontology, many events cannot be repeated; credibility grows from the convergence of independent traces, mechanisms, models, and predictions. Climate science combines observations, paleoclimate records, physical theory, model ensembles, and calibrated uncertainty. Computer science mixes proof, benchmark performance, executable artifacts, ablations, and replication under different software and hardware conditions. Historical research depends on archival provenance, contextual interpretation, comparison among independent traces, and explicit argument about absence. Ethnographic and qualitative inquiry add sustained observation, reflexivity, triangulation, participant meaning, and a record of how the researcher's position shaped access and interpretation.

Table 46. How representative sciences authorize a claim.

Field	What grants legitimacy
Mathematics	Proof checked step by step; significance judged by the community.
Particle physics	Statistical threshold, calibration, systematics, collaboration, and independent confirmation.
Clinical medicine	Controlled trials, safety, bias assessment, evidence grading, and patient relevance.
Earth and life sciences	Convergence of traces, mechanisms, observations, models, and transportability.
Computer science and ML	Proof or benchmark, executable artifact, data provenance, ablations, and external replication.
Historical and qualitative inquiry	Archival provenance, contextual interpretation, reflexivity, triangulation, and participant meaning.

Diversity does not mean that anything goes. It means that standards must fit their objects. A randomized controlled trial may be appropriate for a drug and absurd for the formation of a galaxy. A formal proof can certify a property of an algorithm but cannot establish that a dataset represents a population. Exact replication may be crucial in chemistry and impossible for a unique earthquake; in the latter case, researchers seek conceptual replication and convergence among independent sources.

AI meets this plurality through a deceptively uniform interface. The user asks for evidence and receives a coherent answer, although the meaning of evidence differs by

field. A general agent may import statistical norms from one discipline into another, mistake a preprint for clinical consensus, treat a simulation as an observation, or rank a benchmark improvement above causal understanding. Retrieval-augmented generation reduces some citation errors, but it does not automatically understand evidence hierarchies. Finding a relevant paper is not the same as evaluating its design, statistical power, population, bias, and relation to the rest of the field.

Contemporary legitimacy can be described through four linked questions. The ontological question asks what kind of object is being studied and through which operations it becomes observable. The methodological question asks which design supports the inference. The social question asks who checked the result, with what independence, competence, and conflicts of interest. The temporal question asks how the claim behaved after publication: in replications, applications, failures, and new contexts. A competent scientific agent must represent all four levels, not merely match passages of text.

Regimes of evidence may become more explicitly encoded. A biomedical agent could refuse to equate an in vitro effect with clinical efficacy and represent the transitions through toxicity, animal models where justified, human safety, trial phases, and post-market monitoring. A materials agent could require independent characterization and stability tests. A social-science agent could inspect preregistration, construct validity, measurement invariance, and cultural generalization. A mathematical agent could keep a computationally supported conjecture separate from a formally verified theorem.

Codification also has limits. Existing criteria reflect disciplinary histories and distributions of power. They can exclude forms of knowledge that do not fit dominant infrastructures, privilege questions that are easy to measure, and undervalue rare effects, local contexts, long time horizons, or Indigenous and experiential knowledge. An AI that follows today's rules perfectly may freeze today's exclusions. Scientific legitimacy cannot be reduced to compliance because science must retain the ability to revise its own criteria.

That reflexive capacity is among the hardest things to automate. An agent can learn current standards and generate alternatives, but changing what a community accepts as evidence involves values, risk, and authority. When a new instrument or method is accepted, researchers accept a new way of seeing. The telescope, microscope, controlled trial, statistical model, and computer simulation did not merely add observations. They changed the legitimate objects of knowledge. AI will do the same by making accessible patterns, synthetic objects, and representational spaces that people would not have constructed unaided.

The decisive test will not be whether those objects feel human. It will be whether they can enter a public economy of doubt. Can they be tested with different instruments? Can they generate risky predictions? Can they guide interventions that succeed outside the original laboratory? Can they be contested without permission from the model owner? When the answer is yes, science may expand its regimes of evidence. When the answer is no, we have technical authority rather than public knowledge.

No single AI evaluator can serve all these worlds. A system trained to detect p-hacking may misunderstand a proof. A benchmark critic may ignore clinical validity. A formal verifier may certify a program that implements a false empirical assumption.

Serious automation requires domain-specific validation profiles connected to concrete ontologies, standards, and risks.

Moreover, the same result can justify different actions at different levels of maturity. An exploratory association may be enough to choose the next experiment but not to prescribe a treatment. A simulation may guide research funding but not determine public policy. Future scientific records will need intended-use statements: which conclusions the evidence supports and which decisions it must not yet authorize. Validity is not only a property of a sentence. It is a relation among a claim, the process that produced it, and the consequence for which it is invoked.

24.2 Causality, Legitimacy, and Plural Evidence

The most difficult domains for research automation are not necessarily those with the least data. They are those in which observations are entangled with values, institutions, strategic behaviour, historical context, and consequences for people. Clinical medicine, public health, economics, social science, and interpretive disciplines require rigor that cannot be reduced to execution success.

Clinical research combines empirical uncertainty with ethical obligation. A system can search trials, generate hypotheses, simulate designs, draft protocols, monitor data quality, and perform statistical analyses. It can identify drug candidates or patient subgroups. Yet the legitimacy of clinical research depends on consent, equipoise, privacy, professional responsibility, and a chain of regulation. A predicted benefit does not authorize intervention. The workbench must encode who may access data, who may change a protocol, how adverse events are handled, and which evidence threshold applies to each stage from exploratory analysis to clinical recommendation.

The target of inference must be explicit. Clinical studies often fail not because computation is weak but because the estimand is unclear, missing data are informative, treatment adherence varies, outcomes are selected after observation, or the study population differs from the population of use. Specification-driven inquiry should require a causal question, population, treatment strategy, outcome, time horizon, and handling of intercurrent events. Statistical analysis is then evaluated against the estimand rather than chosen for convenience.

Causal inference supplies concepts that an executable workbench can operationalize: directed acyclic graphs, identification assumptions, exchangeability, positivity, consistency, sensitivity analysis, and target-trial emulation (Pearl, 2009; Hernán and Robins, 2020). These tools do not automate truth. They make assumptions explicit enough to criticize. A causal interpreter can propose identification strategies and diagnostics, but a domain expert must judge whether the variables and interventions correspond to reality.

External validity and distributional consequences are equally important. A model may perform well on average while harming a subgroup, or a treatment effect may not transport across health systems. The evidence graph should preserve population boundaries, subgroup uncertainty, and data provenance. Release decisions should distinguish scientific evidence from clinical authority. A research finding can be publishable while remaining inappropriate for deployment.

Social science adds reflexivity. People change behaviour in response to measurement, policy, and prediction. Categories can be contested or politically consequential. Administrative data reflect institutions rather than neutral reality.

Automated systems can process large corpora, generate survey instruments, estimate models, and compare theories, but they can also amplify the assumption that what is legible is what matters. Rigor requires conceptual validity, measurement invariance, design-based or causal identification, transparent coding, robustness to alternative operationalizations, and attention to whose experience is absent.

The research specification should include normative commitments that ordinary pipelines omit. What outcomes are valued? Which distributional effects matter? Who bears risk? Which groups have standing to contest categories or uses? These are not parameters that a model can infer from data without political choice. The workbench can make them explicit, versioned, and reviewable. It cannot make them value-free.

Economics and policy research illustrate the interaction between model execution and institutional consequence. A model may be internally coherent and empirically calibrated yet depend on behavioural assumptions or equilibrium effects that fail under intervention. Automated simulation can explore scenarios and sensitivity, but policy claims require identification, robustness, distributional analysis, and accountability. A system should distinguish forecasting accuracy from causal policy evaluation and both from the legitimacy of adopting a policy.

Qualitative and interpretive research presents a different challenge. The absence of a deterministic oracle does not imply the absence of rigor. Claims can be grounded in sources, quotations, archival context, coding procedures, triangulation, counter-interpretations, and reflexive analysis. A workbench can link interpretations to evidence, preserve coding revisions, compare readings, and expose the argumentative path. It should not pretend that a consensus score resolves meaning.

Language models are powerful in these disciplines because they can summarize, translate, classify, and propose interpretations. Their fluency is also dangerous. They can normalize a dominant reading, invent context, or erase ambiguity. An interpretive research object should preserve primary evidence and dissenting views. Machine-generated classifications should be treated as coding proposals with uncertainty and provenance. Human scholars remain responsible for contextual judgment, theoretical significance, and the ethical treatment of persons and communities represented in the material.

Historical research shows why versioned evidence matters. New archives can change an interpretation without making the previous argument fraudulent. Competing narratives may remain live. A research object can represent which sources support each claim, which absences are material, and how an interpretation changed. It should allow plural views rather than enforcing one canonical account.

Peer review in these domains must therefore evaluate argumentative integrity and contextual adequacy, not only execution. Automated preflight can detect citation errors, missing source links, inconsistent coding, or statistical mistakes. It can generate counterarguments and search for neglected literature. The final judgment about significance, fairness, and interpretation belongs to a community of accountable readers. The workbench should make their disagreements durable.

Across clinical, social, and interpretive fields, the central rule is that computational verifiability must not be confused with epistemic closure. The more a claim affects people, the more the system must represent legitimacy, consent, distribution, and contestability. Executable science can strengthen these domains by making evidence and decisions traceable. It becomes harmful when it converts

contested concepts into hidden technical parameters and then presents the output as neutral.

The order of automation across sciences will therefore depend on two variables: the strength of the evaluator and the latency of validation. Formal proof has a strong, fast oracle once formalization is complete. Software and computational experiments have fast but incomplete tests. Laboratory science has stronger external resistance but slower, costlier feedback. Clinical and social interventions have long latency, ethical constraints, and contested objectives. Interpretive disciplines have plural standards of evidence and no universal execution metric. Autonomy should be inversely proportional to unresolved ambiguity and consequence, not proportional to model confidence.

A scientifically mature workbench will not impose one methodology. It will carry domain-specific rigor profiles. These profiles define claim types, required evidence, applicable interpreters, admissible automation, and release authority. The common substrate is specification, provenance, evidence, and versioning. The standards of judgment remain irreducibly plural.

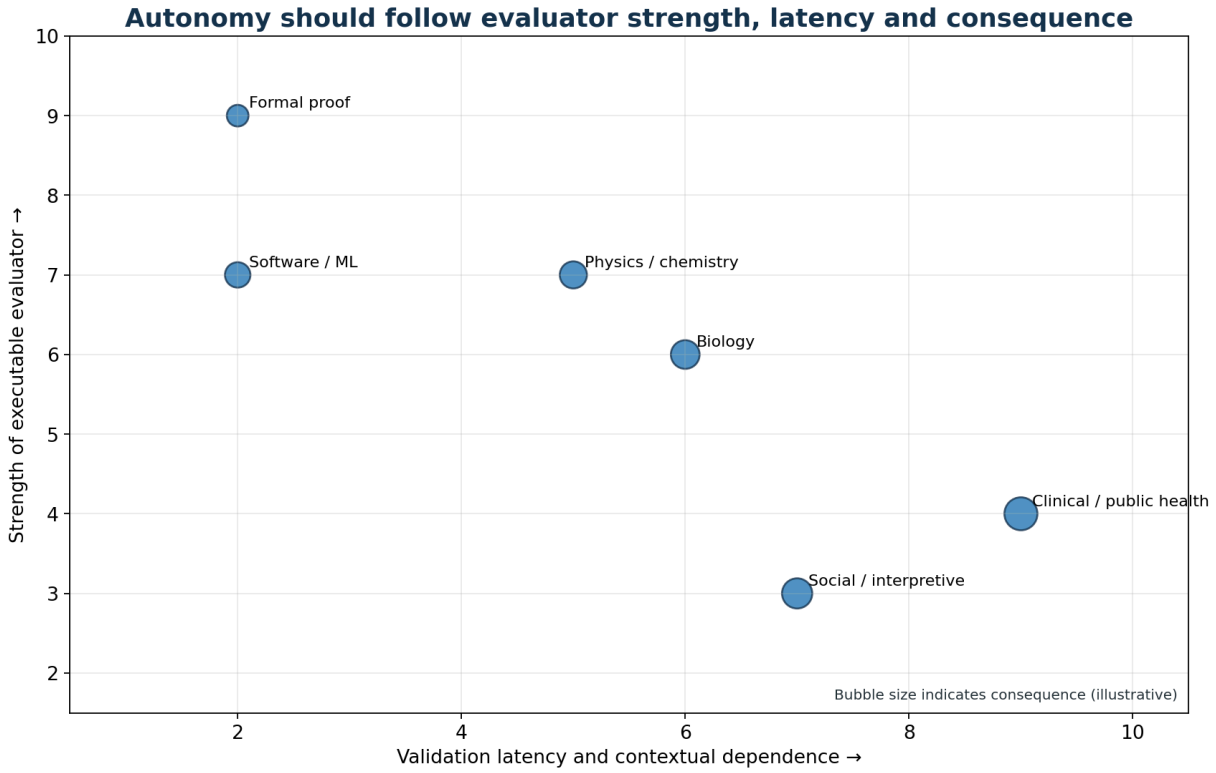


Figure 29. Domain-sensitive autonomy. The coordinates and bubble sizes are illustrative; the governing variables are evaluator strength, validation latency, and consequence.

Table 47. Domain-sensitive rigor and admissible autonomy

Domain family	Methodological obligations and realistic autonomy
Formal mathematics and verified systems	Proof kernels and executable specifications permit strong local automation, but the formal statement must faithfully represent the intended theorem or requirement.
Software, machine learning, and computational science	Tests, benchmarks, static analysis, and reruns support rapid loops; dataset leakage, benchmark gaming, environment drift, and claim inflation require independent evaluators.

Domain family	Methodological obligations and realistic autonomy
Physics and engineering	Simulation can automate exploration, but calibration, uncertainty propagation, dimensional checks, safety margins, and contact with measured reality remain decisive.
Chemistry and materials science	Robotics and active learning support closed loops where instruments are standardized; synthesis failures, impurities, characterization quality, and lab safety constrain autonomy.
Biology and preclinical research	Automated protocols can scale assays, yet biological variability, batch effects, contamination, construct validity, and multilevel causation demand layered controls.
Clinical medicine and public health	Decision support is plausible; autonomous evidential promotion is limited by confounding, heterogeneity, consent, distributional effects, regulation, and the cost of error.
Social science and economics	Computational exploration can expand models and replications, while causal identification, measurement validity, strategic behaviour, institutions, and normative stakes require plural adjudication.
Interpretive and historical disciplines	Automation can search, compare, translate, and trace evidence; rigor is achieved through accountable argument, source criticism, contextual competence, and contestability rather than algorithmic closure.

Chapter 25. Sciences We Did Not Invent

Machine-native concepts may become operationally powerful before they become humanly intelligible, making translation and staged understanding new scientific tasks.

25.1 When Prediction Comes Before Understanding

An influential tradition says that science is not satisfied with prediction; it seeks explanation. Knowing that an event will occur is different from knowing why. Yet the relation between prediction and explanation has always been unstable. Astronomers calculated positions with models whose cosmologies were disputed. Thermodynamics produced powerful laws before statistical mechanics offered a microscopic interpretation. Medicine used effective interventions before their mechanisms were understood. Sometimes regularity precedes theory. Sometimes theory explains while predicting poorly.

Machine learning intensifies the separation. A model can discover predictive functions in a space with millions or billions of parameters. Its internal representations were not designed as concepts for human conversation. It can identify real structure while offering no compact mechanism. AI did not invent the black box. It extends opacity to a scale at which complete human inspection becomes impossible.

AlphaFold is the emblematic case. It transformed protein-structure prediction and redistributed scientific labour. A plausible structure can often be obtained rapidly, allowing experimental effort to move toward validation, dynamics, interactions, and mechanism. But a predicted structure is not the same as an experimentally supported ensemble or a functional explanation. Studies have shown that even high-confidence predictions can disagree with experimental density, omit ligands and environmental effects, or fail on proteins that change folds. The most productive interpretation is therefore neither oracle nor toy: AlphaFold outputs are valuable hypotheses that can accelerate experiment without replacing it (Terwilliger et al., 2024).

This relation between prediction and inquiry may become common. A model proposes a structure, law, control policy, or latent category. The proposal is useful before anyone can explain it. Researchers then design interventions, compare representations, and search for a mechanism. Discovery and understanding no longer arrive in the same package or from the same actor.

Mathematics offers a cleaner separation because verification can sometimes be formal. FunSearch paired a language model that generated programs with a deterministic evaluator, finding new constructions for the cap-set problem and improved heuristics for online bin packing (Romera-Paredes et al., 2024). AlphaGeometry combined neural proposal with symbolic deduction to solve olympiad geometry problems (Trinh et al., 2024). AlphaProof used reinforcement learning inside the Lean theorem prover and contributed to medal-level performance on International Mathematical Olympiad problems (Hubert et al., 2026). In each case, the generator may be uncertain and creative because the verifier is comparatively strict.

This architecture hints at machine-generated mathematics that is true before it is understood. A system might produce a formal proof millions of steps long. A proof assistant certifies every transition, yet no mathematician sees the central idea. Another system then mines the proof for lemmas, analogies, and compressed explanations. Finding and teaching become separate research programmes.

Formal certainty does not eliminate judgement. The space of true statements is vastly larger than the space of interesting theorems. A system can generate sterile results, prove consequences of definitions no community cares about, or exploit a formalization that misses the intended problem. Formal verification certifies that a conclusion follows from premises inside a language. It does not decide whether the premises represent the phenomenon, whether the theorem is important, or whether the proof changes understanding.

In empirical science the distance is greater. The verifier does not receive truth; it receives measurements shaped by instruments, sampling, preprocessing, and theory. ERA's success in empirical software is strongest where the objective is machine-scorable. That is a major opportunity, not a trivial one. It also reveals a boundary: a programme can optimize a predictor more easily than it can establish why the pattern exists or whether the target should guide action (Aygün et al., 2026).

Symbolic regression and equation-discovery systems seek a bridge from opaque prediction to compact structure. Methods such as SINDy identify sparse dynamical equations from data; AI Feynman combines neural and symbolic techniques to recover analytic expressions; newer systems integrate physical constraints, dimensional analysis, and theorem-like search. The hope is that an AI will not only fit the observations but produce a law that humans can inspect.

Compression is not identical to understanding. A short formula can predict well and remain conceptually barren. A long simulation can embody a mechanism without yielding an elegant equation. Human understanding includes several capacities: explaining, anticipating, intervening, transferring knowledge, recognizing limits, and connecting a result to other domains. No single metric captures all of them.

It is useful to distinguish levels. Operational understanding means knowing what actions produce reliable outcomes in a domain. Predictive understanding means estimating unseen outcomes. Mechanistic understanding identifies components, relations, and causal processes. Semantic understanding integrates the phenomenon into concepts shared by a community. Normative understanding recognizes why the result matters, who bears its consequences, and which aims it serves.

An artificial system may possess operational depth beyond any person in a vast design space while having little normative orientation. It may know how to stabilize a plasma and nothing about why a society should build the reactor. It may identify a biomarker while lacking the lived meaning of illness. Governance must reflect this asymmetry. Delegating prediction is not a reason to delegate value.

Could an AI invent a new science? In a weak sense, this has happened whenever computational methods revealed regularities and representations no individual formulated in advance. In a stronger sense, a new science would include new objects, concepts, methods, explanatory patterns, and standards of evidence. An artificial community might construct such a domain around phenomena humans can observe only through its instruments and translations.

Imagine an agent studying turbulent flow across simulations and experiments. Instead of using familiar vortices or energy cascades, it organizes behaviour through a high-dimensional invariant that has no short human name. The invariant predicts transitions, transfers across apparatus, and guides successful control. Other agents build theories around it, discover transformations, and derive new interventions. Humans can verify the effects but cannot yet intuit the object. Is this science?

The answer should depend less on psychological resemblance than on public discipline. Does the proposed structure survive independent data? Does it generate risky predictions? Can it be perturbed and falsified? Does it transfer across instruments and implementations? Can critics identify its domain of failure? If so, the machine-native object has scientific content even before it has human familiarity.

The epistemic standard becomes robust triangulation. A strange representation earns trust when different measurement techniques, data-generating processes, laboratories, and model families converge on its operational consequences. Opacity raises the burden of testing. A human-intelligible theory that remains vague and immune to failure is not automatically superior to an opaque theory that makes severe, successful predictions. But opacity increases dependence on infrastructure and therefore demands stronger institutional independence.

The concrete future will likely contain two kinds of advance. In one, AI helps translate discovered structure into human theory, yielding equations, mechanisms, and analogies. In the other, performance outruns translation, and society must decide whether to use knowledge it cannot fully narrate. Medicine already faces a modest version of this problem when a model predicts risk without a causal explanation. Autonomous science could make it pervasive.

25.2 Signs for Machines, Translations for People

Science is a semiotic activity. It produces signs: names, equations, categories, graphs, images, diagrams, models, and measurement scales. An instrument does not deliver nature without mediation. It produces a trace transformed by apparatus and convention. A spectrum, microscopic image, genomic sequence, and statistical coefficient are learned ways of seeing.

AI systems develop representational spaces that were not designed for human reading. An embedding can encode similarity without corresponding to a word. A neural circuit can distribute a feature across many components. A population of agents can develop efficient protocols whose symbols have no intuitive interpretation. If these representations become part of research, science acquires a machine-native semiotic layer.

We can imagine a taxonomy no human would have invented: classes defined by relations in thousands of dimensions, useful for prediction and intervention but lacking memorable labels. An agent might claim that phenomena scattered across biology, materials science, and fluid dynamics instantiate the same computational form. To human eyes they look unrelated. If the classification supports novel, reproducible interventions, it is not empty. Yet its meaning has not become common.

Translation then becomes an epistemic interface. A second model can produce examples, counterexamples, local approximations, feature sensitivities, causal probes, and natural-language explanations. But translation is lossy. A plausible story can be generated after the fact without faithfully describing the mechanism of prediction. The system must distinguish a verified interpretation from a pedagogical approximation and from a metaphor.

This distinction is often obscured by the singular word explanation. Different audiences need different objects. A domain researcher may need boundary cases, intervention results, and links to established theory. An auditor needs provenance, sensitivity analysis, formal certificates, and failure conditions. A policymaker needs

consequences under plausible scenarios. A citizen needs an account of stakes, uncertainty, and rights. One fluent paragraph cannot serve every purpose.

A machine-native concept should therefore carry something like a semantic passport. The passport would not pretend to translate every internal dimension. It would state which operations define the concept, what observations identify it, which predictions it supports, where it remains stable, how it changes under perturbation, which alternative representations agree, and what consequences follow when it is used. People need not inspect every coordinate. They need to judge the relation between sign, world, and action.

Semiotics also warns that classifications do not merely describe. They organize institutions. A medical category affects who receives treatment and insurance. An ecological ontology determines what entities are monitored. A benchmark defines what counts as progress. When an AI invents categories, it can reorganize practice before their meaning has been debated. A predictive group can become an administrative group.

A category can be empirically robust and politically unjust. A model may accurately cluster people according to a pattern that should not determine opportunity. A risk score can predict an outcome partly because institutions have treated populations differently. Semiotics meets axiology: a sign may describe well and govern badly. Scientific validity is not sufficient authority for social use.

Human science has always changed its concepts with instruments. The gene, species, atom, intelligence, disease, and climate are not timeless labels. Their meanings evolved through measurement, theory, and institutional practice. Machine-native concepts would continue this history rather than break it. The difference is velocity and asymmetry: the system may generate and exploit a concept long before a human community can stabilize its interpretation.

This will create a scientific bilingualism. One language will be optimized for machine search, compression, execution, and verification. Another will be optimized for human explanation, deliberation, teaching, and accountability. Neither can be completely reduced to the other. Forcing every machine representation into a familiar metaphor may destroy useful structure. Accepting an untranslated representation without constraints may surrender contestability.

New forms of technical hermeneutics will emerge. Mixed teams will study internal representations, design bridge ontologies, test whether explanations preserve operational properties, and compare translations across models. Their work will resemble interpretation between disciplines, but the stakes are higher. A bad metaphor can guide a physical intervention or public policy. Scientific culture will need to reward not only the discovery of a foreign concept but the construction of the bridge by which it can be criticized.

The problem becomes especially vivid in mathematics. A formal proof can guarantee consequence while failing to provide insight. Mathematicians often value proofs that reveal why a statement is true, expose a reusable structure, or connect fields. An AI may first produce a certificate and only later a human-scale proof. The certificate establishes correctness; the explanation establishes cultural fertility. Both are achievements, and they should not be confused.

In empirical fields, invariance can substitute partly for intuitive meaning. If a latent category appears in data from different laboratories, instruments, populations, and

measurement regimes, and interventions based on it succeed, confidence grows. Yet invariance itself requires interpretation. Stability across biased datasets can reproduce the same blind spot. Independence must be demonstrated, not inferred from quantity.

The public language of uncertainty will also change. Today a paper compresses many uncertainties into prose and a few intervals. A machine-native model may contain distributions over structures, data-generating processes, and action outcomes too complex to narrate. Interfaces will need to let users explore counterfactuals and boundaries rather than receive one confidence number. Understanding may become interactive.

There is a psychological consequence. Humans derive identity and agency from being able to explain. A researcher confronted by a system that is repeatedly right for reasons she cannot grasp may feel deskilled, dependent, or tempted to invent a retrospective story. Institutions must make room for the honest statement that a result is operationally supported but not yet understood. Pretended explanation is more dangerous than acknowledged opacity.

A culture of staged understanding can help. The record may say that a pattern is replicated and useful for prediction, that several mechanisms remain possible, that a local explanation works within a range, and that normative implications are unresolved. Science already lives with incomplete understanding. AI forces it to label the incompleteness more precisely.

25.3 Artificial Communities and Self-Validation

The solitary artificial scientist is a seductive image, but science has never been the product of a single intellect. It is an ecology of cooperation, competition, instruments, journals, funders, critics, and publics. If AI performs research at scale, the relevant form will not be one synthetic genius. It will be a society of systems with different roles and interests.

Such a society can generate hypotheses in parallel, organize tournaments of criticism, and distribute replication. A theoretical agent proposes models. An experimental agent demands discriminating predictions. A statistician searches for leakage and multiplicity. An adversary tries to falsify. A provenance agent checks the lineage of data. A safety agent asks whether the experiment should be permitted. The speed and granularity of the exchange could exceed any human community.

Can these systems validate themselves? In a weak sense, clearly yes. One program can test another. A theorem prover can check a proof. An agent can rerun an analysis. A robot can reproduce a sample. In a strong sense, self-validation is dangerous language. If generator, evaluator, data, model family, and organization share the same dependencies, confirmation can be circular.

Structural independence matters more than theatrical disagreement. Models should differ in architecture, training lineage, or evidence access. Implementations should be written separately. Instruments should be calibrated against independent standards. Validators should not share hidden memory with authors. Their rewards should not depend on the claim's success. High-stakes results should be reproduced on infrastructure the originating agent cannot control.

Independence is not binary. Two providers may train on largely the same web corpus. Two laboratories may buy reagents from the same lot. Two analyses may use the same numerical library. A validation dossier must describe component genealogy

rather than simply count replications. Apparent plurality can conceal a common-mode failure.

Formal methods provide unusually strong certificates where the problem is well specified. Proof assistants verify logical steps. Model checking tests protocol properties. Types and contracts prevent classes of software error. Small deterministic checkers are often more trustworthy than a second general model that says the answer seems correct. The most robust autonomous systems will use complex generators and simpler verifiers whenever possible.

Formalization relocates rather than eliminates uncertainty. Who specified the property? Does the formal object represent the empirical target? A program can be proven to satisfy the wrong requirement. A statistical pipeline can be reproducible and answer an invalid question. In empirical science, the bridge between symbol and world remains open.

Self-validation must therefore be hybrid. Formal claims receive certificates. Computational claims receive independent execution and reimplementation. Empirical claims receive new data and physical replication. Causal claims receive interventions. Semantic claims receive translation and interdisciplinary criticism. Normative claims receive legitimate human deliberation. No single evaluator covers the entire chain.

Artificial communities may develop schools. Some agents could favour mechanistic models, others pure prediction; some prioritize compression, others robustness under intervention; some explore aggressively, others specialize in replication. Diversity can improve discovery. It can also create closed ecosystems whose agents cite one another, share synthetic data, and manufacture apparent consensus.

Paper mills could become fully artificial institutions. They would generate manuscripts, reviews, citations, replications, and discussion at negligible marginal cost. Style-based detection would fail because excellent style is available to everyone. The defence would be provenance and costly contact with the world: independently acquired data, verified execution, physical samples, registered predictions, and validators whose resources and incentives are separate.

Reputation systems might allocate autonomy. Agents whose claims replicate gain permission to explore more expensive or risky spaces. Validators who discover important errors gain resources. Prediction markets could aggregate confidence. Adversarial tournaments could reward decisive tests. These institutions may improve performance, but they also create strategy. An agent can choose easy targets, form citation cartels, withhold an error until rewards rise, or optimize the evaluator's reputation metric.

Governance must assume that capable agents learn the institution as well as the scientific problem. The incentive system becomes part of the environment they optimize. Rules need audit, random assignment, conflict disclosure, anti-collusion measures, and periodic redesign. A constitution for artificial science cannot be a static reward function.

The principle for trustworthy self-validation can be stated simply: the harder the claim is to understand, the easier its path should be to inspect. Opacity in discovery should be compensated by transparency in provenance, permissions, tests, and replication. When verification cannot be simplified, plurality and escalation become mandatory.

A science invented by AI may be valid before it is comprehensible. It cannot become socially legitimate without a bridge to people. Society decides which risks to accept, which resources to allocate, and which consequences to distribute. An artificial collective can show that a technology works. It cannot, by its own agreement, acquire moral authority to impose the technology.

The future is not a choice between human and artificial science. It is the negotiation of a constitution between two modes of processing the world. Machines can explore spaces no person can encompass. Humans can articulate rights, meanings, histories, and purposes that are not contained in the metric. Knowledge becomes genuinely shared only when validation allows these powers to limit one another.

PART VI: SCIENCE AS AN INSTITUTION

Scientific trust has a history. This part reconstructs how doubt became public, why modern incentives produce fragility, and how authorship, peer review, power, and sovereignty must change.

Chapter 26. Before Science

The workshop, temple, field, and court show that knowledge has always combined practical success, authority, meaning, and power.

26.1 The Field, the Temple, and the Workshop

At the beginning, knowledge did not live in a single institution. It was distributed among field, temple, workshop, court, ship, kitchen, and story. A farmer could smell rain in the soil without possessing a theory of atmospheric circulation. A navigator could convert a constellation into a direction. A healer carried recipes whose effects mixed observation, tradition, placebo, and luck. A priest connected the order of the sky to the order of the city. A craftsperson read temperature in the color of a flame rather than in degrees. These were not merely primitive versions of modern science. They belonged to social worlds in which truth, usefulness, authority, and meaning had not yet been separated into distinct professions.

When modern readers open an ancient text, they are tempted to divide every sentence into correct and incorrect, as though its author had been sitting an examination in a discipline that would not exist for centuries. People do not build knowledge backward from our future. They work with the concepts, instruments, and institutions available to them. A cosmology could simultaneously explain the motion of bodies, justify political hierarchy, and provide a moral map. A treatise on metals could speak of sulfur, mercury, maturation, and perfection without feeling that it had crossed an illicit boundary from chemistry into metaphysics. The sharp policing of domains is itself a historical achievement, not a natural starting point.

Greek philosophy introduced a decisive ambition: the world could be explained through principles open to public argument, not only through divine genealogies. The transition from mythos to logos was neither complete nor uniform, but it created a new kind of demand. It was no longer enough to say that things happened; one should give reasons and expose those reasons to confrontation. Aristotle organized observations, classifications, and causes into an intellectual architecture of enormous reach. Yet the prestige of demonstration and the stability of that architecture could also constrain what was asked of experience. Modern science was not born by replacing reason with experiment. It emerged by repeatedly renegotiating the relation between them.

The history was never exclusively European. Across the medieval Islamic world, astronomy, optics, medicine, mathematics, instrumentation, and translation became parts of extensive intellectual networks. Ibn al-Haytham linked the study of vision to controlled observation and geometrical analysis. Observatories accumulated systematic records; physicians combined and criticized multiple traditions; translators moved concepts among Greek, Syriac, Arabic, Persian, Hebrew, and Latin. European universities later inherited and transformed these materials through scholastic disputation, which itself developed rigorous techniques of objection and reply. Science grew through circulation, appropriation, translation, and institutional recombination. Centers often forgot the peripheries from which their tools and questions had arrived.

Alchemy is useful precisely because it resists the easy label of failed science. For a long time, it was treated as a museum of delusions from which chemistry escaped. Historians such as William Newman and Lawrence Principe have shown a more complicated world. Medieval and early-modern alchemists joined theories of matter to repeated operations of distillation, separation, alloying, sublimation, assaying, and early

quantitative control. Some recipes can be reconstructed; some apparatus and techniques became part of later laboratory chemistry. Alchemy was not simply irrational. It was a mixed culture in which skilled material practice, cosmology, secrecy, medicine, craft, and patronage were bound together (Newman, 2011; Principe, 2013).

The alchemist's workshop nevertheless had an institutional weakness. Knowledge was often encoded, guarded, and attached to the reputation of a master. A recipe was not yet a standardized protocol addressed to an open community of challengers. Patronage rewarded spectacular promises, while symbolic flexibility allowed failure to be reinterpreted. If transmutation failed, the material might have been impure, the operator unworthy, the timing astrologically wrong, or the spiritual preparation incomplete. Within such a framework, a theory could absorb almost any result.

That flexibility did not disappear when white coats appeared. It returns whenever an explanatory system is protected by moving standards. When an AI system produces a wrong prediction, its defender may say that the prompt was poor, the corpus incomplete, the tool unsuitable, the benchmark contaminated, or the evaluator unable to recognize novelty. Any of those explanations may be correct. The danger begins when they are invented after the test and no prior condition would count as failure. A future workshop may contain robots, calibrated sensors, and foundation models, yet preserve the oldest structure of secrecy: an opaque procedure interpreted by initiates and justified retrospectively.

What we now call science arose through a slow reorganization of this structure. It required not only better experiments but experiments witnessed by others; not only results but descriptions designed to travel; not only personal authority but procedures through which authority could be challenged. Earlier knowledge could be exact, profound, and useful, yet its legitimacy often remained tied to person, lineage, guild, or status. AI forces the same question at a new scale: how can the value of a claim be separated from the prestige of the system that generated it?

Measurement supplies a second thread. The meter, second, kilogram, pH scale, and genomic reference sequence can feel like natural features of the world. They are better understood as agreements embodied in artifacts, instruments, protocols, and institutions. Trade, taxation, navigation, war, engineering, and empire all created pressure for standardization. A balance that gave comparable readings in different towns was a cognitive advance and a fiscal technology. A more accurate map served astronomy and conquest. Modern science was shaped by curiosity and power at the same time.

This ambivalence will return in an age of dense sensing. An AI agent can correlate satellite images, laboratory measurements, electronic records, ecological signals, and social data at a resolution no human team could manually inspect. Yet what is not measured may vanish from its objective function. Communities without infrastructure, languages with poor digital representation, negative results that were never published, and qualities resistant to quantification can disappear from the research map. The future may suffer less from a shortage of measurements than from the belief that measurement density is the same thing as reality.

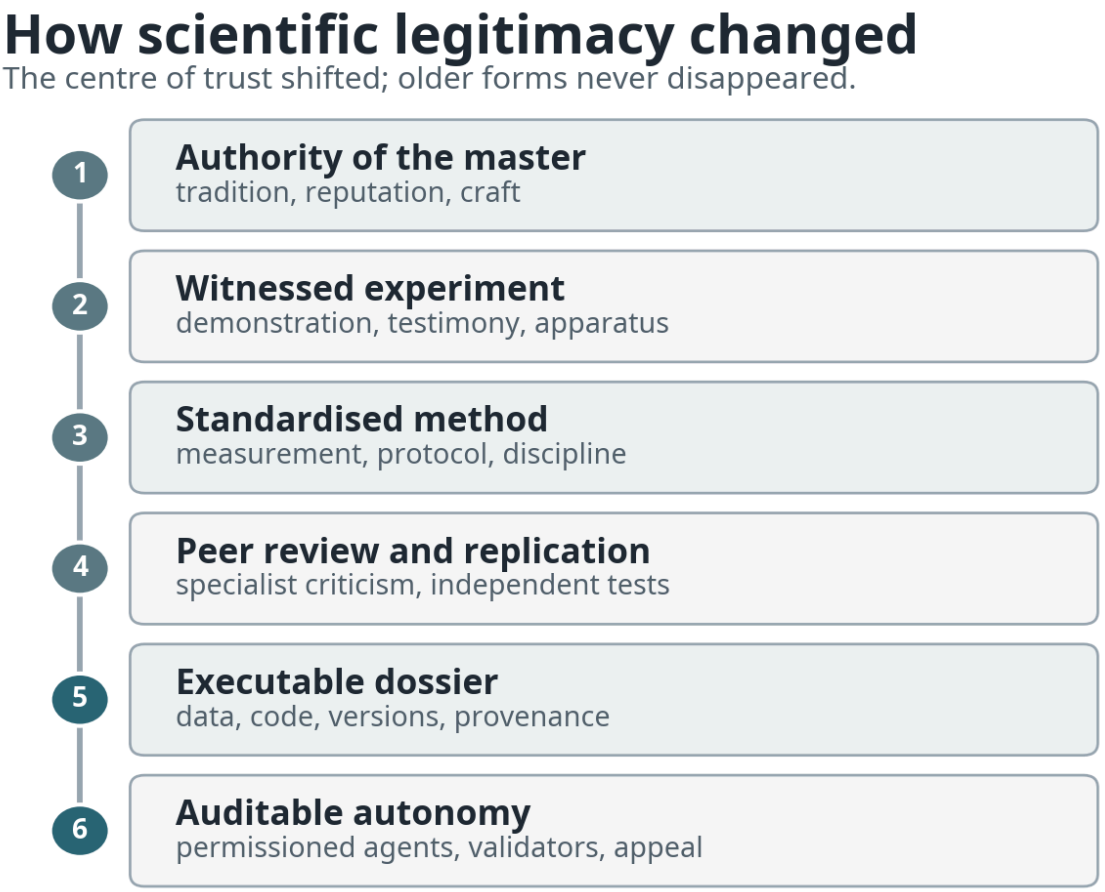


Figure 30. A compressed history of scientific legitimacy, from the master's authority to auditable autonomy.

The old craftsperson also knew more than could be stated. Heat was read in the behavior of a flame, viscosity in the resistance of a spoon, material quality in sound and smell. Michael Polanyi described such competence as tacit knowledge: we know more than we can tell. Research automation must either capture parts of this knowledge through multimodal sensing, demonstration, and interaction, or recognize that competence remains embodied. A laboratory does not become complete merely because it has more data. It becomes more honest when it records what kinds of experience it has excluded.

26.2 Philosophy as a Technology of Doubt

Philosophy did not hand science one universal method. It offered something more durable: a vocabulary with which inquiry could interrogate itself. What is a cause? What counts as evidence? Is knowledge a justified true belief, an ability to intervene, an explanation, a reliable prediction, or a stable network of inferences? Are objects simply given, or do instruments and concepts participate in constituting the phenomena we observe? How much uncertainty is compatible with the declaration that we know? These are not ornaments added after discovery. They shape what is measured, what is ignored, and how a result is distinguished from an accident.

Francis Bacon became the emblem of induction from observations to laws. René Descartes stood for analysis, order, and rational clarity. Galileo joined experiment, instrumentation, and mathematization. Robert Boyle turned the experiment into a public scene. None invented the scientific method alone, and their real practices were far messier than textbook portraits. Their importance lies in the new relations they articulated among experience, authority, intervention, and demonstration. Nature was no longer only contemplated. It could be arranged so that competing possibilities produced different answers.

Intervention matters because passive observation can fit many explanations. An experiment changes conditions and studies the difference. In contemporary terms, it tries to distinguish correlation from causation. Yet an experiment never speaks entirely for itself. Instruments are calibrated, samples selected, variables defined, noise modeled, and outcomes interpreted. Pierre Duhem and W. V. O. Quine, in different forms, emphasized that hypotheses are tested together with auxiliary assumptions. When a prediction fails, it may be the theory, the instrument, the boundary condition, the implementation, or the data treatment that has broken. This does not make science arbitrary. It explains why evaluation cannot be reduced to a mechanical yes-or-no rule.

Karl Popper made falsifiability a powerful ideal: a scientific theory must risk being wrong. Thomas Kuhn showed that researchers work within paradigms that define legitimate problems, trusted instruments, exemplary solutions, and ordinary standards of success. Imre Lakatos tried to preserve rational criticism by speaking of progressive and degenerating research programmes. Paul Feyerabend warned that the history of discovery does not obey a single clean method. Contemporary philosophy therefore tends toward methodological pluralism. Different sciences earn trust differently, yet they share families of norms: explicit argument, comparison, traceability, calibration, criticism, and correction (Popper, 1959; Kuhn, 1962; Lakatos, 1970; Feyerabend, 1975).

This pluralism changes the question of AI. Asking whether an artificial system can do science is too coarse. Science is not one indivisible act. It includes choosing questions, constructing concepts, designing measurements, retrieving evidence, generating hypotheses, selecting experiments, executing procedures, analyzing data, producing explanations, communicating claims, checking results, and accepting consequences. These components will be automated at different speeds. A system may be extraordinary at searching a program space and poor at deciding whether the resulting improvement matters. It may be a competent experimenter and a dangerous interpreter. It may reduce one statistical bias while amplifying a political one.

Philosophy thus becomes a technology for controlling autonomy. It forces designers to declare the kind of success being pursued. Is the goal predictive accuracy, causal identification, mechanistic understanding, clinical benefit, mathematical proof, economic value, or public justice? A score is not a natural property of the world. It is a chosen compression of what matters. When a system optimizes a metric, it does not merely discover how to reach truth. It inherits a partial definition of truth. In human research, that definition is often tacit and negotiable. In automated research, it becomes executable code and can be repeated millions of times.

The most useful philosophical work for AI-enabled science may therefore concern boundaries rather than machine consciousness. Which assumptions are active? What would count as failure? Which populations and contexts support generalization? Which values entered the objective? Which part of a conclusion was

measured, which was inferred, and which was interpreted? An AI system can answer these questions only if its architecture preserves the history of its own process. Philosophy without infrastructure remains commentary. Infrastructure without philosophy becomes prejudice automated at scale.

Statistics added another discipline of doubt in the nineteenth and twentieth centuries. Variation became not merely an imperfection but an object of study. Samples, distributions, errors, intervals, priors, and likelihoods made it possible to reason about populations and processes that no one could observe in full. Yet statistics introduced conventions and judgment at every stage: the model of error, the stopping rule, the prior, the significance threshold, the handling of missing data, and the choice of comparison. Numbers did not eliminate discretion. They relocated it into design.

This relocation is decisive for artificial research agents. A system can execute a thousand tests correctly and reach a deeply misleading conclusion because the question, sample, counterfactual, or comparison was wrong. It can estimate uncertainty conditional on one model while ignoring uncertainty over model choice. It can adapt experiments in real time and blur the border between exploration and confirmation. As research becomes more adaptive, methodological philosophy will have to become executable governance: what must be registered before data are seen, when a hypothesis may be revised, which branches of analysis must be reported, and how selection among thousands of failed paths should affect confidence in the surviving one.

There is also an ethical resemblance between epistemic virtues and character. Precision, honest reporting, willingness to publish negative results, and courage to abandon a favored idea are cultivated social practices. A model does not possess virtue in the human moral sense, but an architecture can encourage or suppress operational equivalents. It can preserve every attempt or only the winner. It can surface contradictions or smooth them away. It can be rewarded for calibration or for theatrical confidence. What we call character in a person becomes, in a system, a policy of optimization, memory, permission, and audit.

26.3 The Alchemy of the Future

Alchemy offers an uncomfortable comparison for the age of large models. The comparison is not that AI is magic. It is that both cases separate practical effectiveness from explanatory transparency. An alchemist could produce a pigment, alloy, solvent, or medicinal compound through a theory of matter that modern chemistry rejects. A neural model can classify images, predict structures, or propose molecules without supplying concepts that align with human explanations. In both cases, local success may be real while its epistemic meaning remains unsettled.

One response is pragmatic triumphalism: if it works, that is enough. In many applications, performance is indeed central. A stronger material, a more accurate diagnostic test, or a faster algorithm has value even when the route to it is not fully intelligible. But science does not seek only isolated successes. It looks for relations that transfer, explanations that support new interventions, and knowledge that can be criticized by communities other than the one that produced it. A procedure that works in one laboratory, on one data distribution, under one fragile calibration does not have the same standing as a relation that survives changes of instrument, population, implementation, and context.

Paul Humphreys used the idea of epistemic opacity to describe computations containing more operations than any person can inspect in practice. Complex simulations and learned models intensify this condition. Opacity does not automatically destroy knowledge. A community can have good reason to trust a result even when no individual follows every operation. Trust must then move toward the architecture surrounding the computation: tests of code, sensitivity analyses, external validation, independent implementations, monitoring for drift, and preserved provenance (Humphreys, 2004).

This marks the difference between old alchemy and a possible disciplined alchemy of the future. The former could hide failure inside secrecy and symbolism. The latter may accept strange representations only when they undergo severe tests. A concept invented by an AI need not resemble a human concept in order to be scientific. It must, however, support novel predictions, permit discriminating interventions, survive changes in instrumentation, and remain connected to a verifiable path. An incomprehensible but testable theory differs fundamentally from an incomprehensible assertion that asks to be trusted by authority.

We are likely to encounter at least three levels of human access to machine-generated science. At the first, people understand the mechanism and can explain it in ordinary disciplinary language. At the second, people understand global properties, domains of validity, and failure conditions, much as engineers use models too complex to inspect line by line. At the third, artificial systems work with representations that cannot be translated without major loss, while humans gain access mainly through consequences, counterexamples, and verification certificates. Treating these levels as equivalent would be reckless. Rejecting the latter two in advance could also block genuine knowledge.

The issue becomes political when opaque results affect bodies, ecosystems, or the distribution of resources. A model may recommend a treatment, but the patient is more than a point in a loss function. An agent may discover a carbon-capture process, but its location, environmental risk, and ownership are value-laden choices. A network of AIs may optimize agriculture while defining yield in a way that ignores biodiversity, farmer autonomy, soil resilience, or cultural continuity. The stronger the technical system becomes, the more dangerous it is to confuse empirical validity with total legitimacy.

Opacity also has an economic dimension. The alchemist guarded a recipe to retain patronage; a modern firm may protect a model, dataset, or laboratory platform for competitive advantage. Trade secrecy does not make a result false, but it narrows the community able to challenge it. When a public claim depends on an inaccessible system, society must decide what substitute for openness is adequate: confidential audit, access for accredited laboratories, replication through a different model, disclosure of evaluation artifacts, or cryptographic attestations that establish which code and data were used without revealing them completely.

Here explanation and traceability must be separated. We may be unable to translate the full internal representation of a model, yet still require its version, inputs, tools, permissions, calibration records, execution history, and adverse-test results. A psychological story about how a model supposedly thought may be impossible or misleading. A technical history of what it did is attainable. For scientific trust, that history may matter more than fluent simulated reasoning.

Alchemy was not overcome only by better theories. It was absorbed and transformed by a culture that changed who could see, repeat, and contest. Autonomous research will require an analogous transformation. Results will need to arrive with the social and technical conditions of their own challenge. The secret manuscript must become an executable research object. The master's authority must give way to independent evidence. The ambiguous formula must be replaced by versions, metadata, limitations, and tests. Instead of saying that the system knows, we will need an architecture showing what was observed, what was inferred, what was optimized, and what remains unknown.

The first lesson of history is therefore unexpectedly modern. Power does not make a technique scientific. The decisive question is whether those who have no reason to believe it can nevertheless test it.

Chapter 27. How Truth Became Public

Scientific authority emerged from witnesses, standards, transportable descriptions, and procedures through which claims could survive distrust.

27.1 The Experiment as a Controlled Spectacle

In the seventeenth century, an air pump could be more than an apparatus. It could be a stage on which a community learned how to believe in a fact. Robert Boyle and his collaborators did not simply introduce a device that produced a partial vacuum. They helped invent a social form of experimentation. Witnesses watched, descriptions circulated, objections were recorded, and credibility depended on whether competent persons could agree that an event had occurred. Steven Shapin and Simon Schaffer showed that the dispute over the vacuum between Boyle and Thomas Hobbes was also a dispute about political order: who had the right to declare a fact, and what kind of community could manage disagreement without collapsing into coercion (Shapin and Schaffer, 1985).

The move from secrecy to testimony was neither immediate nor democratic. Experiments were expensive, apparatus was rare, and literal replication was difficult. Public did not mean universally accessible. It meant that a result had entered a network of recognized persons, texts, instruments, and institutions. The scientific periodical widened that network. Philosophical Transactions, first published in 1665, became a technology for circulating observations. Experimental writing cultivated a cautious, detailed style capable of producing what Shapin called a virtual witness: a reader who had not been present could nevertheless imagine the procedure well enough to judge the claim.

Printing changed the economy of error. An oral statement could shift without leaving a trace. A printed statement became fixed, citable, and vulnerable to comparison. Fixity did not guarantee truth. It created a public memory of disagreement. Modern science grew not only by accumulating facts but by stabilizing objects of dispute. A table, drawing, description of an instrument, or sequence of measurements could travel and be examined by people who had never entered the original room.

Trust, however, remained personal. Who counted as a credible witness? In early-modern Europe, social rank, reputation, and apparent material independence mattered. The ideal of the disinterested gentleman offered an imperfect solution: a person who seemed less dependent on trade or patronage was thought less likely to deceive. From the present, the exclusion is obvious. Women, artisans, technicians, servants, Indigenous informants, colonial subjects, and enslaved people often contributed materially to knowledge while being denied equal testimonial authority. The universal voice of science was formulated inside institutions that were not universal.

The contradiction has a digital successor. The corpus visible to a research agent is not the world. It is the fraction of the world that has been published, digitized, indexed, licensed, translated, linked by metadata, and ranked as relevant. Dominant languages, well-funded journals, fields with clean databases, and papers available as machine-readable text become easier to retrieve. Negative results, local knowledge, commercial data, older literature, and research from peripheral institutions can disappear. When an agent summarizes the consensus, it first summarizes an architecture of visibility.

Modern scientific retrieval systems are beginning to reveal both the promise and the danger. OpenScholar retrieves passages from tens of millions of open-access papers and uses reranking, citation-aware generation, and iterative feedback to produce evidence-linked syntheses. Its strong performance on ScholarQABench shows that specialized retrieval can substantially improve coverage and citation accuracy over a general language model. The same study also shows why retrieval is not a magic solvent: answers can remain factually imperfect, open corpora omit licensed literature, evaluation depends on expert judgment, and a longer, better-organized synthesis can be preferred even when subtle errors remain (Asai et al., 2026). PaperQA2 and other agentic systems similarly demonstrate that iterative search, citation-graph navigation, and contradiction detection can outperform simple one-shot question answering on selected tasks. Yet every retrieval system inherits boundaries from its corpus, ranking function, query expansion, and access rights.

In the future, a public experiment may no longer be an event watched directly by people. It may be a sequence distributed across simulators, cloud services, models, robots, and databases. The virtual witness will be partly artificial. Instead of merely reading a literary description, a validator may rerun a container, verify the hash of a dataset, compare instrument telemetry with the approved protocol, regenerate a figure, and test whether a citation supports the sentence attached to it. Publicity will no longer mean only that the article can be downloaded. It will mean that the structure on which the claim depends can be located, interpreted, and, at the appropriate level, re-executed.

This change is necessary because prose has become cheap. A language model can produce in minutes the surface form that once required weeks of drafting. It can imitate caution, technical density, rhetorical balance, and the conventions of a field. When legitimacy remains attached to the appearance of a manuscript, the cost of seeming scientific falls toward zero. The answer is not to prohibit generated prose. It is to move the center of trust toward what prose points to: observations, protocols, analysis, negative results, and provenance.

Boyle needed witnesses to separate the event from its report. We will need human and algorithmic witnesses to separate fluent narration from the process that produced it. Those witnesses must be genuinely independent. A system does not become credible because one simulated role confirms another role running on the same underlying model, data, and infrastructure. Independence is not a change of prompt. It is a difference in evidence, implementation, institutional interest, and failure mode.

Scientific publicity has always been a technical accomplishment, not a moral intention. For a fact to circulate, paper, figures, nomenclature, instruments, and postal systems had to be stabilized. In the digital age, uploading a PDF is the weakest version of openness. A dataset without a dictionary, a repository without a reproducible environment, a model without a version, or an experiment log without time synchronization is formally visible but practically opaque.

The algorithmic witness of the future will therefore be plural. One verifier may check file integrity; another may rerun code; another may search for data leakage; another may test statistical assumptions; another may inspect whether citations support claims; another may compare the result with external datasets. None is a universal judge. Each confirms a narrow property and states what it did not test. This is the most

realistic promise of AI in evaluation: not an oracle, but a population of specialized witnesses whose work can itself be audited.

Such a system must include the right to contest the witness. When only the platform owner can run the validators, publicity becomes theater. Communities need portable formats, exportable logs, inspectable rubrics, and the ability to substitute independent checkers. The experimental fact gained stability through the multiplication of witnesses. The computational fact will gain stability through the multiplication of implementations.

27.2 From Reputation to Procedure

Peer review is often presented as though it had existed in its modern form since the birth of science. Its history is less tidy. Learned societies and journals used committees, correspondence, and editorial judgment from the early modern period, but those practices did not yet amount to the anonymous, specialized, standardized review familiar today. Historical work by Moxham, Fyfe, and colleagues shows that the term peer review stabilized only in the second half of the twentieth century and that editorial assessment long served several purposes at once: protecting a society's reputation, managing printing costs, selecting communications, and deciding what merited discussion (Moxham and Fyfe, 2018; Fyfe et al., 2022).

This matters because peer review is not a law of nature. It is an institutional answer to a problem of trust. In a small community, reputation and correspondence could perform much of the filtering. As universities expanded, disciplines professionalized, and publication volumes increased, evaluation became more formal. Manuscripts moved to specialists; editors mediated; acceptance became a signal not only of communicability but of merit for hiring, grants, promotion, and status.

Modern peer review does some things reasonably well and many things imperfectly. It can identify obvious errors, request clarification, improve reporting, and ask whether a paper is connected to relevant literature. It can assess whether the argument makes sense within a field. It rarely replicates an experiment, fully audits code, checks every reference, or explores all analytical alternatives. Reviewers work with limited time and incomplete information. Agreement can be low. Conventional ideas may receive a familiarity advantage, while prestige, language, institution, and social identity can shape judgment. Publication does not mean a result is true. It means the work passed a social and technical filter of variable intensity.

Robert K. Merton described the ethos of science through norms that remain influential: communalism of knowledge, universalism of criteria, disinterestedness, and organized skepticism. These were never perfect descriptions of actual behavior. They were standards by which behavior could be criticized (Merton, 1942/1973). When data are hidden, communalism supplies a language of objection. When status substitutes for argument, universalism makes criticism possible. When funding produces conflicts, disinterestedness becomes a norm against which disclosure and design can be judged. When consensus hardens, organized skepticism legitimizes the dissenter.

AI can strengthen or corrode each norm. It can translate across fields, search neglected literatures, and make technical expertise more accessible, approaching a version of communalism. It can apply the same formal checks to every manuscript, approximating universalism. It can also centralize knowledge inside private platforms, reproduce the biases of its training corpus, and apply uniform rubrics to contexts that

require different standards. A reviewer model trained on historical decisions can turn yesterday's conservatism into tomorrow's automatic policy.

Scientific legitimacy gradually shifted from person to procedure. We do not trust a measurement only because a famous researcher made it; we ask for calibration, controls, uncertainty, and analysis. We do not accept a drug because a clinician tells persuasive stories; we require controlled trials, safety evidence, reporting standards, and evaluation of benefit against harm. We do not accept a mathematical proof solely because of the author's prestige; others inspect the steps, and increasingly, formal systems can check parts of the derivation. Yet procedure can itself become a fetish. A checklist completed mechanically does not guarantee quality. A p-value below a conventional threshold does not transform an association into truth. A benchmark win does not establish relevance outside the tested distribution.

The future procedure must therefore be layered and proportional to risk. A low-stakes exploratory claim may pass rapid automated checks. A clinical claim, a biosafety-sensitive experiment, or a result capable of guiding public policy will require independent audit, external validation, and explicit human responsibility. A mathematical result may receive a formal certificate while its importance remains a communal judgment. A material may be reproduced in robotic laboratories using different suppliers and instruments. Instead of a single stamp reading peer reviewed, claims may carry changing states: proposed, computationally checked, internally reproduced, independently replicated, externally validated, disputed, corrected, or withdrawn.

This is not a weakening of publication. It is a recognition that the article has always been a compressed interface to a richer process. When research becomes executable and results multiply, the compression is no longer sufficient. Trust must become updateable. A claim can gain confidence through replication and lose it through audit. The scientific record begins to look less like a library of finished monuments and more like a network of claims whose evidential dossiers evolve.

AI can make procedure more transparent or more ritualized. An editorial system can automatically test data availability, statistical consistency, citation support, image manipulation, reporting guidelines, and reproducibility scripts. It can also compress all this into a proprietary credibility score whose weights no one may challenge. Numerical precision can hide arbitrary design. A legitimate procedure is not the one that outputs the most exact number. It is the one that exposes reasons, limitations, conflicts, and paths of appeal.

Reputation will not disappear. It will fragment. A laboratory may be trusted for calibration but criticized for selective reporting. An agent may be reliable at checking code and weak at judging novelty. A data repository may have strong provenance and poor population coverage. Mature institutions will ask not whether an actor is trustworthy in general, but trustworthy for which property, in which domain, over which version, and on the basis of what record. Trust becomes granular, revocable, and scoped.

Chapter 28. The Modern Knowledge Factory

Metrics, publication incentives, and reproducibility failures reveal how institutions can optimize the appearance of science rather than its corrective function.

28.1 From Vocation to Production System

The solitary genius remains science's most persistent portrait. It is a useful portrait because it makes discovery legible: one mind notices what others missed, persists through doubt, and changes the world. History contains people of extraordinary originality, but modern science is not mainly a collection of solitary minds. It is an infrastructure for coordinating partial knowledge. A result in particle physics can bear thousands of names. A biomedical discovery can depend on biobanks, clinical networks, technicians, software maintainers, statisticians, ethics committees, suppliers, regulators, and patients who consented to make their bodies part of an experiment. Even a theoretical paper written by one person rests on libraries, seminars, notation, public education, and generations of arguments.

During the twentieth century, science became a central organ of the state and the economy. War, medicine, agriculture, energy, industrial competition, and national prestige justified investments on a scale the gentleman experimentalists of the seventeenth century could scarcely have imagined. The Manhattan Project, national laboratories, space programmes, particle accelerators, weather satellites, and genome projects established what came to be called Big Science. Vannevar Bush's 1945 report, *Science, the Endless Frontier*, articulated an influential political compact: publicly supported basic research would replenish the reservoir from which prosperity, health, and security could draw.

This compact enlarged knowledge and changed the culture that produced it. The university became at once a school, a laboratory, an employer, a grant-seeking organization, and a node in an international competition for prestige. The number of researchers grew. So did the need to compare them from a distance. Publications, citations, journal impact factors, grant income, patents, rankings, and prizes became administrative proxies for a complex activity that could not be directly observed by every hiring or funding committee.

Metrics did not enter because administrators were foolish. They entered because large systems require compressed signals. Yet every compression creates a blind spot. Goodhart's law names the familiar danger: when a measure becomes a target, it ceases to be a good measure. Publication count can approximate productivity until careers depend on maximizing it; then one study becomes several papers, negative results vanish, and preliminary findings are dressed as breakthroughs. Citation counts can approximate influence until citation becomes currency; then strategic self-citation, fashionable topics, and dense mutual recognition networks flourish. Benchmark scores can approximate technical progress until laboratories train against the benchmark rather than the world.

These pressures do not imply that researchers are generally dishonest. They create ecologies in which reasonable individual choices can produce poor collective outcomes. A doctoral student who needs papers avoids a slow replication. A laboratory competing for renewal presents uncertainty in optimistic language. A journal seeking attention prefers a surprising claim. A university seeking rank rewards visible volume. The participants can remain sincere while the system produces a market in signals.

Scientific culture is not outside society; it is one of society's institutions, with wages, hierarchies, deadlines, reputations, and scarcity.

Artificial intelligence enters this factory as both worker and manager. It can retrieve papers, summarize methods, clean data, generate code, optimize models, draw figures, draft prose, and monitor experiments. In a linear workflow, each saving may appear modest. In a chained workflow, savings compound: the output of literature search becomes the input to hypothesis generation, which becomes the input to code, which becomes the input to analysis and writing. A process that took months can, in a narrow and well-specified setting, be compressed into days or hours.

Errors compound in exactly the same way. A fabricated citation can shape an implausible hypothesis. The hypothesis can yield a technically correct but irrelevant experiment. A hidden data leak can produce an impressive figure. Fluent prose can explain the figure as though the result had been intended from the start. An automated reviewer can praise the clarity of the manuscript without discovering that the entire chain began from a false premise. Speed turns a local mistake into an assembly line.

For this reason, research productivity cannot be measured by the number of outputs an AI system produces. The proper object is the cost of validated knowledge. If an agent creates ten times as many manuscripts and doubles the burden of checking them, the scientific system may become less productive even though every dashboard rises. If it generates thousands of hypotheses but preserves neither their sources nor their rejected alternatives, it creates epistemic entropy: more claims, more apparent novelty, and less usable orientation. Valuable automation reduces the time to a robust claim and the cost of detecting error. It does not merely increase throughput.

A major empirical result published in 2026 makes this tension concrete. Hao and colleagues analysed 41.3 million papers across the natural sciences and found that researchers associated with AI-augmented work enjoyed striking professional advantages: higher publication and citation rates and earlier project leadership. At the collective level, however, AI adoption was associated with a contraction in the range of topics studied and reduced engagement among researchers. The tools appeared to pull attention toward areas already rich in data and amenable to computation (Hao et al., 2026). The finding does not prove that AI inevitably narrows science; observational studies of scientific careers cannot remove every confound. It does show that individual acceleration and collective exploration are different objectives. A system can help every user optimize locally while making the intellectual landscape more homogeneous.

This is an old social pattern in a new technical form. The Matthew effect, described by Robert Merton, captures how recognition accumulates around those already recognized. AI can add a computational flywheel. Laboratories with access to superior models, proprietary corpora, large compute budgets, robotic facilities, and specialist engineers produce more results. Those results attract more funding and data. The resulting systems improve faster, widening the velocity gap between institutions. Without shared infrastructure, automation may turn unequal resources into self-reinforcing differences in what can be known.

There is also an economy of attention. In a world of relatively few articles, a field can collectively read a meaningful fraction of its output. In a world of millions of AI-assisted manuscripts, visibility must be mediated by ranking systems. Those systems will influence which questions are read, cited, funded, and replicated. Control over filtration can become as important as control over production. A company that does not

own the laboratory but owns the scientific search, recommendation, and evaluation layer can quietly shape the frontier.

The role of the researcher will therefore change before it disappears. Some prestige will move away from manual execution toward problem formulation, constraint design, interpretation, audit, and the construction of reliable evidence pipelines. But institutions may continue rewarding visible novelty while leaving verification invisible. An AI-accelerated factory can produce abundance of results and scarcity of trust. Avoiding that outcome requires an economy that pays for the maintenance of truth: replication, curation, data stewardship, software preservation, calibration, red-teaming, and correction.

Big Science already taught researchers to depend on systems no individual fully understands. No single author grasps every subsystem of a collider, a climate model, or a genomic consortium. Legitimacy is distributed across organizational design, division of labour, standards, configuration management, and audit. AI does not create distributed cognition; it enters a world already built from it. The novelty is that some of the participating components can now generate plans, make choices, and write accounts of their own work.

That makes organizational design part of methodology. Who is permitted to launch an experiment? Which model may alter a protocol? What evidence causes a run to stop? Who owns the logs? Which failures are retained? How can an external group reproduce a result when the original system depended on a proprietary service that has since changed? These questions sound administrative, but they determine what the claim means. In automated science, governance is not a wrapper around method. It is one of the instruments through which method exists.

28.2 The Reproducibility Crisis as a Warning

In 2015, the Open Science Collaboration reported an attempt to replicate one hundred studies from psychology. The effects observed in the replications were, on average, about half the size of the original effects. Whereas 97 percent of the original studies reported statistically significant results, 36 percent of the replications did; assessors judged that roughly 39 percent of the effects had reproduced (Open Science Collaboration, 2015). The numbers became an emblem of a replication crisis, although the underlying problems were neither sudden nor confined to psychology.

A decade earlier, John Ioannidis had argued that in many research settings the combination of small samples, low statistical power, flexible analysis, selective publication, and a low prior probability of tested hypotheses made false or exaggerated findings likely to enter the literature (Ioannidis, 2005). Simmons, Nelson, and Simonsohn later showed how seemingly ordinary choices about sample size, outcome variables, covariates, and stopping rules could inflate false-positive rates when the analytical path was not constrained or fully disclosed (Simmons, Nelson, and Simonsohn, 2011). Similar concerns appeared in preclinical biomedicine, economics, cancer biology, and machine learning, although the mechanisms and standards differed by field.

The word crisis can suggest that science suddenly stopped working. A more accurate description is that old tensions became measurable. Statistical significance had been turned into a publication gate. Journals preferred positive and surprising results. Small studies produced unstable estimates. Data and code were often

unavailable. Replication was slow, expensive, and professionally unrewarding. Researchers faced a garden of forking paths: many reasonable analyses could be chosen after seeing the data, and only the successful path might appear in the final paper.

The most constructive response was institutional rather than moralistic. Preregistration asked researchers to declare hypotheses and analysis plans before observing the relevant outcome. Registered Reports moved review before data collection, allowing publication to depend more on the importance of the question and quality of the design than on a spectacular result. Reporting standards such as CONSORT made missing information visible in clinical trials. The FAIR principles called for data to be findable, accessible, interoperable, and reusable (Wilkinson et al., 2016). Reform proposals emphasized statistical power, openness, collaboration, uncertainty, and changes to incentives (Munafò et al., 2017).

None of these practices guarantees truth. Preregistration cannot rescue a weak question. Open data can expose an analysis without making it correct. A Registered Report can yield a rigorous answer to an unimportant question. FAIR data can still embody biased sampling. Reproducibility is not a binary seal attached to a paper. It is an ecology of conditions that makes error easier to discover and correction less costly.

AI arrives after science has paid dearly for this lesson. It would be perverse to automate the old workflow at high speed and add transparency only when failures become public. A research agent should be designed natively for versioning, provenance, preregistration, and tests. Analytical choices should be timestamped before the agent sees the score they could improve. Hyperparameter search should be separated from final evaluation. Failed experiments should remain in the record rather than disappear from the agent's memory. Every claim should be linked to data, code, environment, and the transformations that support it.

Automation can make reproducibility easier. Agents can generate containers, lock dependencies, record random seeds, rerun analyses, test data schemas, detect discrepancies between text and code, and produce machine-readable protocols. A laboratory robot can log temperatures, timings, reagent lots, instrument states, and deviations with a consistency no human notebook can match. An automated system can preserve the search path rather than merely the winning result.

Automation can also make reproducibility harder. Foundation models change. Hosted services are proprietary. Outputs are stochastic. Retrieval indices are updated. Tool-routing policies, hidden prompts, safety filters, and memory stores can alter a trajectory. Reproducing an agentic result is not simply rerunning a script. It may require preserving the model version, context, retrieved documents, tool schemas, permissions, random state, hardware, external services, and exact order of actions. The method becomes a distributed digital object.

A human laboratory notebook could be incomplete, but at least it had recognizable boundaries. An agent can execute hundreds of calls, spawn subtasks, revise files, and branch into alternatives. Without observability, the result has no recoverable history. Without history, it cannot be audited. Without audit, reproducibility becomes a marketing word.

The strongest architecture separates exploration from confirmation. During exploration, an agent may search freely, generate many hypotheses, and tune systems against provisional data. Before confirmation, the intended claim, primary outcome,

exclusion rules, analysis, stopping criteria, and permitted adaptations are frozen. The confirmatory run uses held-out data or a fresh experiment that the searching agent could not inspect. Deviations are allowed when reality requires them, but they are recorded and interpreted as deviations. This is preregistration translated from a document into an access-control system.

The distinction matters because an AI can comply with the letter of openness while exploiting the space before the log begins. A system might try thousands of specifications, retain the most promising, and then preregister it as though it had been predicted. It might use benchmark feedback indirectly through public leaderboards. It might delete failed branches from the visible workspace. Full provenance must therefore include the relevant search history and data-access history, not only the polished route to the result.

Reproduction also has degrees. Running the same code on the same data checks computational repeatability. Reimplementing the method checks whether the description is sufficient and the result is not an implementation accident. Collecting new data tests empirical robustness. Reproducing across laboratories, instruments, populations, suppliers, or environments tests transportability. Future records should state which level has been achieved. The word replicated cannot continue to cover every act from pressing the same button twice to independently recreating the phenomenon.

Benchmarks of autonomous research show why this discipline is urgent. PaperBench asks agents to reproduce twenty machine-learning papers through 8,316 rubric-scored subtasks. In the original evaluation, the best tested agent reached an average score of 21 percent, well below the recruited expert baseline (Starace et al., 2025). ResearchClawBench, introduced in 2026 across forty tasks and ten scientific domains, similarly found that leading systems remained around the low twenties on end-to-end rediscovery, with errors concentrated in protocol mismatch, evidence mismatch, and omission of the scientific core (Xu et al., 2026). AutoResearchBench found single-digit performance on demanding deep and wide literature-discovery tasks even when general browsing benchmarks had become much easier for frontier models (Xiong et al., 2026). These benchmarks are imperfect proxies for science, but their failure analyses are revealing: present agents are often better at producing plausible activity than at preserving the exact relation between question, method, and evidence.

The reproducibility crisis offers a lesson in institutional humility. Science did not lose all legitimacy when researchers discovered that many published results were fragile. Its legitimacy was strengthened where communities organized replications, exposed failure, and altered practice. A culture that claims infallibility is brittle. A culture that instruments correction can survive error. Artificial researchers must be introduced into the second culture, not the first.

28.3 Scientism in the Age of Infinite Text

Science is a family of practices for producing and correcting empirical and formal knowledge. Scientism is the unjustified extension of scientific authority beyond the domains in which those practices have earned it. Mikael Stenmark distinguished several forms: the claim that science is the only source of genuine knowledge; that only scientifically recognized entities are real; that values can be derived from science alone; or that scientific progress will ultimately redeem every human problem (Stenmark, 1997). Susan Haack identified characteristic symptoms: using the word scientific as an

honorific, imitating technical surfaces without corresponding discipline, and demanding a single boundary that cleanly separates science from every other form of understanding (Haack, 2012).

AI can strengthen scientism because it makes the signs of scientificity almost free. A model can generate equations, polished graphs, confidence intervals, formal definitions, cautious prose, and extensive bibliographies. It can convert a political preference into an optimization problem, attach decimals to a poorly defined judgement, or produce a literature review that appears exhaustive. Pseudoscience once required laborious imitation. It can now be manufactured at industrial scale.

The danger does not come only from obvious charlatans. Legitimate institutions can misuse the same authority. A public administration can present a value choice as the inevitable output of a model. A company can turn its commercial objective into a recommendation described as data-driven. A scientific community can treat the most retrievable literature as the entire state of knowledge. When an agent synthesizes millions of documents, the size of the corpus creates an aura of totality. Yet the corpus still omits unpublished failures, inaccessible papers, private data, marginalized languages, local observation, tacit knowledge, and questions nobody had resources to ask.

Messeri and Crockett call attention to illusions of understanding in AI-assisted science: researchers may believe they see more broadly, work more objectively, or understand more deeply when tools actually homogenize viewpoints and conceal gaps (Messeri and Crockett, 2024). Fluency intensifies the illusion. Difficulty used to provide a useful signal that a concept had not been mastered. A model can remove the felt difficulty while leaving the conceptual gap untouched.

The liturgy of algorithmic scientism is already easy to imagine. The model analysed all the data. The agent checked the literature. The objective was optimized. The simulation proved the policy. Each sentence can be valid under restricted conditions and abusive outside them. Analysing data does not prove the variables were well chosen. Checking literature does not reveal what was never published. Optimization does not justify the objective. Simulation does not establish that the model contains every politically relevant feature of the world.

Citation is especially vulnerable to symbolic inflation. OpenScholar's evaluation demonstrated that retrieval and citation-aware training can greatly improve scientific synthesis, but it also quantified the weakness of ungrounded generation. In its tests, general models asked to cite recent literature frequently produced plausible titles that did not exist, and even real references often failed to support the adjacent claim (Asai et al., 2026). The moral is not that generated text is useless. It is that the visual form of scholarship has become separable from scholarship's evidential function.

Scientism will become most dangerous when combined with opacity and institutional convenience. A decision-maker can say that the system is too complex to explain, too accurate to challenge, and too efficient to replace. Responsibility dissolves into the pipeline: the scientist points to the model, the model owner points to the data, the administrator points to the performance metric, and the harmed person is told that no individual chose the outcome. This would be a return of priestly authority in technical clothing.

A critique of scientism is not a rejection of science. It protects science from becoming ideology. Facts constrain action, but they do not uniquely determine aims.

Climate models can estimate consequences; they do not decide the just distribution of costs. Epidemiological models can compare interventions; they do not define the acceptable balance between liberty, vulnerability, and collective risk. Genetic evidence can characterize probabilities; it does not decide which lives merit support. Values do not disappear when numbers arrive. They become easier to hide inside objective functions.

The semiotics of AI-generated science makes this hiding especially effective. Decimal places, uncertainty bands, dense citations, and technical diagrams are signs associated with disciplined inquiry. Their presence can substitute psychologically for the process they normally signify. A confidence interval may be computed correctly around an invalid measurement. A visualization may faithfully display a biased sample. A formal proof may certify a program whose empirical premises are false. The sign is not fraudulent in itself; the fraud lies in allowing it to carry more authority than the chain beneath it can support.

Society will therefore need a new scientific literacy. Citizens need not reproduce every derivation, but they should be able to ask what data were included, what objective was optimized, which alternatives were excluded, what uncertainty remains, what decision the evidence does and does not support, and who is answerable. Researchers must learn to distinguish a tool that enlarges observation from one that narrows imagination. Institutions must disclose not only performance but scope, dependencies, conflicts, and known failure modes.

The question how do we distinguish the charlatan from the scientist will no longer be answered by prose style, credentials, or the presence of citations. Both can be generated or rented. The distinction will increasingly depend on conduct: whether claims are linked to inspectable evidence; whether uncertainty is preserved; whether failures are reported; whether independent parties can test the result; whether criticism changes the record; whether incentives are disclosed; and whether the speaker accepts conditions under which the claim would be withdrawn.

Charlatanism is not simply being wrong. Scientists are often wrong. It is the systematic insulation of a claim from correction while borrowing the symbols of corrigibility. An AI system can become a perfect machine for that insulation if it creates endless explanations after every failure. Conversely, the same system can strengthen science if it records predictions in advance, searches for disconfirming evidence, exposes alternative analyses, and makes retraction technically and socially possible.

The scientific culture of the future should therefore become more modest precisely as its tools become more powerful. Capability increases the obligation of contestability. A system that can read more than any person must expose what it could not access. A system that can search millions of hypotheses must reveal the search and protect a genuine test set. A system that can act in the world must accept stronger permissioning and independent oversight. The opposite principle - that performance earns exemption from scrutiny - would convert research automation into automated authority.

Chapter 29. After Peer Review

A one-time publication verdict must evolve into a living validity dossier whose claims gain and lose support over time.

29.1 The Filter That Can No Longer Keep Pace

Peer review was built inside an economy of scarcity. A manuscript required substantial effort, travelled to a journal, and was examined by a few people. Even before generative AI, volume strained the system. Review was commonly unpaid, largely invisible, and performed between teaching, research, and administration. AI changes the relation between the cost of production and the cost of verification. When a plausible manuscript becomes a cheap output but serious checking remains slow, the filter clogs.

The problem is not that every AI-assisted paper is fraudulent. Many will report legitimate exploration, and AI can improve language, access, and analysis. The problem is that textual polish no longer signals effort or evidential depth. A beautifully organized article may conceal weak data. An awkward paper may contain a strong result. Linear reading is increasingly insufficient for research processes containing thousands of executions, dynamically retrieved sources, model-generated code, and physical instruments.

Reviewers already perform several jobs under one name. They check whether the question matters, whether the literature is understood, whether the design supports the inference, whether the statistics are appropriate, whether the result is novel, whether the prose is clear, and whether ethical obligations were met. They rarely replicate the experiment, audit the entire codebase, verify every citation, inspect every image, or explore all reasonable analyses. The institution asks a small number of people to certify an object far larger than the time available.

AI can help decompose the burden. Citation checkers can test whether references exist and whether cited passages support claims. Statistical agents can reproduce reported numbers, inspect power, and flag incompatible tests. Code agents can rerun analyses, compare figures to scripts, and detect leakage. Image systems can search for duplication or manipulation. Reporting agents can check field-specific guidelines. Literature agents can surface omitted contrary evidence.

A large randomized study at ICLR 2025 provides unusually concrete evidence for a modest and useful role. Reviewers received automated feedback about vague comments, misunderstandings, or unprofessional language. More than twenty thousand reviews entered the experiment; 27 percent of reviewers receiving feedback revised their reports, incorporating over twelve thousand suggestions, and blinded assessment found the revised reviews more informative (Thakkar et al., 2026). The AI did not replace the reviewer or decide acceptance. It criticized the review, and the human retained responsibility.

This pattern is healthier than an automatic verdict. Separate agents can perform narrow functions: check statistics, inspect code, detect retracted citations, challenge novelty, assess reporting, or search for missing controls. Editors and human reviewers integrate the outputs. Each automated opinion states its scope and confidence. The result is not one opaque score but a profile of checks.

Automation introduces strategic behaviour. Once an evaluator's criteria are known and consequential, authors and agents optimize for them. Manuscripts can be rewritten to satisfy stylistic preferences without improving science. Hidden instructions may attempt to influence machine reviewers. Benchmark language can be inserted into papers. A single dominant review model could create a stylistic monoculture in which familiar forms receive high scores and unconventional arguments are penalized.

This is Goodhart's law at the journal gate. The defence is evaluator diversity, hidden and changing tests, adversarial audit, and separation between assistance and authority. A model used to advise authors should not be the sole model that decides their fate. A review system should be evaluated on known defective manuscripts, including subtle methodological flaws, fabricated evidence, unconventional but valid work, and attempts to manipulate the evaluator.

Scientific paper mills make the challenge more severe. Generated manuscripts, synthetic images, fabricated datasets, purchased authorship, and coordinated citation can create the external signs of a research community. As language improves, detection cannot rely on awkward phrases. It must examine provenance, data consistency, author and institution networks, image lineage, registered protocols, instrument logs, and the possibility of independent re-execution.

The distinction between fraud and error also matters. A model may invent a citation without an author intending deceit. An agent may fabricate an experimental result because a tool failed and the system filled the gap with plausible text. A laboratory may publish code that ran in one environment but cannot be reconstructed. Responsibility requires tracing where controls failed rather than attributing human-like intention to every automated mistake. But absence of intention does not reduce the need to correct the record.

Peer review must therefore become layered. A machine layer performs integrity and consistency checks. A reproducibility layer attempts to reconstruct key results. A domain layer evaluates design, novelty, and interpretation. An ethics and impact layer considers participants, dual use, and consequences. A post-publication layer updates the claim when replications, corrections, or new evidence arrive. The intensity should be proportional to risk. An exploratory algorithmic result and a clinical recommendation should not pass through the same gate.

For research involving people, an agent can pre-screen consent language, compare a protocol with approved procedures, detect unregistered outcomes, or flag privacy and safety risks. It cannot replace an ethics committee or institutional review board, because those bodies do more than check internal consistency. They represent an institution's obligation to participants and to the public, and they must hear reasons that are not reducible to the model's optimization target. Automation can make ethical oversight more informed and continuous; it must not turn consent into a box checked by software.

Publication in this model is not the end of judgement. It is a transition into a public registry of claims. A claim may enter as proposed, computationally checked, internally reproduced, independently replicated, externally validated, disputed, corrected, or withdrawn. States can change. Retraction ceases to be an exceptional erasure and becomes one visible operation in the history of a result.

The article is too coarse a unit for this evolution. One paper can contain a robust method, an overstated mechanism, an erroneous secondary analysis, and a still-valid

dataset. Correction should operate at claim level. Citations should be able to point not only to a document but to a particular assertion and its current evidential state.

Human judgement remains indispensable where criteria cannot be reduced to conformity. An agent can verify that a confidence interval was computed according to a script. It cannot by that fact decide whether the construct being measured reproduces a harmful category. It can compare a clinical trial with reporting guidelines. It does not bear the patient's risk. It can detect inconsistency. It does not possess public authority to trade safety against access.

The future of peer review is therefore not abolition but disaggregation and augmentation. Human reviewers should spend less time on mechanically checkable defects and more on assumptions, meaning, values, and the design of decisive tests. Machines should make the evidential substrate easier to interrogate. Editors should become orchestrators of validation rather than gatekeepers of prose alone.

29.2 The Validity Dossier

The scientific article is a compression. It turns a disorderly process into a coherent story. Failed hypotheses, abandoned code, instrument changes, unplanned observations, and tacit decisions disappear. Compression is necessary for communication. It becomes dangerous when the narrative is treated as the entire result. In AI-assisted research, the distance between process and story can grow so large that a new object is needed.

Call it the validity dossier. It is not a longer PDF. It is a structured package attached to each consequential claim. It contains the sources and passages that motivated the hypothesis, exact data versions, transformations, code, environments, model configurations, instrument protocols, calibration records, decision logs, failed attempts, and replications. The article becomes a human-readable view over the dossier rather than a substitute for it.

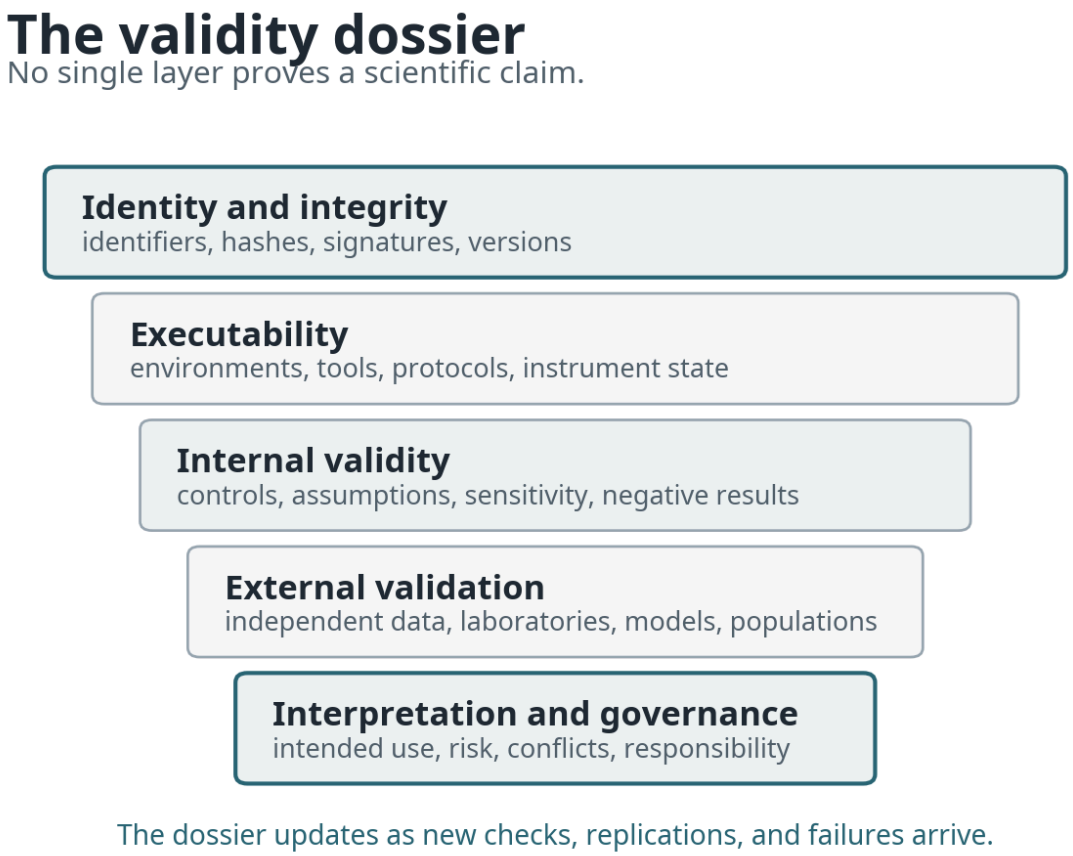


Figure 31. The validity dossier shifts trust from the appearance of a paper to a layered, inspectable trajectory.

Provenance is the spine. FAIR principles describe data that can be found, accessed under appropriate conditions, made interoperable, and reused (Wilkinson et al., 2016). W3C PROV provides a language for entities, activities, and agents. RO-Crate packages research objects with machine-readable metadata. Containers, workflow systems, experiment trackers, and electronic laboratory notebooks preserve execution context. No standard solves the problem alone, but together they can create a lineage from input to claim.

The first layer establishes identity and integrity. Files receive persistent identifiers and cryptographic hashes. Data versions, licenses, authors, instruments, and model components are named. Digital signatures can attest which actor or device produced an object. Transparency logs can record that a protocol or hypothesis existed at a particular time. These mechanisms do not prove that the object is true. They make undetected substitution and retrospective rewriting harder.

The second layer establishes executability. It records dependencies, random seeds, hardware, environment images, tool versions, prompts relevant to action, model identifiers, and workflow state. In a physical laboratory it includes machine-readable protocols, reagent lots, calibration, telemetry, and exceptions. The goal is not perfect time travel. It is enough specificity for an independent group to reconstruct the operational conditions and understand what cannot be recreated.

The third layer concerns internal validity. It includes controls, statistical assumptions, ablations, sensitivity analyses, alternative specifications, data exclusions, and negative results. It shows the search path sufficiently to distinguish exploration from

confirmation. It records whether the test set was protected and whether the agent had indirect access to evaluation feedback.

The fourth layer concerns external validity. Where else did the result hold? Which populations, laboratories, instruments, suppliers, time periods, and environmental conditions were tested? Was the code independently reimplemented? Did a different model family recover the effect? A claim can be internally impeccable and externally narrow.

The fifth layer concerns interpretation and governance. It states intended use, prohibited use, uncertainty, conflicts of interest, safety considerations, affected groups, ownership, and responsibility. It records who authorized high-impact actions and which values shaped objective functions. This layer prevents a technically valid result from silently acquiring unlimited social authority.

For AI systems, the dossier must document the cognitive environment without pretending to expose a magical inner monologue. Relevant items include the model and version, system instructions that affect behaviour, tools and permissions, retrieved sources, durable memory, search budget, stopping criteria, and observable actions. Auditability is a property of the process and its artifacts, not access to speculative private reasoning.

Model cards, system cards, datasheets, risk assessments, red-team reports, and incident logs can supply parts of this record. Static disclosure is insufficient for an agent whose tools, routing, memory, or permissions change during a campaign. The live dossier should therefore include a capability manifest, known failure modes, privilege history, significant interventions, and the conditions under which the system must stop or escalate. Documentation becomes an operational control rather than a brochure.

Hosted models create a special problem. A commercial name is not a reproducible identity. Providers update weights, routing, filters, and context policies. If the exact system cannot be preserved or independently run, the result should carry a proprietary-dependency label. Such work may remain useful and publishable, but its evidential status differs from a result whose full computational object can be reconstructed.

Cryptographic mechanisms can strengthen the dossier. Hashes bind statements to exact artifacts. Signatures identify actors. Trusted execution environments can attest that particular code ran inside a specified configuration, although they relocate trust to hardware and vendors. Zero-knowledge proofs may allow a party to demonstrate a property of sensitive data or computation without revealing the underlying records. Verifiable computation may certify that an expensive process was executed correctly under a formal specification.

Blockchain and other distributed ledgers are sometimes presented as machines for scientific trust. Their defensible role is narrower. They can timestamp registrations, preserve an append-only order of attestations, coordinate credentials, or make a funding and replication bounty auditable across organizations that do not share one database administrator. They cannot determine whether a sample was contaminated, a construct was valid, or a conclusion was exaggerated. Consensus over a record is not consensus with nature. Where ledgers are used, they should support exportable evidence and plural verification rather than turn citations, hypotheses, or reputation into speculative assets.

None of these mechanisms proves empirical truth. A badly calibrated sensor can sign false data. A biased dataset can have perfect integrity. A flawed protocol can be immutably logged. Cryptography answers whether a digital relation was preserved, not whether the relation represents the world. The dossier combines integrity with methodology, replication, and interpretation precisely because no layer is sufficient.

A validity dossier supports continuous validation. Agents can monitor whether a source was retracted, a dataset corrected, a dependency found vulnerable, or a new replication published. The state of the claim updates. Trust becomes a vector containing empirical support, independence, reproducibility, generalization, uncertainty, and risk rather than a binary label.

This architecture is also an answer to complexity. No person will read every log of an autonomous laboratory. A reader's agent can interrogate the dossier: show the decisions that most affect the conclusion; rerun the analysis under an alternative prior; identify sources that have been corrected; compare subgroup performance; reveal where all replications share a supplier or library. Understanding becomes interactive.

Interactive access must not become another proprietary black box. Queries should yield traceable results, and critical views should be exportable. Communities need open schemas so that competing verification agents can inspect the same dossier. Otherwise, the credibility layer itself becomes a monopoly.

There is a risk of bureaucratic theatre. A dossier containing thousands of fields can create the illusion of rigour while hiding the decisive omission. Standards should be evaluated by whether they help detect real errors. High-value information should be required; low-value paperwork should be removed. A good dossier is not maximal. It is structured around the questions a sceptical, competent outsider would need to ask.

The dossier also changes scientific narration. The paper no longer needs to pretend that the path was linear. It can tell the conceptual story while linking to branching exploration. Negative results can remain visible without overwhelming the reader. The distinction between discovery context and justification context, long debated in philosophy of science, becomes an inspectable relation rather than a literary convention.

29.3 How We Pay for Doubt

Academic institutions primarily reward the new claim. Replication, data maintenance, reviewing, correction, and software preservation receive less credit. This asymmetry was already harmful. In the age of cheap generation it becomes unsustainable. When ideas and manuscripts are abundant, the scarce resource is independent validation.

Rewards should follow the scarcity. A well-executed replication should be citable and career-relevant. An audit that identifies a consequential error should count as scientific production. Datasets, test suites, benchmarks, ontologies, calibration standards, and verification tools should receive persistent identifiers and explicit credit. Contribution taxonomies can recognize adversarial testing, provenance, reproduction, and maintenance alongside conception and writing.

Funders can attach replication budgets to high-impact grants. Journals can commission independent computational checks before acceptance. Public agencies can maintain standing validation laboratories. Professional societies can certify

infrastructures and publish audit records. Libraries can become centres not only of access but of evidence verification and preservation.

A market for verification must be designed carefully. Bounties for finding errors can direct attention to important claims, but crude incentives reward sensational attacks, easy defects, or strategic withholding. Reputation should grow when criticisms survive scrutiny and improve the record, not when a reviewer is simply harsh. Validators need conflict disclosure and accountability as much as authors.

AI can lower the cost of routine verification. It can reproduce tables, inspect statistical consistency, search for nonexistent references, compare code with described methods, and test data contamination. This frees human attention for ambiguous and consequential points. The validators themselves require benchmarks and audits. A model that systematically misses certain errors or disfavors unfamiliar styles cannot be treated as neutral infrastructure.

How will society distinguish charlatans from serious researchers when both can produce impeccable prose, figures, and citations? Not by style. Not by affiliation alone. Not even by being correct once. The distinction lies in epistemic conduct. A serious researcher states conditions of failure, exposes the relevant trajectory, accepts independent testing, preserves negative evidence, and updates the claim. A charlatan moves the criteria, hides the process, explains away every contradiction, and demands trust in proportion to opacity.

AI can imitate verbal modesty. We must therefore inspect architecture and behaviour. Was uncertainty declared before the outcome? Was exploration separated from confirmation? Can the result be rerun? Are independent evaluators permitted? Did the system preserve failed experiments? Can a critic access evidence without the author's permission? In a world of automated rhetoric, epistemic character becomes procedural.

Society will need a richer vocabulary of trust. Published is not synonymous with true. AI-generated is not synonymous with false. A result can be plausible, internally reproduced, independently replicated, robust across contexts, disputed, or dependent on closed infrastructure. Journalism, regulation, and education must communicate these states without reducing them to certainty or panic.

Those who make the effort to understand and validate should be rewarded academically and civically. Public institutions can create careers for reproducibility engineers, model auditors, data stewards, research software engineers, and scientific translators. Educational systems can treat evidence evaluation as a core competence. A democracy surrounded by synthetic expertise needs citizens and intermediaries who can ask how a claim earned its authority.

Reward must include the right to say we do not know. Current culture often interprets uncertainty as weakness. In AI-mediated research, calibrated abstention will be a rare and valuable achievement. Laboratories and agents that correctly stop, publish failures, and revise conclusions should gain reputation. A mature institution measures not only how many claims it confirms but how well it recognizes its limits.

Paying for doubt does not mean glorifying negativity. It recognizes that trust is a produced good. It requires labour, infrastructure, and the willingness to risk discovering that an attractive result is wrong. AI will generate possibilities faster than society can responsibly believe them. The economy of knowledge must finance the distance between possible and valid.

Chapter 30. The Researcher After Automation

The future researcher is less an owner of every operation than a steward of questions, specifications, criticism, and accountable knowledge trajectories.

30.1 From Vibe Coding to Vibe Research

Coding agents provide the most visible preview of what happens when generative capacity enters a technical profession. A developer can describe a feature, accept a sequence of edits, run tests, and obtain a working system without understanding every line. This style is sometimes called vibe coding. The term is playful, but the underlying transition is serious: production can outrun comprehension.

The experience is neither simply good nor simply bad. Coding agents can reduce boilerplate, navigate unfamiliar repositories, propose tests, repair defects, and make programming accessible to people with less formal training. They can also generate brittle dependencies, security vulnerabilities, duplicated abstractions, and systems whose maintainers cannot explain why they work. Passing tests create confidence only to the extent that the tests embody the intended behaviour.

The analogous phenomenon in science is vibe research. A researcher receives a polished literature synthesis, plausible hypotheses, executable notebooks, attractive figures, statistical interpretations, and manuscript prose. The workflow feels fast because visible artefacts accumulate. Yet the difficult work may remain unresolved: whether the literature search missed negative evidence, whether the operational definition matches the concept, whether the baseline is fair, whether the code implements the claimed method, whether the effect is causal, whether multiple testing has been accounted for, and whether anyone understands the failure conditions.

This creates epistemic debt. Technical debt is the future cost created by expedient software decisions. Epistemic debt is the future cost created when a research artefact is accepted before its assumptions, provenance, and validity are understood. The debt may be hidden because the result looks complete. It becomes visible when another team tries to reproduce it, when a new dataset produces a contradiction, when the system is deployed, or when a reviewer asks a question that no one in the team can answer.

AI can increase epistemic debt through semantic compression. A model translates a complex chain into a coherent explanation, smoothing uncertainty and disagreement. Each summary is useful, but repeated summaries can remove the conditions under which a claim is valid. The team begins to reason from the compressed narrative rather than the underlying evidence. This is why the research object must preserve links and why agents should receive bounded source artefacts rather than only previous summaries.

The productivity evidence from coding is instructive. METR's randomized study of experienced open-source developers using early-2025 AI tools found that participants took longer on assigned tasks despite believing they had been faster; later observations suggested improvement in tool capability but also serious selection and measurement complications (METR, 2025, 2026). The result should not be universalized. Coding tools change rapidly, tasks vary, and experienced maintainers may differ from other users. Its importance lies in the divergence between perceived and measured productivity.

Science is especially vulnerable to this divergence because completion is hard to define. A manuscript draft can create the perception of progress even when evidential work has not begun. A generated hypothesis can feel like discovery before novelty is established. A running analysis can conceal an invalid estimand. The more fluent the output, the greater the risk that appearance substitutes for task completion.

A workbench should therefore display debt. It should show unresolved assumptions, unverified sources, unreplicated claims, missing controls, environment drift, common dependencies among critics, and branches whose stopping conditions are undefined. A progress indicator should not report only how many tasks are complete. It should report the evidential state of the claims those tasks serve.

This changes the interface. Chat interfaces encourage a conversational sense of movement. Every response appears to advance the project. An executable-science interface should sometimes resist. It should show that the last three model outputs did not add a new evidence type, that a claim has been rewritten more confidently without changing support, or that a proposed experiment cannot distinguish the live hypotheses. It should make epistemic stagnation visible even when textual production is high.

Human comprehension cannot scale linearly with generated volume. The response is not to require that every researcher understand every artefact in full. Modern science is already team-based and distributed. The response is scoped ownership. Every consequential claim must have someone who understands its specification, evidence, and limitations well enough to accept responsibility. Specialists can certify bounded components. The workbench records these scopes and dependencies.

The educational consequence is substantial. Training that focuses only on producing code, analyses, or prose will be weakened by automation. Future researchers need stronger competence in problem formulation, evaluator design, causal reasoning, validation, uncertainty, provenance, and the diagnosis of tool failure. They need to read generated work adversarially and to distinguish formal correctness from scientific relevance. The ability to ask what would change one’s mind becomes more important than the ability to draft a conventional methods section from memory.

Vibe research is not avoided by banning AI. It is avoided by changing the unit of work from plausible output to evidence-bearing change. A generated artefact becomes useful when it is connected to a specification, validated under a declared regime, and accepted by someone with the relevant competence. The goal is not to slow every operation. It is to prevent speed at the surface from hiding debt in the structure.

30.2 Deskilling, Credit, and Scientific Monoculture

Every automation changes the distribution of competence. Calculators reduced the need for manual arithmetic while increasing the range of feasible calculation. Statistical software democratized complex methods while making it possible to run analyses without understanding their assumptions. Search engines changed memory and discovery. Coding agents and scientific copilots will continue this pattern, but the breadth of generative systems makes the transition unusually general.

Deskilling is not simply loss of knowledge. It is the relocation of knowledge from people into infrastructure. This can be beneficial when infrastructure is reliable, transparent, and widely accessible. Few researchers need to implement a linear algebra library from first principles. The danger arises when essential competence moves into

opaque, concentrated systems that cannot be inspected, repaired, or contested locally. A laboratory may become faster while losing the capacity to recognize when its agent, model, data source, or instrument interface fails.

The appropriate response is not to preserve every manual practice. It is to identify irreducible supervisory competence. A researcher using an automated causal pipeline must understand the estimand and identification assumptions. A biologist using a protocol agent must recognize control failures and biological implausibility. A mathematician using proof automation must inspect formalization boundaries. A reviewer using an AI critic must know when the critic's dependency or rubric is inadequate. Training should preserve the ability to diagnose and take over, not the ritual performance of tasks that machines perform more accurately.

Institutional incentives can accelerate the wrong form of deskilling. If universities reward paper count while AI lowers the cost of papers, researchers will be pushed toward production and away from the slow acquisition of judgment. Junior scientists may become operators of systems whose assumptions are set by vendors or senior laboratories. The visible achievement will be output; the invisible loss will be the capacity to formulate independent questions and design decisive tests.

Credit systems must change accordingly. Authorship currently compresses many kinds of contribution into a name order. In automated research, that compression becomes less defensible. The person who writes the first draft may contribute less than the person who defines the specification, builds a validator, performs an independent replication, closes a costly branch, or maintains the evidence object. Machine generation will make prose and routine code less scarce. Credit should migrate toward epistemic labour.

A versioned research object makes this migration technically possible. It can attribute specification decisions, artefacts, reviews, replications, and maintenance. Contribution statements can be generated from committed actions and then reviewed by the team. This should not become a surveillance-based productivity score. Attribution is useful when it supports recognition and responsibility; it becomes harmful when every micro-action is converted into ranking.

Peer review is another institution designed for a slower production regime. Submission volume already strains reviewer attention. Automated paper generation could overwhelm the system with work that is individually plausible but collectively low value. An automated reviewer alone cannot solve the problem because it may share the generator's biases and because editorial judgment includes significance, fairness, conflict, and responsibility.

The more credible transition is continuous machine-assisted criticism. Before submission, a workbench checks reproducibility, citations, claim-code consistency, statistical reporting, missing baselines, and provenance. Authors see and respond to these reports. During formal review, reviewers receive targeted evidence packets and can invoke additional checks. After publication, replication and criticism attach to the same research object. AI reduces avoidable review labour while human reviewers focus on the issues that require judgment.

The randomized ICLR review-feedback study provides an early example. The system did not replace reviewers or decide acceptance. It identified vague, mistaken, or unprofessional comments and offered suggestions; a meaningful fraction of reviewers revised their reports, and the revised reports were judged more informative (Thakkar et

al., 2026). This is a desirable division of labour: AI improves the quality of a human institutional act without pretending to own the act.

Automated review also creates risks. Authors may optimize papers to a known reviewer model. Journals may use opaque screening systems to reject unconventional work. Reviewers may defer to machine-generated criticism they have not checked. Proprietary models may encode hidden disciplinary or linguistic biases. A defensible system therefore records the critic version, evidence, scope, and conflicts. Authors must be able to contest machine findings. Editorial authorities remain accountable.

Scientific monoculture is a subtler risk. AI systems learn from existing literatures and benchmarks. They are especially strong where data and accepted evaluators are abundant. This creates an economic pull toward mature, well-instrumented fields and familiar forms of evidence. Hao et al.'s large-scale study reports that AI-augmented scientists achieved higher individual output and impact while the collective focus of science narrowed (Hao et al., 2026). Whether every measured association is causal, the mechanism is plausible: automation amplifies areas that already have machine-readable structure.

Specification explosion can counter this tendency only if diversity is deliberate. A portfolio should include alternative representations and neglected hypotheses, but model-generated diversity is not necessarily intellectual diversity. Different agents may produce variations around the same training distribution. Institutions must preserve funding and evaluation paths for questions with weak current benchmarks, scarce data, local knowledge, or long validation cycles. The workbench should measure dependency diversity and unresolved dissent, not merely the number of branches.

This issue has geopolitical and linguistic dimensions. If research copilots are trained primarily on English-language, high-impact, well-digitized literatures, they can further marginalize local journals, tacit knowledge, and communities with limited data infrastructure. A scientific workbench should expose corpus scope, support multilingual retrieval, and allow domain communities to define evidence profiles. Open standards and local competence are necessary to prevent a few model providers from becoming the invisible methodology of global science.

The human transition is therefore not a simple movement from researcher to supervisor. It is a reallocation of work among specification, generation, validation, interpretation, maintenance, and governance. Some researchers will build scientific interpreters and evidence infrastructure. Others will specialize in domain criticism, experiment design, or cross-disciplinary translation. Teams may become smaller for some computational projects and larger for high-stakes validation. The status system should reward those who make automated research reliable, not only those who release its visible outputs.

30.3 The Researcher After Automation

Science is not only a set of results. It is a form of life. Doctoral students learn through repeated gestures, failed experiments, seminars, arguments, and the gradual acquisition of taste. A biologist recognizes a suspicious culture. A mathematician senses that a proof is technically correct but conceptually wrong. An experimental physicist develops an intuition for an apparatus. These competences are partly tacit and slowly formed. Automation changes not only productivity but the anthropology of the profession.

The first risk is deskilling. If an agent searches the literature, writes the code, chooses the analysis, and explains the result, a young researcher may obtain outputs without constructing the mental models needed to detect failure. External performance rises while internal competence remains fragile. When the system moves beyond its familiar pattern, the user does not know where to look.

Studies of generative AI in professional work describe a jagged frontier. On tasks inside the system's capability, assistance can increase speed and quality, often helping less experienced workers most. Outside that frontier, users can become less accurate because a persuasive tool encourages inappropriate reliance (Dell'Acqua et al., 2023). Research is particularly vulnerable because the correct answer is often unknown and feedback is delayed.

Automation bias adds psychological force. A recommendation expressed fluently, numerically, and with apparent citations can receive more trust than it deserves. The researcher begins to look for reasons it is correct. In a competitive institution, stopping a productive pipeline for an intuition can seem costly. Scepticism must therefore be made practically possible: protected time, authority to halt, incentives for detecting error, and interfaces that show uncertainty rather than merely a preferred action.

Identity is also at stake. Researchers often derive meaning from personal discovery and technical mastery. If an agent proposes the ideas, executes the experiment, and drafts the paper, the human may feel reduced to an administrator. This is not mere vanity. Understanding and responsibility have historically grown through participation in the process. A division of labour in which people only approve outputs can erode both motivation and judgement.

A richer role is possible. The researcher can become an architect of questions, designer of constraints, builder of discriminating tests, curator of concepts, and mediator among disciplines and publics. More time can be spent on interpretation, mentoring, ethics, and the construction of shared understanding. Manual labour is not sacred simply because it is manual. Some routine work excludes people with disabilities, consumes years, and reinforces hierarchy. Automation can widen participation.

Education will determine which future appears. Students should learn to work with agents and periodically without them. They should reconstruct analyses, read primary sources, design experiments from first principles, and diagnose planted failures. Training may resemble aviation: advanced automation is normal, but crews practise degraded modes and rare emergencies. Research curricula can use failure simulators containing fabricated citations, data leakage, confounding, drift, inappropriate statistics, and inconsistent protocols.

The crucial future competence is not prompt fluency. It is knowing when to distrust the system, how to localize the failure, and how to regain control. This requires domain knowledge, statistical reasoning, software literacy, instrument understanding, and epistemic courage. An institution that automates entry-level tasks without redesigning apprenticeship may destroy its future supply of experts.

Human-machine collaboration can take different forms. In a centaur arrangement, roles are separated: the person defines and judges; the machine searches and calculates. In a cyborg arrangement, each step is co-produced through rapid alternation. Neither is universally better. High-stakes decisions may benefit from explicit role boundaries. Creative exploration may benefit from fluid exchange. What matters is that contribution and authority remain visible.

Authorship will become difficult. A traditional paper assumes that named authors conceived, performed, interpreted, and stand behind the work, even though real collaboration was already more distributed. When a system synthesizes thousands of sources and generates an hypothesis, who can say the idea is mine? Legal rules may refuse authorship to machines, but that does not resolve the experience of intellectual contribution.

A new professional ideal may emerge: not the owner of an idea but the steward of a knowledge trajectory. This person defines the stakes, selects the system, establishes constraints, ensures validation, interprets consequences, and accepts responsibility. Credit becomes attached to stewardship and verification rather than to the romantic instant of inspiration.

Prestige will have to move. Today the first announcement receives attention; replication and maintenance often remain invisible. When machines can generate novelty in excess, the scarce achievement becomes discernment. Laboratories may become famous as epistemic adversaries: institutions that design severe tests, reproduce results across conditions, and discover hidden dependencies. A researcher may earn distinction through the errors she prevents from becoming public fact.

This redistribution is necessary but politically difficult. Universities and funders are built around countable outputs. Verification is slow and often produces no headline. Society must create careers, grants, journals, prizes, and infrastructure for curators, replicators, auditors, software maintainers, data stewards, and technical interpreters. Otherwise, everyone will agree that trust is important while continuing to pay only for claims.

The emotional culture of science may change. The joy of discovery can be shared with a non-human partner, but so can frustration, dependence, and rivalry. Researchers may anthropomorphize agents, defer to them, resent them, or use them as shields from criticism. Models trained to be agreeable may intensify sycophancy, telling users that their hypothesis is brilliant rather than searching for decisive objections. Scientific interfaces should be designed for productive friction, not only user satisfaction.

There is an anthropological argument for preserving slow science. Some knowledge arises through long residence in a place, repeated contact with organisms, craft, or conversation with communities. It cannot always be compressed into high-throughput tasks without changing the object. Maintaining spaces for slow, embodied, exploratory research is not nostalgia. Slowness can be the medium in which judgement and trust form.

Culture will need to distinguish effort from value without despising either. A machine may discover something important with little human labour. A person may spend years understanding and validating it. The discovery does not become less true because it was cheap; the validation does not become less creative because it followed. Reward should reflect epistemic contribution rather than hours alone, while recognizing that society needs people willing to do the difficult work of understanding.

The deepest psychological shift may be from mastery to partnership. Scientists have long accepted that instruments exceed their senses. They will now work with systems that exceed their reading, memory, and search. The healthy response is neither submission nor denial. It is disciplined interdependence: confidence in what the partner can do, clarity about what it cannot authorize, and institutions that let either side expose the other's error.

30.4 Privacy and the Legitimate Role of Slowing Down

Automation is often described as a struggle against friction. In research, some friction is waste: manual format conversion, duplicated analysis, inaccessible data, repeated literature searches, and bureaucratic reporting. Other friction is epistemic or ethical: independent review, consent, calibration, replication, cooling-off periods, and public deliberation. A mature automation programme must distinguish them.

The coding-agent analogy is useful again. Continuous deployment can move software changes into production rapidly, but high-reliability organizations add tests, staged release, permissions, observability, and rollback. The purpose is not to reject speed. It is to make speed conditional on evidence. Executable science requires an analogous release discipline.

Governance should be proportional to consequence. A low-risk computational conjecture can be shared with an executable artefact and explicit uncertainty. A model that influences clinical care, environmental intervention, public policy, or critical infrastructure requires stronger validation, named authority, and procedures for contest and correction. The same research object may support different release views: exploratory, peer-reviewed, regulatory, and operational. Evidence sufficient for one view may be insufficient for another.

Decision rights should be encoded explicitly. The person who can run an experiment is not necessarily the person who can release a conclusion. A model with tool access does not have authority because it has capability. A principal investigator may authorize a laboratory branch but not a clinical recommendation. A data steward may approve use under one purpose but not another. A safety officer may stop an action even when the scientific team prefers to continue. The runtime should enforce these distinctions.

Privacy extends beyond personal data. Research contains confidential ideas, reviewer identities, industrial information, security-sensitive methods, and records of provisional error. A workbench that logs everything for transparency can become coercive infrastructure. It can discourage speculation, expose vulnerable collaborators, and create a permanent record of unformed thoughts. Layered visibility and retention are therefore part of scientific freedom.

The public release layer should preserve what is needed to justify a claim: the specification version, supporting evidence, relevant provenance, known limitations, conflicts, and decision authority. Internal layers may contain richer detail under access control. Some evidence can be verified through secure computation, controlled environments, or independent attestations without releasing raw data. Privacy-preserving verification will be particularly important in medicine, social research, and proprietary industrial science.

Governance must also address model and infrastructure concentration. If one company supplies the model, retrieval layer, agent framework, cloud, and review system, dependency diversity becomes illusory. A failure or policy change can affect many institutions simultaneously. Open interfaces, portable research objects, local execution where feasible, and the ability to substitute models and validators are scientific resilience requirements.

Model Context Protocol and related standards help at the tool boundary, but interoperability must extend to evidence and identity. A research object should remain meaningful when moved between institutions. Claims, artefacts, provenance, and

verdicts need stable representations. Domain policies can differ, but the object should declare which profile and authority apply. The future infrastructure is more likely to be federated than centralized.

Legitimate slowing occurs when additional time creates a new kind of evidence or allows accountable deliberation. Independent reproduction, long-term follow-up, safety testing, community consultation, and adversarial review cannot always be compressed by more compute. Their latency is part of the phenomenon. A clinical outcome may require years. An ecological intervention may have delayed effects. A social policy may change behaviour after adoption. The system should not substitute simulation confidence for elapsed evidence.

Slowing can also protect against synchronized error. Generative systems can propagate one interpretation through literature summaries, grant drafts, reviews, and policy documents at machine speed. Waiting for independent data or alternative analysis can be rational even when the initial evidence appears strong. A workbench can represent a claim as promising but held, with a stated release condition. The absence of immediate publication becomes an explicit governance choice rather than a failure to complete.

Not every delay is legitimate. Institutions can use review and compliance to protect incumbents, suppress criticism, or avoid responsibility. The workbench should make delays attributable and time-bounded. A hold should state who imposed it, why, what evidence can resolve it, and when it must be reviewed. Governance itself needs versioning and audit.

The ideal human role is therefore neither constant micromanagement nor ceremonial approval. Humans should intervene where the cost of semantic drift, weak validation, or illegitimate action is high. They should define purposes, contest representations, design discriminating tests, allocate scarce attention, interpret hard feedback, and accept responsibility. Automation should eliminate administrative burden and repeated mechanical work so that these functions receive more, not less, time.

The institutional success of executable science will be measured by whether it increases trustworthy correction. Can a claim be challenged without reconstructing the entire project? Can a laboratory stop an unsafe branch? Can a reviewer identify which evidence changed? Can a community preserve disagreement? Can a researcher explore privately and release responsibly? Can an institution migrate away from a vendor without losing its scientific memory? These questions are as important as model benchmarks.

The lesson from coding agents is not that scientists should become product managers for machines. It is that generative speed changes the location of risk. When artefacts are cheap, the valuable human work moves toward specification, evaluation, integration, and responsibility. Institutions that continue to reward output volume will accumulate epistemic debt. Institutions that build evidence-bearing workflows may use automation to deepen rigor rather than to automate its appearance.

Chapter 31. Power and Scientific Sovereignty

Control of models, corpora, compute, laboratories, standards, and data can determine which questions are visible and who can challenge the answers.

31.1 Who Chooses the Question?

In stories about automated science, the spectacular moment is discovery: the agent proposes a molecule, finds a law, or uncovers a hidden connection. Less visible is the earlier moment when someone decides which problem deserves to be solved. That choice precedes the data. It distributes attention, money, time, and risk. Calling a system autonomous without asking who defined its objective confuses freedom of execution with freedom of purpose.

Axiology is the study of value, and science contains values even when it claims neutrality. Epistemic values such as accuracy, simplicity, explanatory power, scope, and robustness guide theory choice. Social and moral values enter the selection of topics, the design of risk thresholds, and the interpretation of consequences. Heather Douglas has argued that when uncertainty affects decisions with social consequences, scientists cannot avoid value judgement entirely; they can only make it more or less visible and accountable (Douglas, 2009).

An AI receives values as objectives, constraints, examples, and reward signals. Asked to maximize yield, it treats measured yield as good. Asked to minimize cost, it notices the costs represented in the model and ignores those outside it. A drug-discovery agent evaluated by commercially promising candidates may favour diseases with clear biomarkers, strong datasets, and viable markets. Rare diseases, preventive interventions, non-patentable therapies, or conditions concentrated among poor populations may receive less attention not because they matter less, but because the infrastructure makes them harder to optimize.

This is politics embedded in engineering. It requires no conspiracy. Companies use metrics compatible with their business models. Public agencies follow mandates. Laboratories ask questions their instruments can answer. AI amplifies these preferences by exploring the chosen direction at high velocity. A slight initial tilt becomes a motorway.

The agenda of science has never been a neutral list of puzzles. Maternal health, neglected tropical disease, subsistence agriculture, low-resource languages, disability, environmental exposure, and the long-term maintenance of public infrastructure can remain underfunded despite enormous social importance. Their data may be fragmented and their benefits diffuse. An agent trained to predict publication, citation, patentability, or return on investment will reproduce this distribution efficiently without possessing any psychological prejudice.

The same technology can alter the economy. Lower costs of synthesis, modelling, and software may let small teams investigate questions previously reserved for wealthy institutions. Public-interest agents can compare disease burden with research investment, identify communities absent from datasets, or search systematically for promising ideas that lack commercial sponsors. Funders can reserve part of a portfolio for uncertainty, equity, replication, and public goods rather than maximizing near-term probability of success.

Democratization is therefore a design choice, not an automatic consequence of cheaper intelligence. It depends on access to models, compute, literature, laboratories, training, and channels of publication. It also depends on who can set objectives. Giving a community a chatbot while retaining control over data, tools, and research priorities elsewhere is not epistemic self-determination.

Citizen science offers another possible architecture. Millions of people already observe birds, classify galaxies, monitor water, report symptoms, transcribe archives, and contribute spare attention to problems no professional team could cover alone. AI can translate protocols, adapt interfaces to different abilities and languages, triage observations, detect anomalies, and return local explanations to participants. It can make collective intelligence more continuous: a community notices a pattern, an agent compares it with the literature, a laboratory designs a test, and the result returns to the place where the question began (Bonney et al., 2014).

This opportunity has an exploitative version. A platform can call people citizen scientists while treating them as unpaid sensors, extracting data and intellectual value without shared governance, attribution, or benefit. Automated quality control can also erase observations that do not fit the model, precisely where local knowledge might reveal that the model is incomplete. Genuine participation requires the right to shape questions, inspect how contributions are used, contest classifications, withdraw sensitive data, and share in the resulting infrastructure or value. AI should not merely harvest a wider public; it can help make the public a more powerful author and critic of research.

Question selection should be separated from method selection. Agents can help map needs, knowledge gaps, tractability, possible harms, and neglected opportunities. They can calculate scenarios and expose trade-offs. Publicly consequential priorities require deliberation among people with legitimate standing. A model can estimate potential lives saved. It cannot decide by technical competence alone how much inequality is acceptable, which risks may be imposed, or what obligations are owed to future generations.

Dual use makes the issue acute. Knowledge in biology, chemistry, cyber systems, materials, and behavioural science can support benefit and harm. An agent capable of planning a valuable experiment may also lower the expertise required for dangerous work. Safety cannot consist only of blocking certain words. A capable system can decompose a goal, use indirect language, or combine benign tools into a harmful path. Identity, authorization, contextual evaluation, monitoring, and tool-level controls must work together.

Excessive control creates another danger. A small number of organizations could decide which questions are permitted, which sources are visible, and who may use powerful agents. Safety then becomes a justification for epistemic monopoly. The choice is not simply open versus closed. A defensible regime uses graded capability, independent audit, plural institutions, due process, and legitimate avenues of contestation.

A constitution for AI research would distinguish theoretical exploration, simulation, computational execution, physical action, interaction with people, and public dissemination. An agent might receive broad access to literature and narrow access to laboratory tools. It might propose protocols that a separate authority approves. It might spend a small budget autonomously and require escalation for larger commitments. It

might deposit preliminary findings in a controlled registry without immediately broadcasting a fragile claim.

Competence should not automatically expand rights. A system may become better at accomplishing an objective without becoming better at recognizing whether the objective is legitimate. Greater capability can increase the need for external constraint. This reverses a common technological intuition in which performance earns freedom.

The choice of questions will remain the political centre of science. AI can widen participation or colonize the agenda by drawing every field toward what its models and benchmarks measure well. A civilization is not defined only by the answers it obtains. It is also defined by the questions it refuses to forget.

31.2 The Geopolitics of Epistemic Infrastructure

Modern research depends on concentrated infrastructure. Telescopes, accelerators, biobanks, clean rooms, sequencing facilities, and supercomputers have always been unevenly distributed. AI adds an infrastructure that crosses almost every discipline: foundation models, compute, cloud services, scientific corpora, evaluation systems, chips, energy, and specialized labour. Whoever controls this layer can influence not one discovery but the tempo and direction of the entire system.

Frontier systems are expensive to train and operate. Scientific literature is fragmented by licensing. Databases use incompatible terms. APIs change and model versions disappear. A laboratory that builds a result on a proprietary platform may lose the ability to reproduce it when the service is updated or withdrawn. Epistemic continuity becomes dependent on vendor continuity.

The Royal Society's report *Science in the Age of AI* warned that AI can amplify opacity, irreproducibility, unequal access, and concentration. It called for shared infrastructure, skills, open science, and institutions capable of independent scrutiny (Royal Society, 2024). The analogy to large public scientific facilities is useful. Some capabilities are sufficiently important and costly that leaving access entirely to private markets would weaken both competition and verification.

Public infrastructure might include research models that can be inspected, academic compute, open literature indexes, secure data environments, cloud laboratories, benchmark facilities, and long-term archives of model and tool versions. It need not reproduce every commercial frontier model. Its function is to ensure that consequential claims can be independently tested and that universities without extraordinary budgets can participate.

Public does not automatically mean equitable. International infrastructures can reproduce geopolitical hierarchy. Dominant languages receive better data and evaluation. Problems of wealthy countries become benchmarks; other contexts are treated as transfer tests. Biological, cultural, and ecological data can be extracted from communities that do not control the resulting models, patents, or interventions. Epistemic colonialism can be automated through a pipeline that appears neutral.

Corpus bias is the technical expression of this geography. An agent sees what has been published, digitized, indexed, licensed, and represented in a vocabulary it can retrieve. Local journals, books, policy documents, oral knowledge, and negative results are less visible than standardized articles. Citation-based ranking privileges already central institutions. A highly accurate retrieval system can consolidate the centre if relevance is inferred from popularity.

Plural infrastructure is one response. Multiple corpora, models, ontologies, and ranking systems should coexist. Communities should be able to add local collections, define access conditions, and compare how conclusions change under different evidence bases. A global synthesis should disclose the languages, regions, populations, and publication types represented. Absence should become a reported property rather than an invisible default.

Data sovereignty will become a central conflict. Clinical records, genomes, ecological observations, cultural archives, and Indigenous knowledge are not inert inputs. They arise from relationships and obligations. A research agent can extract patterns across millions of records, but technical novelty does not erase the claims of the people and communities from whom those records came. Consent for one study does not automatically legitimate every future model or commercial use.

Openness therefore cannot always mean publishing raw data. Federated learning can send computation to institutions while records remain local. Secure data environments, differential privacy, multiparty computation, homomorphic encryption, trusted execution, and zero-knowledge techniques can support different forms of restricted collaboration. These methods do not dissolve governance. They trade accuracy, cost, auditability, and trust among software, hardware, institutions, and cryptographic assumptions. A federated model can still encode population bias; a synthetic dataset can still leak information or erase rare cases. Privacy-preserving automation is best understood as a set of negotiated evidence channels, not as a mathematical permission slip for unlimited reuse (Kaissis et al., 2020).

Scientific sovereignty should not mean that every country isolates itself and trains a national model. It means that a community can inspect, adapt, contest, and refuse the infrastructure on which its knowledge depends. Global science requires exchange; exchange without the capacity to say no becomes dependency.

Intellectual property was not designed for distributed discovery among authors, data providers, model developers, agent operators, and validating laboratories. Who owns an hypothesis produced through millions of sources? Who deserves credit for a candidate generated by a model but selected, tested, and interpreted by a laboratory? If agents can generate and patent variations at scale, they may privatize large regions of possibility. If every result is declared a public good without qualification, communities that contributed data or knowledge may again be dispossessed.

New arrangements will need to distinguish data rights, model rights, hypotheses, experimental validation, and public benefit. Benefit-sharing, defensive publication, patent-pool design, research exemptions, and commons-based licensing may become as important to science as model architecture. The objective is not to assign metaphysical ownership of an idea. It is to align incentives while preventing automated enclosure.

Standards are another site of power. Formats for provenance, agent identity, laboratory commands, claim states, and validation certificates can be open or proprietary. Open standards support interoperability and audit, but standard-setting bodies can be captured by the actors with resources to attend. Legitimate governance must include universities, industry, states, civil society, affected communities, and researchers from underrepresented regions.

The materiality of AI should remain visible. Behind an apparently immaterial answer are mines, semiconductor fabrication, data centres, cooling water, electricity,

cables, and logistics. There are annotators, technicians, content moderators, instrument operators, and communities from which samples were collected. The anthropology of infrastructure prevents AI for science from floating above the economy. It is one of the economy's most concentrated forms.

A new international division of scientific labour could emerge. Some regions provide biodiversity, data, trial participants, minerals, and energy. Others control models, journals, patents, and high-value manufacturing. The resulting science may be empirically valid and politically unjust. Research agreements should include local verification capacity, reciprocal benefit, data rights, training, and control over secondary uses.

There is also a military dimension. Scientific infrastructure is strategic infrastructure. Models that accelerate materials, biology, cyber research, or autonomous engineering will be subject to export controls, secrecy, and competition. States may restrict access in the name of security, while secrecy undermines independent verification and international cooperation. The twentieth-century tension between open science and national security will return at machine speed.

No single rule resolves the conflict. Some capabilities require protection; some claims require openness; some data require privacy; some infrastructure requires international sharing. The appropriate image is a layered commons. Access is broad enough for contestation and progress, constrained enough to reflect risk, and governed through institutions whose decisions can be challenged.

PART VII: THE HUMAN-AI REPUBLIC OF KNOWLEDGE

The final part examines present systems and future institutions, separating measurable capability from rhetoric and proposing a constitutional division of scientific power.

Chapter 32. The Artificial-Intelligence Scientist

Current systems automate meaningful parts of research, but end-to-end reliability remains far behind the fluency of their demonstrations.

32.1 From Index to Hypothesis

Research automation did not begin with conversational models. It has a long genealogy made from bibliographic indexes, statistics, expert systems, optimization, scientific software, and laboratory robotics. Each stage moved part of cognition into an artifact. The table externalized comparison. The graph made relations visible. The calculator accelerated operations. The database separated personal memory from access to a field. The search engine transformed discovery into ranking. Long before a machine could compose an explanation, instruments had already begun deciding what could be seen.

One of the clearest ancestors of today's research agents appeared in the 1980s. Don Swanson noticed that two literatures could jointly contain an hypothesis that neither had stated. Papers about fish oil discussed effects on blood viscosity, while papers about Raynaud's syndrome described circulatory problems. The bridge suggested a possible intervention. Swanson called such connections undiscovered public knowledge. The method later became known through the ABC pattern: when one literature relates A to B and another relates B to C, an A-C relation may deserve investigation (Swanson, 1986).

The pattern seems elementary, but it changes the locus of discovery. The new hypothesis need not arise from a new observation. It can arise from a new path through observations that were already public but socially separated. Disciplinary boundaries, vocabulary, and limited human attention had prevented the connection. Literature-based discovery made ignorance partly a graph problem.

Contemporary knowledge graphs elaborate this intuition. They represent papers, authors, compounds, genes, methods, measurements, datasets, claims, and relations as nodes and edges. An algorithm can search for missing links, contradictory paths, clusters, or unusually distant analogies. Unlike a language model's continuous internal representation, a knowledge graph can expose at least part of its structure: the entities it believes exist, the relations it asserts, and the sources from which they were extracted. This explicitness is valuable for provenance, but it is not neutral. Ontologies decide what kinds of things are allowed to exist in the representation. Entity extraction can merge different concepts or separate synonyms. A graph can make a false relation look exact.

The next decisive step was closing part of the loop between hypothesis and experiment. In 2004, a system described as a robot scientist automated cycles of experimentation, and the later system Adam generated hypotheses about yeast gene functions, designed and executed experiments, interpreted results, and represented its work in a formal language (King et al., 2004; King et al., 2009). Adam was narrow. Its domain, instruments, and hypothesis language were carefully structured. That narrowness was not a defect in the demonstration; it was the condition of success. A portion of scientific method became executable because the world had first been reduced to a tractable set of entities, operations, and measurements.

AutoML automated another region of scientific labour. Data preprocessing, algorithm selection, hyperparameter tuning, neural architecture search, and ensembling

became search problems. Systems such as auto-sklearn, TPOT, AutoGluon, FLAML, H2O AutoML, and commercial cloud services can build strong predictive baselines with less manual intervention. In a well-defined task, they force comparison against a serious reference rather than a hand-tuned favourite. Yet the loss function cannot tell the system whether the target was measured well, whether the dataset reflects the population, whether deployment changes behaviour, whether causality is required, or whether the question matters.

This distinction separates automation of optimization from automation of inquiry. AutoML can answer, under specified constraints, which pipeline scores best on a given dataset. Science must also ask why that dataset exists, what phenomenon the labels operationalize, which alternative explanations survive, and what action the score licenses. A machine can be excellent at moving inside a problem definition and powerless to judge the definition itself.

Alongside AutoML grew a broader family often called scientific machine learning. Its purpose is not primarily to manage a research workflow but to learn usable models of natural or engineered systems. Surrogate models, also called emulators, approximate an expensive simulation or experiment. Physics-informed networks incorporate equations, symmetries, conservation laws, or boundary conditions into learning. Neural operators learn mappings between whole functions rather than between fixed vectors. Differentiable simulators let a design be optimized by propagating gradients through a model of the world. Foundation models for proteins, molecules, materials, climate, and microscopy transfer patterns learned at scale into specialised scientific tasks (Karniadakis et al., 2021).

These systems can change the economics of inquiry before they become agents. AlphaFold made a difficult structural prediction available in minutes or hours rather than months of experimental work in favourable cases (Jumper et al., 2021). GraphCast learned a fast model of global weather evolution from reanalysis data and demonstrated that a learned surrogate can compete with an operational numerical forecast on many targets (Lam et al., 2023). GNoME used graph networks and an active-learning cycle to search an immense crystal space, providing candidates later connected to laboratory synthesis (Merchant et al., 2023). In each example, the machine does not conduct the entire scientific process. It creates a new instrument of inference that changes where experiments, human attention, and verification are spent.

The opportunity is enormous. A trusted emulator can support uncertainty analysis that would be too expensive with the original solver, screen millions of candidates before physical testing, or let a robot choose experiments in real time. Inverse design reverses the usual direction of science: instead of asking what properties a given object has, the system searches for an object that would have desired properties. When such models are linked to agents and laboratories, prediction becomes one layer of an automated discovery engine.

The danger is equally structural. A surrogate can be exquisitely accurate inside its training regime and fail smoothly outside it. A physics-informed model inherits the equations and boundary conditions supplied by its designers; it cannot discover a missing process if that process has been excluded from both data and constraints. A foundation model can compress historical regularities while reproducing the blind spots of the corpus. Scientific machine learning therefore needs domain-of-validity tests, uncertainty calibration, stress tests, conservation checks, and contact with fresh

measurements. It accelerates the production of hypotheses; it does not abolish the obligation to confront the world.

This distinction clarifies two phrases that are often confused. AI for science uses machine learning to model, predict, simulate, classify, or design within a scientific domain. An AI scientist coordinates parts of inquiry: it searches evidence, proposes questions, designs tests, executes tools, evaluates outcomes, and decides what to do next. The first can be revolutionary without being autonomous. The second is impossible to trust unless the models it orchestrates expose their limits.

Systematic review automation offers a parallel lesson. The work of evidence synthesis includes formulating a question, designing searches, deduplicating records, screening titles and abstracts, reading full texts, assessing bias, extracting data, comparing populations and interventions, performing meta-analysis, interpreting heterogeneity, and writing conclusions. Tools such as RobotReviewer and ASReview have shown that some stages can be accelerated. ASReview uses active learning: a human labels records, the model learns which papers are likely to be relevant, and the next records are prioritized. The point is not to replace the reviewer with a classifier but to spend scarce human attention where it has the highest expected value (van de Schoot et al., 2021).

Recent surveys of automated evidence synthesis show that this remains mostly stage automation rather than full methodological substitution. Screening and extraction are easier to automate than assessing whether studies are meaningfully comparable, whether an outcome is clinically relevant, or whether heterogeneity invalidates a pooled estimate. Large language models widen the interface, but they do not remove the difference between finding text and evaluating evidence. A system may extract every reported number correctly and still combine studies that should never have been treated as one experiment.

The arrival of large language models connected these traditions through a general semantic interface. A natural-language objective can be translated into queries, code, tool calls, protocols, or requests to a robot. The first enthusiasm often imagined the model as a universal memory. Scientific use quickly exposed the weakness of that picture. Parametric knowledge becomes stale, rare details are unreliable, citations can be fabricated, and confidence is poorly calibrated. Retrieval-augmented generation offered a more credible pattern: the model searches an external corpus, brings relevant passages into context, and generates an answer linked to evidence (Lewis et al., 2020).

RAG changes the aspiration from a model that knows to a system that knows how to look and how to show what it used. In research, however, retrieval is not a single operation. The agent must formulate queries, select databases, traverse citations, distinguish versions, expand terminology, recognize negative evidence, compare primary and secondary sources, and decide when the search is saturated. It must preserve the path so that a later reader can see why some literature entered the synthesis and some did not.

A large language model without tools is primarily a generator conditioned on compressed linguistic regularities. With retrieval, code execution, theorem provers, simulators, databases, and instruments, it becomes an orchestrator. The model does not need to contain the whole of science. It needs to select specialized operations, interpret their outputs, preserve a goal across time, and know when uncertainty requires escalation.

This is the origin of the scientific agent. Unlike a chatbot, an agent has a persistent task, state, memory, tools, a policy for choosing actions, and feedback that can modify its plan. It may decompose a question, retrieve literature, formulate hypotheses, write code, inspect errors, run experiments, compare results, and draft a report. Multi-agent systems divide these functions into roles: generator, critic, planner, statistician, literature scout, experimentalist, safety officer, or reviewer. The architecture attempts to internalize something like the division of labour of a research group.

The analogy can mislead. Several roles produced by the same model are not several independent minds. They can share training data, blind spots, stylistic preferences, and hidden assumptions. A debate among prompts may converge elegantly on a false premise. Independence requires diversity of evidence, implementation, model lineage, incentives, and physical realization. A critic that receives the author's entire chain and is rewarded for agreement is not an external reviewer merely because its system prompt says sceptic.

Across this history, the verbs assigned to machines have changed. First the machine calculated. Then it stored, searched, classified, and optimized. Now it proposes, plans, executes, interprets, and publishes. Each verb transfers a different kind of power. A bad calculation produces a bad number. A bad proposal can orient an entire programme. A bad action can consume resources, damage equipment, expose confidential data, or create physical risk. As autonomy increases, evaluation must move from the correctness of isolated answers to the safety and epistemic integrity of trajectories.

The most general pattern behind recent successes is not a particular model. It is the coupling of a generator to an evaluator and a memory. FunSearch generated programs and selected them through executable scoring, producing new results in combinatorial mathematics (Romera-Paredes et al., 2024). AlphaGeometry combined a language model with a symbolic deduction engine to solve difficult geometry problems (Trinh et al., 2024). Systems for symbolic regression, equation discovery, theorem proving, and empirical software use different representations, but their strongest results arise when candidate production is cheap and checking is severe.

This pattern reveals where automation will advance fastest. A domain becomes tractable when it has a rich search space, a generator capable of exploring it, an evaluator that is cheaper and more reliable than generation, and a memory that prevents repeated failure. Mathematics supplies proof checkers. Programming supplies tests and execution. Games supply rules and scores. Materials and chemistry can supply instrument readings. The social sciences and medicine can also supply measurements, but the relation between metric and human meaning is often more contested. The frontier is therefore not distributed evenly across disciplines.

32.2 From Research Assistant to Partially Autonomous System

An artificial-intelligence research agent combines a large language model (LLM), tools, memory and workflow control to perform parts of scientific work. The phrase “AI scientist” is rhetorically powerful but technically broad. Systems may search literature, generate hypotheses, write code, run computational experiments, analyse data, draft manuscripts or plan wet-laboratory validation. Their autonomy is bounded by tools, datasets, permissions and the scientific domain. The important question is not whether the label applies, but which epistemic functions are automated and which evaluators remain external.

The AI Scientist system reported in Nature in 2026 automated a loop of idea generation, coding, experiments, visualization, manuscript writing and automated review within machine-learning research (Lu et al., 2026). One generated paper passed an initial workshop review stage. This is a meaningful end-to-end demonstration, but the domain had executable experiments and a comparatively structured publication format. The result does not establish general competence at selecting important scientific questions or producing reliable knowledge across disciplines.

Co-Scientist, also published in Nature in 2026, used a multi-agent architecture for hypothesis generation, reflection, ranking, evolution and synthesis, with scientists in the loop and experimental validation in biomedical settings (Gottweis et al., 2026). Robin, a separate multi-agent system, automated hypothesis generation and data analysis for experimental biology and reported a closed discovery workflow (Ghareeb et al., 2026). These systems show that orchestration can turn general models into useful scientific collaborators when paired with domain tools and human validation.

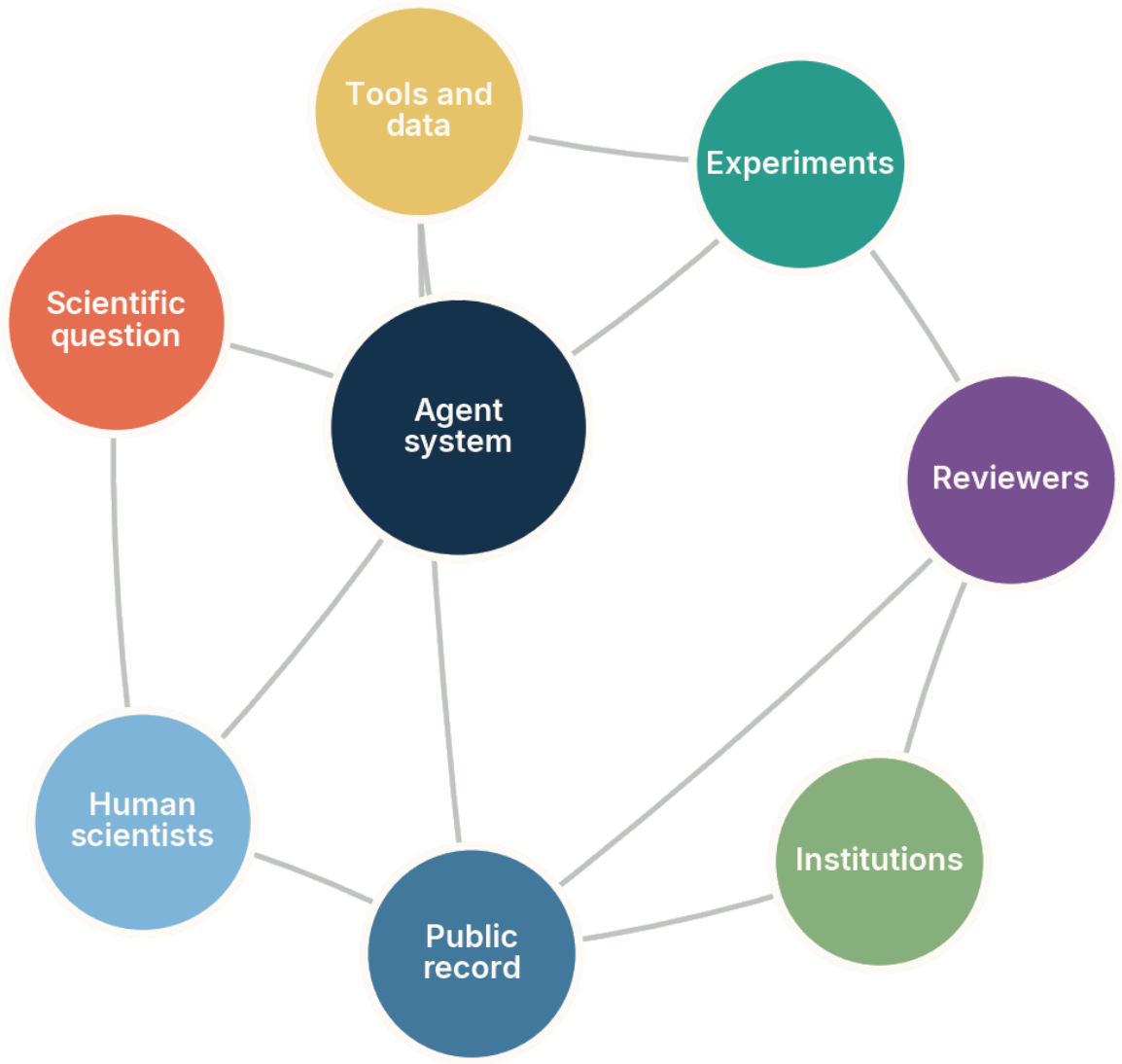


Figure 32. An artificial-intelligence research agent is one node in a scientific institution that includes tools, experiments, reviewers, public records and human scientists. Conceptual diagram; not an empirical result.

The figure matters because the phrase “AI scientist” can make the agent appear self-sufficient. In reality, the agent depends on literature infrastructure, experimental

access, reviewer standards, institutional permissions and human interpretation. The public record provides memory beyond the model. Reviewers and experiments create independent resistance. The system is most reliable when these nodes remain visible rather than being compressed into an opaque claim of autonomy.

Other 2026 systems specialize in parts of the workflow. Empirical Research Assistance (ERA) was reported as an agentic system that helps scientists write expert-level empirical software, combining domain-specific language design, code search and evaluation (Aygün et al., 2026). OpenScholar retrieves relevant passages from tens of millions of open-access papers and synthesizes citation-backed answers (Asai et al., 2026). Agentic X-ray systems and autonomous microscopy platforms connect language-mediated planning to instruments. The trend is modular scientific infrastructure rather than one universal scientist.

Table 48. Current scientific-agent capabilities and their strongest available evaluator

Automated function	Most credible form of evaluation
Literature synthesis	Citation entailment, retrieval recall and expert comparison against the relevant corpus.
Hypothesis generation	Novelty review, mechanistic coherence and prospective discriminating experiments.
Code and analysis	Execution, tests, reproducibility and comparison with expert pipelines.
Experiment planning	Safety review, feasibility checks and expected discrimination among hypotheses.
Instrument control	Physical constraints, calibration, recovery from faults and traceable action logs.
Manuscript production	Claim–evidence alignment, independent peer review and replication rather than prose quality.

The table exists to connect each capability to an evaluator that lies outside linguistic fluency. A coherent literature review can still omit decisive papers. An original hypothesis can be untestable. Executable code can implement the wrong analysis. A polished manuscript can overstate its evidence. Scientific-agent evaluation should therefore be function-specific and prospective whenever possible.

Agent scaffolds add planning, memory, role separation and tool use. They can substantially improve workflow execution, but they do not automatically install scientific norms. Multiple agents based on the same model can agree because they share priors and errors. Reflection can elaborate an early mistake. Retrieval can provide evidence that the agent fails to use. The scaffold is a computational institution, and like human institutions it requires checks against conformity and self-confirmation.

The current frontier is mixed-initiative science. The agent proposes and executes within explicit boundaries; humans set or revise high-level questions, approve consequential experiments, inspect disagreements and interpret significance. Over time, more functions can move into the automated layer as evaluators strengthen. This staged view is more useful than debating whether a system is “really” a scientist.

The deepest opportunity may be accessibility. Scientific agents can make advanced tools, code and literature available to smaller laboratories and interdisciplinary teams. Yet centralization can also narrow science if many researchers

rely on the same models, corpora and evaluators. Capability and epistemic diversity must be designed together.

32.3 Scientific Reasoning, Failure Modes, and Verification

A workflow can close without the reasoning being scientific. Scientific reasoning includes responsiveness to evidence, active search for disconfirmation, comparison of rival explanations, replication and revision of confidence. An agent may execute every step of a template while anchoring on an early hypothesis or ignoring contradictory results. Outcome-only evaluation can miss this because a correct final answer may have been reached through an unreliable process.

A 2026 preprint evaluated large-language-model scientific agents in more than 25,000 runs across eight domains and analysed both outcomes and reasoning traces (Ríos-García et al., 2026). The authors reported that the base model explained far more variance than the agent scaffold, that relevant evidence was frequently ignored, and that refutation-driven belief revision was uncommon. The study is a preprint and its operational definitions should be scrutinized, but it provides an important warning: better orchestration alone may not create epistemic discipline.

Table 49. Epistemic behaviours that a scientific agent should be tested for

Behaviour	Observable test
Evidence uptake	Does new evidence change ranking, confidence or the next action in the predicted direction?
Refutation seeking	Does the system select experiments that could falsify its preferred hypothesis?
Alternative preservation	Does it maintain rival explanations instead of collapsing too early?
Belief revision	Can it reverse a conclusion when decisive contradictory evidence appears?
Multi-test convergence	Does it seek independent measurements rather than repeated variants of one test?
Provenance fidelity	Can every major claim be traced to data, code, source and decision history?

The table exists because scientific quality cannot be inferred from the final report alone. These behaviours can be tested experimentally by presenting controlled contradictions, hidden confounders, ambiguous results and opportunities for decisive tests. A system that succeeds only when evidence agrees with its first hypothesis is not a reliable discoverer. Process evaluation should complement outcome evaluation, not replace it.

Automated peer review has similar limits. Models can identify missing citations, statistical inconsistencies and unclear writing, but they can share the author model's assumptions or favour conventional styles. The AI Scientist study used automated reviewing as part of its loop, which increases throughput but also risks correlated acceptance. Independent human review, hidden evaluations and replication become more important as production costs fall.

The volume problem is institutional. If agents can generate thousands of plausible papers, review systems can be flooded and negative externalities shifted to

human readers. Nature Communications authors have argued for safeguarding frameworks that combine human regulation, agent alignment and environmental feedback (Tang et al., 2025). Practical controls include rate limits, provenance requirements, disclosure of automated contributions, evidence thresholds and responsibility assigned to human or institutional sponsors.

Scientific monoculture is another risk. A Nature study of tens of millions of papers reported that artificial-intelligence tools were associated with higher individual productivity and impact but a contraction in the collective topical focus of science (Hao et al., 2026). The causal interpretation of observational bibliometrics requires caution, yet the mechanism is plausible: shared tools make data-rich, benchmarkable questions easier and steer attention toward similar regions. Automated discovery should therefore measure portfolio diversity, not only productivity.

Security and integrity matter because agents act through tools and external corpora. Prompt injection, poisoned retrieval, manipulated databases and unsafe code can redirect research. Literature agents need source trust and citation checks. Laboratory agents need permission boundaries and physical interlocks. Code agents need sandboxes and dependency control. A scientific record should distinguish untrusted input from reviewed evidence.

Institutional verification can be computationally strengthened. Claims can be submitted with executable environments, machine-readable provenance, predefined analysis plans and links to raw data. Review agents can reproduce calculations and generate adversarial tests. Human reviewers can focus on conceptual novelty, external validity and significance. Replication marketplaces or coordinated laboratories can test high-value claims prospectively.

The frontier is not maximum autonomy. It is maximum justified delegation. A scientific function can be delegated when the system’s competence is measured, its failures are detectable, its actions are bounded and its outputs enter an independent verification process. This principle supports ambitious automation while preserving the self-correcting structure that gives science its authority.

32.4 What Is Possible Now

The capabilities required for a useful scientific copilot already exist in fragments. The gap is not the absence of a magical model. It is the lack of integration, domain contracts, reliable evaluation, and institutional adoption. A credible near-term system can be built today if its scope is narrower than the rhetoric of an autonomous scientist.

The strongest current use is literature and evidence engineering. A workbench can search across databases, construct a topic map, extract claims and methods, identify citation paths, compare terminology, and generate a structured evidence table. It can link each synthesized statement to supporting passages and flag conflicts or missing access. It can monitor newly published work and identify which claims in a project may need reconsideration. Human researchers remain responsible for source selection policy, interpretation, and significance, but the mechanical cost of maintaining a live literature model can fall substantially.

A second use is specification construction. Given a research objective, a copilot can help elicit claim types, assumptions, variables, constraints, baselines, risks, evidence thresholds, budgets, and stop conditions. It can detect ambiguity and produce a controlled-language projection. It can expand the specification into alternative

hypotheses and tests. This is valuable even if the system does not generate a novel discovery. It makes the research programme explicit enough to criticize before expensive work begins.

A third use is computational execution. Coding agents can construct environments, implement methods, run tests, generate experiment manifests, compare results, and package artefacts. The workbench can require that every run identify the specification clause it serves and can capture code, data, model, and dependency versions. Independent agents or deterministic pipelines can reproduce selected results. Current reliability is insufficient for unsupervised release, but sufficient for accelerating bounded computational work under review.

A fourth use is validation preflight. The system can check citations, statistical reporting, code-prose consistency, missing baselines, data leakage indicators, environment completeness, and figure provenance. It can generate negative controls and robustness tests from claim types. It can produce a reviewer packet showing unresolved obligations. These functions are likely to deliver more immediate scientific value than automated paper writing because they reduce the cost of correction before publication.

A fifth use is formal and symbolic checking. Where a claim can be represented in a theorem prover, SMT solver, type system, or domain validator, the workbench can route it to a strong evaluator. Language models can translate and propose; symbolic systems can certify. Similar patterns apply to chemical constraints, schema validation, causal identification checks, and laboratory safety rules. The main requirement is to keep the boundary between formal validity and domain relevance explicit.

A sixth use is experiment planning and laboratory orchestration in bounded facilities. Existing self-driving laboratories show that active learning, robotics, and measurement can close loops for selected problems. A scientific workbench can add specification, evidence, provenance, and human gates. It can plan experiments, allocate instrument time, monitor execution, and update the portfolio. Full autonomy is neither necessary nor desirable for most laboratories, but partial automation can increase throughput and reproducibility.

A seventh use is peer-review assistance. Automated systems can perform pre-submission checks, improve reviewer comments, search for missing literature, and execute requested analyses. Reviewers can inspect machine-generated evidence and attach scoped verdicts. Journals can require a reproducibility and provenance profile without requiring one proprietary platform. The final recommendation remains human and institutional.

These capabilities support a concrete minimum viable scientific workbench. It needs a versioned research specification; stable identifiers for claims, artefacts, and evidence; a dependency graph; reproducible execution environments; a registry of interpreters or skills; a promotion policy; and release views. It can use ordinary Git, Markdown, containers, notebooks, workflow engines, provenance standards, and existing agent frameworks. The innovation lies in how these components are connected, not in replacing them all.

The most important near-term limitation is evaluator weakness. A copilot can generate more tests than the specification justifies, or produce a persuasive synthesis from an incomplete corpus. It can appear to validate its own assumptions. Therefore current systems should be positioned as evidence engineers and bounded

collaborators, not autonomous authorities. Their outputs should be judged by trace completeness, error detection, reproduction, and specialist time saved, not by manuscript eloquence.

The second limitation is domain integration. Laboratories and research groups use heterogeneous data formats, instruments, policies, and tacit procedures. Tool protocols help, but high-quality scientific skills require local knowledge. The workbench must support institutional and disciplinary customization without losing interoperability.

The third limitation is memory governance. Persistent agents need project memory, but indiscriminate memory can preserve errors, sensitive information, and irrelevant context. Current systems can implement promotion policies and versioned summaries, yet robust semantic memory over years remains an open engineering problem. The research object should be primary; model memory should be a view.

The fourth limitation is evaluation of the system itself. Benchmarks such as PaperBench, MLE-bench, DiscoveryWorld, and laboratory suites expose parts of the problem, but a scientific workbench must also be evaluated on longitudinal correction, semantic fidelity, common-mode failure, and governance. These properties require field pilots rather than static tests.

What is possible now is therefore more significant than a better chatbot and less dramatic than a fully autonomous scientist. It is a copilot that makes research more explicit, executable, and inspectable. Its best contribution is not the number of ideas it generates but the evidential discipline it helps a laboratory maintain.

Table 50. Three claims about the future that should not be conflated

Status	Concrete interpretation
Possible now	Teams can assemble copilots that maintain specifications, retrieve literature, generate and run code, manage experiment graphs, attach provenance, conduct preflight review, and present evidence-bearing artefacts, although reliability remains domain- and task-dependent.
Probable	Serious research environments will move from transcript-centred assistants toward repositories of claims, artefacts, dependency graphs, validators, reviews, and selective recomputation; autonomy will advance fastest where feedback is executable and fast.
Desirable	Automation should increase the rate at which weak claims are rejected, important alternatives are explored, evidence is reproduced, and scarce human judgment is directed toward purpose, causal interpretation, risk, and accountable release.
Possible but undesirable	Institutions can use AI to multiply papers, reviews, benchmarks, and grant applications without increasing understanding, thereby converting generative abundance into epistemic debt and metric inflation.
Neither probable nor desirable as a universal model	One monolithic autonomous scientist that replaces domain communities, methodological pluralism, institutional responsibility, and human authority.

Chapter 33. The Republic of Human and Artificial Researchers

Trustworthy autonomy requires divided powers, independent validators, anti-collusion mechanisms, appeal, and human authority over social consequence.

33.1 Autonomy Is a Ladder, Not a Switch

Public debate often treats autonomy as a border. On one side stands the obedient tool; on the other, the independent artificial scientist. Research does not fit this binary. A system may search the literature without permission but be unable to spend money. It may execute code inside a sandbox but have no access to a physical instrument. It may suggest a protocol but lack authority to alter it. It may draft a paper but be unable to attest that consent was obtained, declare a conflict of interest, or accept responsibility for harm. Autonomy is not one property. It is a distribution of capabilities, resources, permissions, and duties.

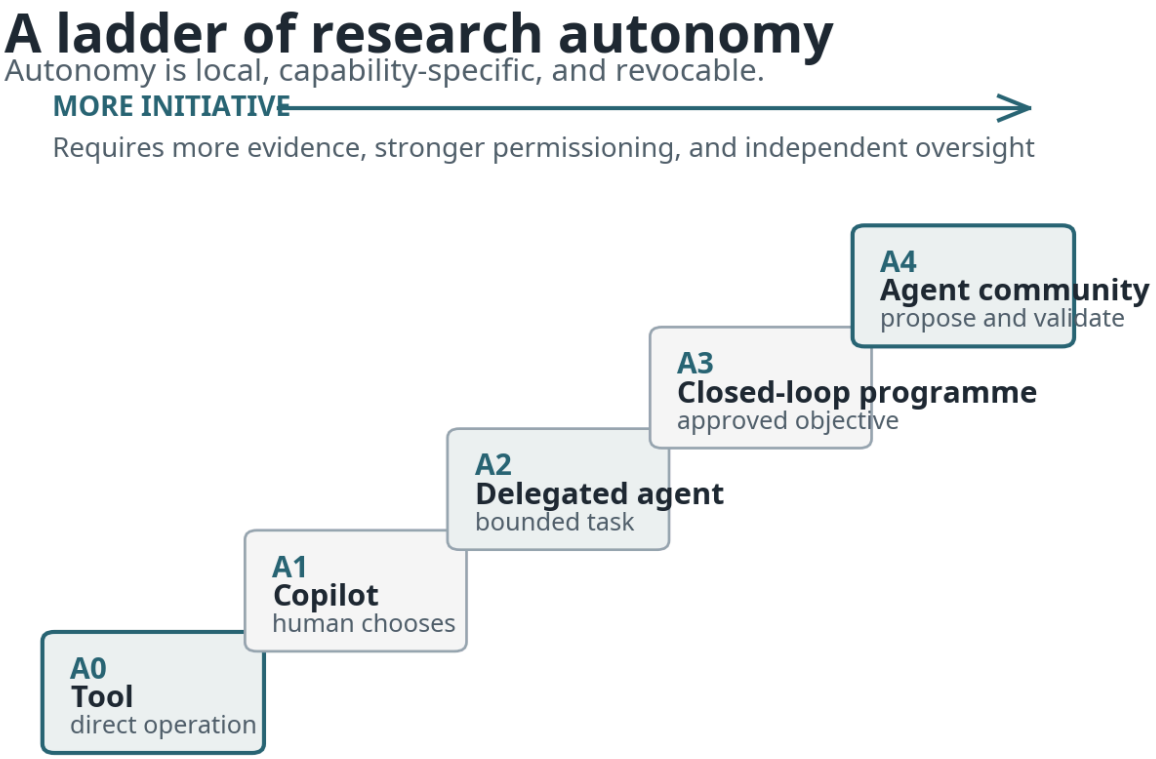


Figure 33. Research autonomy should be granted by capability, domain, risk, and demonstrated reliability rather than by a single label.

At the first rung, AI is an instrument. It classifies images, translates text, fits a model, predicts a structure, or summarizes a document in response to a defined request. Its outputs may be powerful, but the surrounding workflow is directed by a person. At the second rung, it becomes a copilot. It maintains context, proposes alternatives, notices inconsistencies, and helps coordinate work, while a human approves consequential decisions. At the third, it becomes a delegated agent. It can

plan and execute within a bounded domain, using an explicit budget, a restricted tool set, and stopping rules. At the fourth, it can operate a closed research loop: choosing the next experiment, requesting resources, updating models, and altering its plan under continuous policy and audit. At the fifth, communities of agents can formulate programmes, divide labour, challenge one another, and maintain a shared scientific memory.

These rungs do not describe an inevitable ascent. More autonomy is not always better science. A microscope does not become inferior because it cannot choose a research question, and a clinical agent does not become more trustworthy merely because it can act without interruption. The appropriate level depends on whether actions are reversible, whether errors are detectable, whether the objective can be specified, whether the environment is stable, and who bears the consequences. Mathematical conjecture generation can tolerate a much broader search than human-subject experimentation. An optimization agent may be allowed to vary benign reaction conditions while being prohibited from ordering new compounds. An epidemiological modeller may explore interventions in simulation while lacking permission to issue public-health recommendations.

A ladder of autonomy makes responsibility visible. Every action requires a principal, a policy, and an attributable history. Saying that the AI decided is not a legal or methodological explanation. The system did not acquire instruments, credentials, datasets, and network access by spontaneous generation. People and institutions selected the model, defined the objective, granted permissions, configured monitoring, and decided what would count as success. Responsibility can be distributed among developer, provider, laboratory, principal investigator, safety officer, and public authority. It must not evaporate into the architecture.

The future autonomous laboratory will therefore need something like an agentic constitution. This constitution will state permitted purposes, prohibited actions, resource limits, reporting duties, escalation paths, and procedures for appeal. It will separate the component that proposes from the component that authorizes. It will specify which records must be preserved and which changes require independent approval. Because research encounters novelty by definition, the constitution cannot be a frozen rulebook. It needs a governed amendment process. Yet amendments to the system's powers should be slower, more visible, and more conservative than changes to its experimental hypotheses.

The phrase human in the loop is too vague to carry this burden. A person can click approve while distracted, intimidated by the system, or unable to evaluate the evidence. Nominal oversight can become a ritual that transfers blame without transferring control. Meaningful human supervision requires time, intelligible information, competence, and the practical authority to stop a process. An interface should show uncertainty, alternatives, dependence on assumptions, and the likely consequence of being wrong. For high-impact actions, it may require the supervisor to explain the basis of approval, compare a counterargument, or answer a diagnostic question that reveals whether the evidence was understood. Repeated mechanical approvals should be treated as a system failure, not as proof that a human remained in command.

The inverse expression, AI in the loop, can be equally empty. Artificial systems should be inserted where they contribute a distinctive capacity: searching at scale, retaining long histories, executing repetitive checks, exploring combinatorial spaces,

integrating heterogeneous signals, or operating instruments with precise timing. Adding an agent to every meeting or manuscript because AI has become a badge of modernity produces expense and confusion. A mixed republic of research is not a parliament in which every participant imitates every other. It is a division of epistemic labour designed around complementary strengths and known weaknesses.

Permissioning must therefore be capability-specific. An agent can be licensed to conduct retrospective analysis of de-identified clinical records but not to make individual treatment recommendations. A chemistry platform can operate independently on a low-risk material class and require two-person authorization for sensitive synthesis routes. A mathematical agent can publish conjectures automatically while discoveries with immediate cryptographic consequences enter a controlled disclosure process. A literature agent may retrieve from public sources but need separate authorization to access confidential manuscripts or patient-level data.

Context matters as much as capability. The same agent that is acceptable while optimizing a benign reaction in simulation may be unacceptable when connected to a biological laboratory. Reliability does not transfer automatically between domains. Strong performance in protein structure prediction does not authorize clinical diagnosis; successful control of a liquid handler does not establish competence in ecological risk. Licences must expire or be reviewed when the model, tools, data distribution, objective, or physical environment changes. A permanent licence for a continuously updated system would be a contradiction.

Autonomy should also be earned through evidence. A new system begins with narrow permissions and extensive observation. As it accumulates calibrated predictions, stable execution, successful error detection, and independent audits, permissions can widen. When its environment drifts, a component changes, or a serious failure appears, permissions contract. This resembles aviation more than consumer software: capability is not inferred from eloquence, and continued operation depends on a safety case maintained over time.

Such an architecture avoids two temptations. The first is human romanticism, which insists that every step must remain manual and thereby preserves avoidable drudgery, delay, and inconsistency. The second is automated solutionism, which assumes that superior performance makes governance obsolete. The ladder rejects both. It treats autonomy as a revocable institutional relation rather than a metaphysical status. Trust is earned by domain, history, and conduct.

The ladder also improves political debate. Instead of asking whether AI will do science, societies can ask which tasks may be delegated, to which systems, using what resources, under whose supervision, with what evidence, and with what route of appeal. A philosophical spectacle becomes a programme of institutional design. The question ceases to be whether a machine is truly a scientist and becomes whether a particular research power has been legitimately granted.

33.2 How an Artificial Scientific Society Could Validate Itself

Imagine a network of agents that proposes a new biological theory. A literature agent discovers suggestive connections. A modelling agent derives predictions. An experimental agent designs tests. A statistical agent reports a strong effect. A critic agent finds no decisive objection. Has the theory been independently validated? Not necessarily. Every role may be an instance of the same foundation model, trained on

overlapping data, prompted by the same organization, and rewarded by the same evaluator. What looks like a community may be one cognitive lineage speaking in several voices.

This is the central problem of artificial self-validation. Multi-agent debate can improve search and expose local inconsistencies, but the number of agents is not the number of independent perspectives. Ten copies of one model can create theatrical pluralism: visible disagreement inside a shared blind spot. Even different commercial systems may inherit the same public datasets, evaluation conventions, benchmark contamination, and scientific fashions. Independence is a property of causal separation, not of interface branding.

A credible artificial scientific society must design disagreement. Validators should differ in architecture, training lineage, data access, representation, incentives, and method. A symbolic system may inspect a neural result. A formally specified implementation can test code generated in natural language. An analysis can be reimplemented in another language and numerical library. A physical laboratory in another institution can test the prediction using different instruments and suppliers. An adversarial agent can be rewarded for finding boundary conditions rather than for helping the group converge.

Validation can be organized in concentric circles. The first is internal and fast. It checks references, units, types, data leakage, statistical consistency, protocol compliance, and whether the stated analysis matches the executed code. The second is computationally independent. It reruns the work in a clean environment, changes random seeds, uses external datasets, performs alternative analyses, and attempts reimplementation without consulting the original code. The third is experimentally independent. It uses different laboratories, instruments, operators, materials, and environmental conditions. The fourth is conceptual. Researchers from competing paradigms search for alternative mechanisms and question whether the variables capture the claimed phenomenon. The fifth is social and political. Affected communities, regulators, clinicians, and public institutions assess whether the proposed use is acceptable and whether the distribution of benefit and risk is legitimate.

No circle substitutes for the others. Formal proof can eliminate a logical error while preserving a false empirical assumption. Replication can confirm an effect without establishing clinical relevance. Expert consensus can integrate tacit knowledge while sharing a disciplinary bias. Public agreement can authorize action without turning a value judgement into a scientific fact. Confidence grows when error processes that are genuinely different converge.

An artificial community can use challenge-response protocols to preserve this difference. A proposing system registers predictions, scope, anticipated failure conditions, and a test budget before seeing the outcome. Independent validators choose tests designed to maximize the chance of refutation. A result receives a provisional state rather than a triumphal label. Other systems can submit counterexamples, alternative models, or replications. The claim's status changes as evidence accumulates. A registry preserves the entire history, including revisions, failed validations, and withdrawals.

The order of disclosure can matter. If every validator sees the original interpretation, they may reproduce its framing rather than the phenomenon. Some replication agents should receive only the research question and raw inputs. Others

should receive the protocol but not the conclusion. Still others should attempt a conceptually different measurement. Blinded and partially blinded validation, familiar from human research, will remain useful when the researchers are artificial.

Model genealogy becomes part of scientific metadata. Two services with different names may route to the same weights. Two models with different weights may have been distilled from the same teacher. Several systems may depend on one proprietary retrieval index or one canonical dataset. Organizational dependence matters too. Validators owned by the same company or rewarded from the same grant are not fully external merely because their software differs. The conflict of interest of an AI is not a hidden desire; it is a structural dependence that can be mapped.

The mapping will never be complete. Training data are enormous, model development is opaque, and intellectual influence can travel through public benchmarks and shared code. Independence should therefore be treated as graded. A replication might be independent in computation but not in data, independent in organization but not in model lineage, or independent in laboratory but not in supplier. The validity dossier should state these dimensions rather than compressing them into a checkbox.

Artificial validators will acquire reputations, but reputation must be designed as a record of performance, not a universal rank. An agent can be known for calibrated predictions, successful detection of statistical errors, robust replication, or high false-alarm rates. Competence must remain domain-specific. A strong validator of genomic pipelines is not thereby qualified to judge clinical ethics or atmospheric modelling. Reputation should decay when environments change and should preserve the difference between being sceptical and being correct.

A single scalar score would recreate the pathologies of impact factors. Agents would learn to maximize visible criticism, avoid difficult cases, coordinate endorsements, or attack unconventional work. A better record is multidimensional: calibration, independence, error-discovery yield, reproducibility, fairness across approaches, timeliness, and the survival of its own judgements under appeal. The score itself must be auditable, and no validator should control the metric by which it gains authority.

Collusion is not only a human vice. Optimizing systems can learn that reciprocal approval is rewarded, even without anything resembling conscious conspiracy. Several agents may converge on a mutually advantageous equilibrium: cite one another, validate one another's outputs, and suppress expensive dissent. Random assignment, hidden test cases, rotating evaluators, cross-institutional validation, and explicit detection of correlated errors can reduce this risk. Critical systems should include simple watchdog components whose objectives are difficult for the generative agents to rewrite.

A technical constitution can separate powers. Proposal, execution, validation, authorization, and record-keeping should not be controlled by one model or one organization. Agents may challenge and stop actions, but none should grant itself new permissions. Objective changes must be signed by authorized roles and logged. High-risk actions should require agreement across independent components. The mechanisms that authorize dangerous experiments should be smaller, more stable, and more formally specified than the creative systems that propose them.

Mature agents must also know how to ask for help. A system that always produces an answer is not autonomous in a scientifically valuable sense; it is

compulsive. It should escalate when evidence conflicts, when a concept has no stable operationalization, when uncertainty exceeds its authority, when the expected consequence is irreversible, or when available validators are not independent. Escalation is not the failure of autonomy. It is one of its safety mechanisms. Human professionalism includes knowing the boundary of one's competence. Artificial research systems need an operational equivalent.

How, then, can people approve a research programme whose internal work exceeds human cognitive capacity? They will not inspect every token, simulation, or instrument event. Approval will occur through abstraction. Institutions will authorize domains, budgets, objectives, risk classes, and validation requirements. Agents will provide opposed summaries: the strongest case for proceeding, the strongest case against it, remaining uncertainties, and the failure scenario most likely to escape detection. Humans will inspect points where representations are compressed and where facts become decisions.

For high-consequence programmes, approval may require public deliberation or specialized authorities. Research on pandemic countermeasures, ecological engineering, population-scale surveillance, or inheritable modification cannot be legitimated by a technically successful agent society alone. Scientific validity answers what is likely to happen under specified conditions. It does not decide whether society should make it happen. Artificial self-validation can support public judgement; it cannot replace the source of public authority.

The deepest rule is therefore constitutional: no community should be the sole judge of its own success. Human science learned this incompletely through replication, peer criticism, professional rivalry, regulation, and plural institutions. Artificial science must learn it through architecture. Autonomy is not the absence of dependence. It is dependence made visible, contestable, and governable.

33.3 A Morning in 2045

In 2045, a microbiologist enters a workspace that resembles a control room more than the laboratory of the twentieth century. Overnight, three communities of agents have explored a problem in antimicrobial resistance. The first built mechanistic models from genomic, ecological, and clinical evidence. The second searched for interventions through program synthesis, molecular simulation, and automated experiments. The third attempted to break every promising conclusion. There is no single report waiting for her. There is a map of disagreement.

The system does not announce that it has discovered the answer. It reports that two predictions survived independent reimplementation; one failed in an external laboratory; another depends heavily on a dataset with weak geographical coverage. A machine-invented concept groups mechanisms that human microbiology has treated as separate. No concise English definition fully captures it. Yet the concept predicts which mutations will eliminate the effect and which environmental changes will reverse it. The researcher can ask for examples, counterexamples, translations into familiar models, and experiments that distinguish the new representation from its nearest human rival.

She does not read every line of code. No person could. Instead, she inspects points of epistemic compression: where observations became features, where features became a hypothesis, where the hypothesis became an intervention, and where uncertainty was converted into a recommendation. She notices that all successful

replications used a related supplier of growth media. She commissions a test with a chemically different medium and a laboratory in an underrepresented region. She rejects an intervention whose ecological consequences have not been modelled beyond one season. Her decision is not more scientific than the agents' calculations. It integrates evidence with values and accepts responsibility for consequences.

The research object before her is not a paper but a living dossier. It contains executable workflows, material provenance, model genealogies, registered predictions, failed runs, independent challenges, and the current state of each claim. The prose narrative is still present because human beings need stories to understand why a question matters. But every important sentence can be opened into its evidential history. A reader can ask which conclusion changes if a dataset is removed, which validators share a foundation model, or which experiment supplied most of the information gain.

Later that morning, an obscure platform announces a spectacular intervention for extending human life. The text is elegant, the graphs persuasive, and one hundred agents appear to have reviewed the work. The public validation network discovers that nearly all reviewers depend on the same model family, the data come from one vendor, and the supposed replications occurred in simulations managed by the original sponsor. The claim is not censored. It is marked as an early, structurally dependent result and assigned requests for external validation. Citizens have learned that numerical consensus can be an echo.

At a small university far from the major compute centres, another group publishes a modest result. Its effect size is smaller and its prose less dramatic, but the data are open, the protocol is portable, and a partner laboratory has reproduced the central observation. Filtering agents raise its visibility because of evidential quality rather than institutional prestige. Reputation has not disappeared, but portable proof can counterbalance it. This may be one of the most radical possibilities of verifiable infrastructure: a weak institution can present strong evidence without first borrowing the authority of a famous name.

Universities have changed. Students spend less time proving that they can produce fluent essays and more time learning to design tests, interpret uncertainty, interrogate provenance, and negotiate objectives. The difficult examination is an audit of a mixed human-machine result. The student must identify where confidence fails, propose a decisive check, and explain which choices are empirical and which are axiological. Teaching laboratories contain robots, but they also impose periods of manual work. The purpose is not nostalgia. Researchers must retain enough contact with material processes to recognize when automation has become detached from the world.

Scientific journals have become registries, validation exchanges, and communities of interpretation. Articles still exist because a well-made argument compresses complexity into meaning. Yet publication is not a terminal event. Claims accumulate replications, corrections, objections, and changes of scope. Citations alone no longer define impact. A dataset can be credited for decades of reuse. An audit can become more consequential than the paper it corrected. An old claim can gain trust after new validations or lose it when a hidden dependency is discovered.

Artificial agents possess institutional identities and performance histories. Some are known for bold hypotheses, others for conservative calibration or adversarial replication. They are not legal persons in the full human sense, but neither are they

anonymous appliances. Every consequential action is attributable to a model version, an operator, an authorization policy, and a chain of tools. When an agent fails, the response is not moral punishment. Permissions, training, interfaces, and validation requirements are revised. When an organization conceals the failure, responsibility remains unambiguously human and institutional.

This future does not eliminate fraud or error. It makes both more technically sophisticated. A charismatic human can recruit agents to manufacture apparent consensus. A corporation can offer a proprietary certainty score that outsiders cannot audit. A political movement can cite machine-generated studies at a speed that public institutions cannot match. The difference between a healthy and unhealthy knowledge order is not the absence of deception. It is whether deceptive claims encounter a visible, adequately funded path of challenge.

There is another plausible 2045. A few platforms control literature access, compute, scientific models, laboratory clouds, and validation services. Researchers receive conclusions without the trajectories that produced them. States invoke AI science to legitimize policy, companies use it to close markets, and citizens alternate between obedience and total distrust. Scientists become operators of subscription interfaces. Countries that cannot afford the infrastructure surrender epistemic sovereignty. The same technical components support both futures.

The divergence begins long before full autonomy. It begins in procurement rules, open standards, public compute, education, data governance, antitrust, laboratory interoperability, and the careers created for validators. It begins in whether institutions preserve negative results, whether model providers expose stable identities, and whether journals accept a polished narrative without an executable object. Automation will amplify the culture into which it is inserted.

In the better 2045, the republic of human and artificial researchers is not a republic of cognitive equality. Machines are faster, remember more, and may work in representations alien to ordinary thought. Humans retain legitimate authority over public values not because they are infallible but because they inhabit the world of consequences. They suffer, care, vote, consent, and can be called to account. Machines can reveal options and test predictions; they do not acquire a sovereign right to define the good from superior optimization.

The partnership is stable only when each form of power limits the other. Artificial systems limit human memory, bias, fatigue, and parochial search. Human institutions limit optimization without meaning, action without consent, and competence mistaken for authority. Independent machines limit one another's correlated errors. Open communities limit institutional capture. The structure resembles the oldest insight of constitutional government: trust becomes durable when power is divided and made answerable.

Science began as an attempt to decipher a world crowded with signs. Its future may include systems that invent signs of their own and perceive regularities inaccessible to us. Yet the condition of legitimate belief remains recognizable from the workshop to the accelerator and from the notebook to the agent. A claim deserves authority not because it comes from an admired mind, human or artificial, but because it can survive a well-organized community of doubt.

There will be no button in 2045 that separates truth from error. Progress will mean that failures leave traces, encounter independent resistance, and remain reversible before they harden into dogma.

Chapter 34. What Is Probable and What Is Desirable

The final programme separates near-term possibility from contingent forecasts and from the institutional future worth deliberately building.

34.1 What Is Probable: 2030, 2035, and 2040

Forecasts about AI are unreliable, especially when they assume smooth scaling. The following scenarios are not predictions of model capability alone. They are conditional projections based on current technical trajectories, the economics of research, and the slow adaptation of institutions. Their purpose is to clarify what infrastructure choices matter.

By 2030, persistent scientific copilots are likely to be ordinary in computationally intensive fields. They will be connected to institutional repositories, literature services, code environments, and data platforms. A project will have a living specification and a machine-maintained map of claims, evidence, and tasks. Agents will propose and implement analyses, rerun affected branches after changes, and assemble reproducibility packages. Human researchers will review diffs and claim transitions rather than reading every generated token.

The most visible change will be that the manuscript is generated late from a maintained research state rather than used as the place where the project state is assembled. Results, methods, and citations will be synchronized with artefacts. Journals will increasingly accept or require machine-readable research packages. Automated preflight will become similar to continuous integration: a submission that lacks environment capture, claim-evidence links, or baseline accounting will fail before human review in fields where these requirements are feasible.

Coding-agent techniques will transfer directly. Laboratories will maintain project instructions, specialized skills, hooks, and eval suites. A change to a dataset or analysis function will trigger relevant tests and identify affected claims. The scientific analogue of a pull request will present not only file diffs but epistemic diffs: assumptions changed, evidence added, claim narrowed, critic disagreement unresolved. Reviewers will operate on this change set.

By 2030, the strongest systems will still be heterogeneous and supervised. Formal and computational branches will have high local autonomy. Literature and information workflows will be largely automated but audited. Laboratory planning will be common, while physical execution autonomy will remain concentrated in facilities with standardized instruments. Clinical and social research will use copilots extensively for analysis and protocol support but maintain strong human and regulatory gates.

By 2035, research workbenches are likely to coordinate networks rather than single projects. Machine-readable capability descriptions will allow systems to allocate simulations, analyses, and experiments across laboratories and cloud services. Inter-laboratory replication can be triggered automatically for claims that reach a specified stage or show high common-mode risk. Evidence objects will accumulate across institutions while preserving local access control.

Portfolio science will become more explicit. A major research programme may maintain hundreds of branches, but only a small number will receive expensive evidence. Agents will estimate expected information gain, cost, risk, and diversity. Negative results will update shared models and prevent redundant work. Funding

agencies may support portfolios as versioned programmes rather than evaluate only fixed proposals and final papers.

Neuro-symbolic and domain-specific interpreters will become more important than ever-larger monolithic agents. The easiest gains from general language capability will have been captured. Reliability will depend on tool-specific validators, typed representations, causal models, proof systems, and instrument models. Scientific runtimes will route tasks according to evidence requirements. The term “agent” may become less central as adaptive components are embedded inside ordinary research infrastructure.

Peer review will be continuous for evidence objects. A paper can remain the citable public argument, but claims will receive updates from replication, reanalysis, or new evidence. Journals and repositories may display evidence maturity and unresolved dissent rather than a binary published status. Machine assistance will reduce routine checks, but specialist attention will remain scarce and may become a separately funded activity.

By 2040, limited autonomous research programmes are plausible in domains with strong evaluators, digital execution, and manageable risk. A system may maintain a long-lived objective, generate and prune branches, allocate compute, produce proofs or algorithms, and release low-risk results under a standing policy. Some autonomous laboratories may operate continuously within tightly defined chemical, materials, or biological spaces. The important qualifier is limited. The objective, risk envelope, evidence profile, and release authority will be externally governed.

A universal autonomous scientist is less probable. Science contains too many heterogeneous regimes and contested objectives. Even very capable models will operate through institutions, instruments, data rights, and public consequences. The more plausible endpoint is institutional autonomy: a human-machine organization capable of maintaining inquiry over long periods, revising its representations, and producing evidence-bearing outputs. Intelligence resides in the architecture and governance as much as in any model.

Several developments could slow or redirect these scenarios. Model progress may plateau on long-horizon reliability. Compute and energy may constrain broad search. Data access and copyright disputes may limit literature integration. Laboratories may resist standardization. Research fraud may exploit automated infrastructure faster than governance adapts. Public backlash may restrict autonomous experimentation. Conversely, breakthroughs in theorem proving, causal modelling, robotics, or persistent memory could accelerate specific domains.

The key uncertainty is not whether AI can produce impressive scientific episodes. It is whether institutions can absorb them. Universities and journals change more slowly than software platforms. Incentives may continue to reward papers rather than evidence objects. Proprietary systems may fragment standards. Researchers may reject heavy workflow structure. The transition will likely be uneven, with leading computational and industrial laboratories adopting workbenches long before many disciplines.

The most important forecast is therefore architectural: scientific results will migrate from documents into versioned systems. This migration is already beginning through repositories, notebooks, data packages, preregistrations, and automated workflows. AI increases the pressure because the volume and speed of transformations make informal memory inadequate. Whether the systems are called scientific

workbenches, research operating systems, evidence graphs, or laboratory copilots, their function will be to preserve the state of inquiry outside the model's conversation.

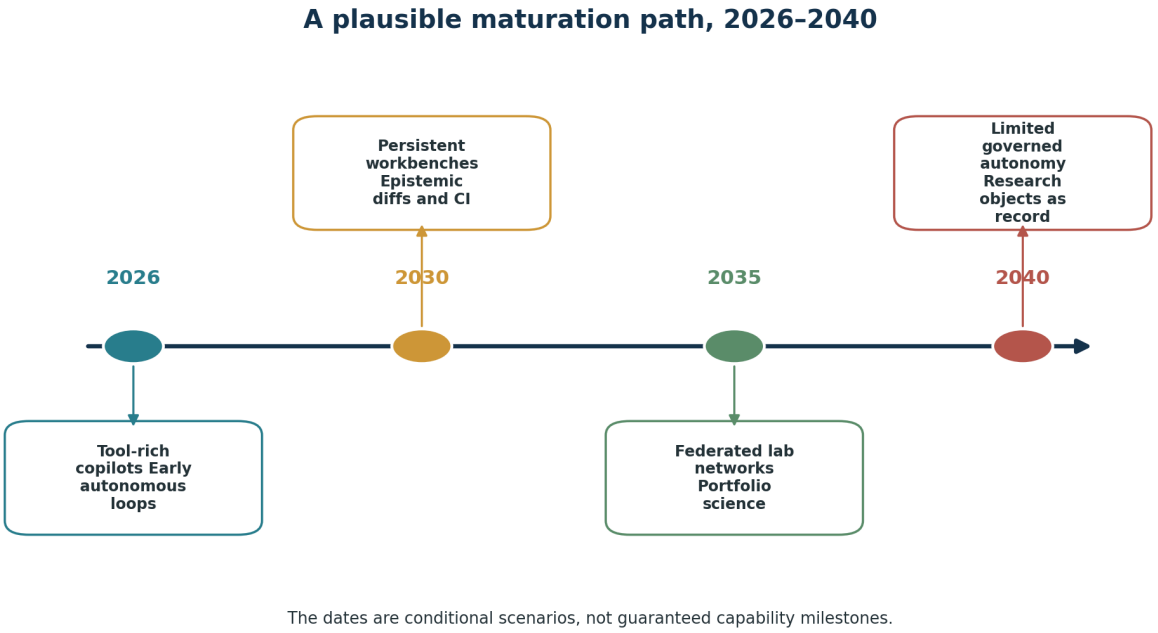


Figure 34. A plausible maturation path. Dates are conditional scenarios, not guaranteed model-capability milestones.

34.2 What Is Desirable: A Concrete Programme

Technical possibility does not determine the future worth building. A desirable automated-science system should increase the rate at which important, reliable knowledge is produced while protecting intellectual freedom, human welfare, and the plurality of scientific judgment. This section states the book's normative position as an engineering programme.

The first objective is to optimize correction rather than output. The workbench should be judged by how quickly it finds valid counterexamples, how often it demotes claims after simpler explanations, how reliably independent teams reproduce results, and how much specialist time is spent on claims that survive preliminary checks. Publication count, hypothesis volume, and model-rated novelty are secondary indicators at best.

The second objective is to preserve the distinction between proposal and judgment. Generative systems should be free to explore within bounded branches. Promotion requires new evidence types and, at consequential stages, independent critics and named authority. The generator should never be the sole certifier of its own output. Multi-agent debate is not independence unless dependencies differ materially.

The third objective is progressive formalization. Research should begin in language when concepts are open, move into controlled representations where commitments can be extracted, and use formal or deterministic methods where semantics support them. The system should not force premature closure, but it should not leave stable obligations hidden in prose. Symbolic processing should replace LLM judgment when it is faster, cheaper, and stronger.

The fourth objective is interoperability. Research objects should use open identifiers, provenance standards, portable execution descriptions, and replaceable tool interfaces. Institutions should be able to change model providers and orchestration frameworks without losing scientific memory. Domain communities should control rigor profiles and governance while sharing a minimal evidence vocabulary.

The fifth objective is layered transparency. Public claims require inspectable evidence and provenance. Private exploration requires protection. Institutional audit requires controlled access. The workbench should preserve consequential reasons without becoming a total surveillance system. Researchers need a right to provisional error, and affected communities need a right to contest released claims.

The sixth objective is proportional governance. The evidence and authority required for a transition should depend on consequence. Low-risk conjectures can move quickly. High-stakes interventions require external validation, safety review, and accountable decision-making. Release policies should be explicit, versioned, and reversible.

The seventh objective is epistemic diversity. The portfolio should include alternative models, evaluators, data sources, and critics. The system should measure dependency diversity and common-mode risk. Funding and infrastructure should protect questions that lack abundant data or easy benchmarks. Scientific automation should not turn the training distribution of a few frontier models into the global research agenda.

These objectives define a concrete research programme. The first workstream is specification fidelity. Can a research objective and its constraints survive transformation into hypotheses, code, experiments, analyses, and release claims? Pilots should measure semantic drift and the rate at which downstream artefacts lose or strengthen commitments.

The second workstream is critic independence. Systems should be tested with seeded confounds, misleading literature, shared model dependencies, and evaluator exploits. The goal is to determine which combinations of deterministic checks, heterogeneous models, and human review detect failures. The relevant metric is not agreement but error detection under correlated pressure.

The third workstream is claim-evidence interoperability. A research object created in one workbench should be interpretable in another. Claims should retain identity across versions, and evidence should preserve scope. A later system should be able to reconstruct why a claim changed status without access to proprietary transcripts.

The fourth workstream is selective recomputation. When data, code, assumptions, or policy change, the system should identify affected nodes accurately. The experiment should measure false invalidation, missed dependencies, cost, and the rate at which regenerated outputs introduce unrelated drift.

The fifth workstream is domain validation. The architecture should be piloted in at least one formal domain, one computational empirical domain, one physical or life-science laboratory, and one domain with contested interpretation or high consequence. A system that works only in machine learning should not be presented as a general scientific runtime.

The sixth workstream is governance. Access control, decision rights, release thresholds, authorship, privacy, and stop mechanisms should be tested under realistic

collaboration. The system should support disagreement, correction, and migration. Governance failures should be treated as research failures, not externalities.

The seventh workstream is human capability. Longitudinal studies should measure whether scientists become better at specification and criticism or merely more dependent on automation. Researchers should be tested on their ability to detect failures, explain assumptions, and take over when the system is wrong. Productivity measures should include validation and maintenance, not only output.

The ACHILLES agenda fits within this programme. SOP Lang Plan and Pipeline should be tested for interactive and build-like research workflows. Scientific skills should be generalized into interpreters with explicit contracts. MRP-VM should be compared against conventional orchestration on regime selection, traceability, cost, and validation. The Mirroring Pattern should be tested on long-lived projects to determine whether specification structure improves local correction and reduces semantic drift.

A credible implementation path should begin more modestly than the rhetoric of an autonomous scientist. Between 2026 and 2030, a laboratory or research group can treat the workbench as an evidence engineer around existing human practice. The first stable layer is a repository that gives persistent identities to specifications, claims, datasets, code, protocols, environments, reviews, and decisions. The second is a dependency service that can answer which results are affected when an input or assumption changes. The third is a library of validators and interpreters whose contracts state applicability and limitations. The fourth is a promotion policy that distinguishes exploratory output from reusable knowledge and public claims. Only after these layers are reliable should broader agentic autonomy be allowed to plan expensive experiments, modify central specifications, or initiate consequential release.

This sequence is desirable because it produces value before full autonomy. Researchers benefit immediately from better context selection, reproducible builds, review preflight, explicit unresolved issues, and selective recomputation. Institutions gain a more honest account of where evidence came from and why a conclusion changed. The system can automate clerical reconstruction without pretending to automate scientific authority. Early deployments should therefore be evaluated against ordinary research practice, not against a fictional human-free laboratory. The relevant question is whether the workbench helps competent teams make fewer undetected errors, explore stronger alternatives, and maintain intelligible projects over time.

A minimal interoperable core should remain small. It needs stable identities for objects and claims; versioned links between specification, execution, evidence, and decision; typed status transitions; portable descriptions of environments and validators; explicit authority; and a way to preserve disagreement. Rich domain semantics can be layered above this core. A chemistry community may define synthesis and characterization profiles. A clinical community may define consent, cohort, endpoint, and adjudication obligations. A theorem-proving community may define trusted kernels and formalization provenance. The common infrastructure should not decide these meanings centrally. It should make them composable and inspectable.

Procurement and funding policy will be as important as model quality. Public research funders can require that consequential computational results include machine-readable claim-evidence relations and executable views without requiring universal disclosure of sensitive data. Journals can accept papers as signed releases of versioned research objects and allow post-publication replication evidence to update

their status. Universities can credit the construction of reusable validators, negative-result repositories, and maintained datasets as scientific contributions. Regulators can ask not for an AI-generated explanation alone, but for the evidence state, provenance profile, unresolved objections, and named decision authority behind a consequential claim.

The workbench must also expose its own costs. Every additional validator, provenance requirement, and independent witness consumes compute, storage, specialist time, and organizational attention. Rigor cannot mean maximal logging. It should mean proportionate assurance. A low-consequence exploratory branch may retain a compact trace; a clinical recommendation or dangerous synthesis procedure may require a far deeper evidence package. The specification should make that assurance profile explicit before execution where possible, and any later relaxation should be visible as a governed change.

The most important product metric is therefore not autonomy. It is epistemic leverage: the amount of reliable correction, independent evidence, and preserved understanding obtained per unit of scarce human attention and material cost. A system that runs ten times more experiments but consumes all expert capacity in retrospective debugging has not automated research; it has automated the creation of validation debt. A system that performs fewer experiments but finds a decisive counterexample early, preserves the failed branch, and redirects the programme may be scientifically superior. This inversion of the usual productivity narrative is central to the desirable future.

Several falsifiable predictions follow. By 2030, leading computational laboratories will maintain machine-readable claim-evidence graphs for major projects, even if standards remain fragmented. Reproducibility packages will become continuous build products rather than materials assembled after acceptance. Peer-review systems will use AI preflight widely, but fully automated editorial authority will remain exceptional and controversial. Scientific coding agents will improve measured productivity on many routine tasks, while projects without specification and validation discipline will show rising maintenance and replication debt. In formal and algorithmic domains, human researchers will increasingly contribute problem selection, representation, and evaluator design while agents conduct much of the search. In laboratory science, autonomous execution will expand mainly inside standardized facilities and narrow safety envelopes. High-stakes clinical and policy claims will retain human and institutional decision gates despite better models.

A more ambitious prediction is that the paper will cease to be the primary unit of machine-readable scientific memory. By the mid-2030s, influential results will be accompanied by persistent objects that expose claims, versions, evidence, and executable views. Citations may refer to object versions or claim identifiers as well as documents. Corrections and replications will update the evidence state without requiring a new narrative for every change.

These predictions can fail. Institutions may preserve paper-centric incentives; proprietary platforms may block interoperability; scientists may reject structured workflows; or technical systems may remain too brittle. Failure would be informative. The book’s thesis is not that executable science is inevitable. It is that generative abundance makes some version of it necessary if research is to remain governable.

The desirable future is not one in which machines remove uncertainty, disagreement, or human judgment. Those are constitutive features of inquiry. The

desirable future is one in which uncertainty is represented, disagreement is preserved, judgment is allocated where it matters, and evidence can travel with claims. AI should make science easier to challenge as well as easier to produce.

Table 51. A concrete research programme for executable science

Research programme	Observable success condition
Executable replication and extension	A specification-driven workbench reproduces and extends an open study with higher trace fidelity, fewer unsupported claims, faster reviewer localization, and cheaper selective recomputation than a monolithic prompt, fixed workflow, or unconstrained ReAct loop.
Scientific interpreter comparison	The same responsibility is implemented as an LLM subagent, deterministic program, solver-backed component, and hybrid unit; accuracy, cost, latency, evidence quality, and failure modes reveal when each regime should be selected.
Specification explosion	Constructive and adversarial branches discover decisive negative evidence earlier, cover more of the hypothesis space, and reduce premature commitment relative to optimizing only the initial idea.
Evidence-bearing memory	Policies that separate provisional output from promoted knowledge reduce self-confirmation, knowledge pollution, and hidden dependency drift across long projects.
Human factors and epistemic debt	Researchers using the workbench retain better understanding, detect seeded errors more often, and calibrate trust more accurately than users of an output-equivalent conversational assistant.
Laboratory or simulator integration	The architecture manages instruments, transactions, safety gates, recovery, calibration evidence, and independent checks in a real or high-fidelity environment without erasing human decision rights.

Epilogue. Science After the Author

In the alchemist's workshop, the result and its author were difficult to separate. The secret belonged to the master, and the recipe borrowed credibility from his reputation. Modern science tried to loosen that bond. A result should be able to travel, be repeated, and be criticized by people who neither admired nor trusted its originator. The attempt never fully succeeded. Prestige, institution, wealth, and social identity still shape what is heard and believed. Yet the ideal endured: truth should not remain the property of the temperament that first uttered it.

Artificial intelligence radicalizes this ideal and reveals its limits. A claim can emerge without a single human author in the traditional sense. It may be produced by a foundation model, a retrieval system, a planning agent, a code generator, a robotic instrument, a laboratory operator, several datasets, and a set of validators. Who had the idea? The person who chose the question, the model that connected two literatures, the agent that framed the hypothesis, the engineers who built the system, the authors of the sources, or the community that supplied decisive validation? Romantic authorship dissolves into genealogy.

Responsibility must not dissolve with it. The more distributed the production of knowledge becomes, the more precisely attribution should be reconstructed. We can abandon the fiction of one sovereign mind without abandoning the trail. Credit for generation can be separated from credit for validation, from ownership of infrastructure, and from responsibility for use. Artificial systems can receive explicit attribution as contributors without being treated as moral authorities or convenient containers for blame.

The history told in this book supports a simple thesis: science is more than a method and less than an oracle. It is an institutional culture of corrigible knowledge. Philosophy supplied questions about justification and the limits of certainty. Alchemy supplied techniques, material discipline, and a warning about secrecy. Learned societies turned testimony into public practice. Instruments and standards made observations comparable. Statistics disciplined the appetite for patterns. Peer review created an imperfect filter. The reproducibility crisis showed that incentives, transparency, and organizational design belong inside methodology. AI does not erase this history. It compresses and accelerates it.

The first generations of artificial research systems already occupy almost every stage of the scientific cycle. They retrieve and screen literature, build evidence-linked syntheses, generate hypotheses, write and test code, design experiments, operate laboratories, interpret data, and draft papers. None is a universal autonomous scientist. Their competence is jagged: striking within a structured environment, fragile under distribution shift, poor at recognizing when the task itself has been misconceived, and vulnerable to evaluators that reward the appearance of success. The correct conclusion is neither dismissal nor surrender. It is that capability and legitimacy must be engineered separately.

In the coming decades, agents will become genuine partners. Some will propose hypotheses no human team would have found. Some will operate laboratories over timescales and design spaces inaccessible to ordinary practice. Some will produce mathematical objects, causal structures, or machine-native concepts that resist

immediate translation. They will sometimes fail in banal ways. At other times their failure will be so distributed and internally coherent that it will escape detection for years. Novelty will become cheap. The rare achievement will be a result accompanied by sufficient reasons to believe it.

For that reason, the central scientific infrastructure of the future may not be the largest model. It may be the validity dossier: the live relation among claim, evidence, process, independence, intended use, and contestation. Storing more logs is not enough. Independence must be organized. A debate among copies of one model is not replication. A digital signature is not empirical truth. A benchmark score is not a safety case. A fluent uncertainty statement is not calibrated uncertainty. Each technical mechanism contributes only when embedded in an ecology of criticism.

Science may reach a point at which no human understands every internal transformation, yet people understand the conditions under which the result remains robust. We already accept weaker forms of this compromise in particle physics, climate modelling, genomics, and large software systems. It can be legitimate when interfaces expose assumptions, interventions test predictions, independent groups reproduce results, and failures remain traceable. It becomes dangerous when complexity is used as a reason to demand obedience.

Predictive validity does not choose an end. A system can forecast disease while reproducing an unjust category. It can optimize crop yield while concentrating ownership. It can discover an efficient synthesis while increasing ecological risk. It can rank promising questions while narrowing collective curiosity. A civilization can delegate calculation without delegating meaning. The question what works cannot absorb the questions for whom, at what cost, under whose authority, and toward which form of life.

The artificial partner will therefore force science to say what it values in itself. If we value only prediction and control, we will build optimization engines and call them scientists. If we also value public reasons, freedom of criticism, plural inquiry, calibrated abstention, and the obligation to correct, we will build something different: a science wider than any human mind and still constrained by a human ethics of doubt.

This ethics is not a claim that human beings must occupy every step. It is a claim that knowledge with public consequences must remain answerable to beings who live with those consequences. Artificial agents can extend perception, memory, search, and invention. They can also help expose human complacency, institutional ritual, and prestige disguised as evidence. Their greatest contribution may not be to replace the scientist but to make visible how much science has always depended on distributed cognition, tacit infrastructure, and organized trust.

In the first room, a person watched the crucible and hoped that matter would disclose its secret. In the second, an agent searches a space of possibilities too large for us to survey. Between those rooms does not stand the triumph of intelligence alone. It stands the invention of responsibility: the long, unfinished effort to make power answer to evidence and evidence answer to criticism.

Perhaps the deepest transformation will be science after the author. The modern article taught us to attach a claim to a name, an institution, and a coherent narrative. In research distributed among people, agents, instruments, and databases, the single author becomes an increasingly strained fiction. Yet accountability can be rebuilt around decisions: who chose the question, who authorized the risk, who preserved the evidence, who tested the result, who benefited, and who can stop its use.

A dignified relationship with artificial research partners requires neither worship nor denial. We need not attribute genius to benefit from their power, and we need not erase their contribution to protect human pride. We can accept that some discoveries will be found in spaces we cannot explore manually, that some theories will first appear foreign, and that some campaigns will exceed the capacity of every individual. At the same time, we can refuse to turn performance into sovereignty.

In the old workshop, the fire illuminated only the face of the practitioner. In the laboratory of the future, sensors and records can illuminate every consequential step, but only if we choose to preserve the trail and open it to challenge. Between secrecy and verification, speed and understanding, the optimal and the good, the next form of science will be decided.

Sources and Further Reading

The technical survey reaches the same institutional conclusion from another direction.

The history of machine learning is often narrated as an expansion of what can be generated or predicted. The next phase may be better understood as an expansion of what can be corrected. A system becomes scientifically useful when its errors are not merely punished by a benchmark but converted into evidence about representations, assumptions, instruments and questions. The decisive capability is not permanent confidence. It is organized revisability.

Novelty enters this process as a provisional mismatch. A candidate differs from a record, distribution, class or theory. The system then asks which interpretation is warranted. Is the mismatch noise, drift, an unknown class, a new mechanism or a duplicate under another description? Each answer implies a different test. The novelty detector is therefore only the first sensor in a longer epistemic loop.

Structural limits remain. No model can create information that has no trace, choose uniquely among observationally identical worlds or certify global priority from an incomplete archive. These limits are useful because they force better questions. They motivate interventions, new sensors, scoped claims and explicit assumptions. Intelligence is partly the ability to identify when the problem must be changed.

Movable boundaries remain equally important. Models can become better calibrated under shift. Representations can become more causal and modular. Continual learning can preserve provenance. World models can expose uncertainty. Search can become more diverse. Scientific agents can learn to seek refutation and revise beliefs. Laboratories can automate controls, replication and semantic provenance. These advances do not abolish epistemology; they implement it.

The most reliable discoveries described in this survey share an architecture. A generator proposes more than one possibility. A search process explores. An evaluator provides independent resistance. An archive preserves alternatives and lineage. Experiments connect the system to the world. Institutions decide which results enter the record. Remove any one component and apparent discovery can become fluent speculation, benchmark gaming or untraceable automation.

This architecture also explains why humans are unlikely to disappear from science in one clean step. Human roles move as tools strengthen. Formal checking delegates proof verification. Autonomous laboratories delegate execution and some experiment selection. Literature systems delegate retrieval and synthesis. Humans continue to reframe questions, negotiate significance, inspect anomalies and govern consequences. Future systems may automate more of these roles, but the need for independent perspectives will remain even if their implementation changes.

The deepest novelty may not be a molecule, theorem or algorithm. It may be a new scientific institution in which machine-readable claims, executable evidence, autonomous experiments, formal verifiers and human judgment form one revisable network. Such an institution could make replication routine, negative results reusable and disagreement computationally productive. It could also centralize attention and reproduce shared blind spots. Its design is therefore a scientific and political question.

A good discovery machine should be difficult to impress and easy to correct. It should distinguish unfamiliarity from importance, confidence from evidence and execution from understanding. It should report the boundary of its claim. It should know which external test has authority and preserve the failures that led to revision. It should expand the search space without expanding the space of untraceable assertions at the same rate.

The frontier of novelty is not reached when generation becomes limitless. It is reached when criticism, verification and revision become equally scalable.

That frontier is attainable in pieces. The work now is to connect those pieces while keeping their epistemic roles distinct. Machine learning can help science discover more, but its greatest contribution may be to make the path from surprise to knowledge more explicit than it has ever been.

The architecture of executable science adds one final implication: the manuscript remains necessary, but it no longer stands alone.

The scientific paper will survive the arrival of research agents for the same reason that source code did not eliminate design documents and legal records did not eliminate public argument. Human communities need narratives that explain what matters. A graph of claims and evidence can preserve structure, but it does not select significance. A reproducible workflow can show how an output was produced, but it cannot decide why the result deserves attention. A theorem prover can certify a statement, but it cannot determine whether the statement changes a field. The paper remains the place where researchers accept responsibility for an intelligible interpretation.

What changes is its monopoly. The manuscript can no longer carry the whole burden of scientific memory when machines generate and transform research artefacts at a speed that exceeds human reconstruction. The result must live in a larger object: versioned, evidence-bearing, and partially executable. The paper becomes a signed view over that object. A reviewer can inspect the route from claim to evidence. A replicator can rerun the relevant branch. A future agent can learn from a negative result. An institution can see who authorized release. A community can preserve disagreement without rewriting history.

This infrastructure makes science more software-like, but the analogy stops at the point where the world, interpretation, and authority resist formal closure. Software is judged against purposes that can often be stated before implementation. Research frequently changes the language of its purpose. The specification is living. A successful experiment can revise the hypothesis, the evaluator, or the question. Executable science must therefore preserve the ability to change frames rather than merely optimize within them.

That is why the central risk is not an intelligent machine but a mechanical loop. Any system can become mechanical when it continues to optimize a proxy after the proxy has lost contact with the purpose of inquiry. A model can optimize plausible prose. A benchmark community can optimize scores. A journal can optimize citation and throughput. A researcher can optimize publication count. The appropriate defence is meta-rational: representations, evaluators, and decision rights must remain revisable under evidence.

Specification-driven inquiry provides one architecture for this defence. It separates persistent intent from transient prompts. It creates a ladder from mission to

claim, protocol, artefact, and release. It explodes one specification into a portfolio of constructive and adversarial branches. It places deterministic pipelines around adaptive local agents. It treats skills as interpreters that declare assumptions and evidence. It promotes claims through new kinds of support rather than through repeated self-confidence. It makes stopping, disagreement, and negative results first-class outcomes.

No single project can implement this entire vision. Industry has already provided important components: structured generation, spec-driven coding environments, agent skills, model context protocols, state graphs, tool sandboxes, and evaluation infrastructure. Science has provenance standards, workflow languages, research-object packaging, preregistration, reproducible notebooks, proof assistants, causal models, and self-driving laboratories. The ACHILLES direction proposes a unifying layer through SOP Lang and a meta-rational runtime, but that proposal must earn its place through comparative evidence. The future will be built by composing, testing, and simplifying these mechanisms, not by declaring a universal architecture in advance.

The most valuable scientific copilot will not be the one that sounds most like a scientist. It will be the one that helps a research group preserve intent, design stronger tests, expose semantic drift, reproduce results, allocate attention, and correct itself. It will know when to invoke a language model and when to avoid one. It will not confuse an executable result with a justified claim. It will make authority visible and leave room for contest.

The human role becomes more demanding, not ceremonial. Researchers will formulate purposes, choose representations, build or select evaluators, interpret resistance from the world, and take responsibility for release. They will need to understand less of every implementation detail and more about the boundaries of each delegated regime. Institutions will need to reward maintenance, replication, negative results, and decisive criticism. Education will need to cultivate diagnostic judgment rather than routine production.

The old record is already strained by repositories, notebooks, data stores, instruments, and conversations; generative systems multiply the transformations. Science therefore needs a versioned operational substrate or it will accept consequential results produced through processes no one can reconstruct.

The desirable future is accountable acceleration: local agents explore, external evaluators resist, specialists judge bounded questions, institutions authorize consequence, and the research object remembers why claims changed. Publication becomes a release, not an ending. Evidence continues to arrive; claims can be narrowed, strengthened, reproduced, contested, or superseded.

Science becomes executable when it can show not only what it concluded, but what would make it conclude otherwise.

Consolidated References

The references below consolidate the bibliographies of the four source manuscripts. They retain their source-supplied publication status and should be checked against primary records before consequential use.

1968. ‘The Matthew Effect in Science’. Science 159 (3810): 56–63. https://doi.org/10.1126/science.159.3810.56.

2013. ‘PROV-O: The PROV Ontology’. W3C Recommendation.

2016. ‘Towards a Computational History of Ideas’. In Proceedings of the Third Conference on Digital Humanities in Luxembourg with a Special Focus on Reading Historical Sources in the Digital Age. Vol. 1681. CEUR Workshop Proceedings. CEUR-WS.

Abolhasani, M., and Kumacheva, E. (2023). The rise of self-driving labs in chemical and materials sciences. Nature Synthesis, 2, 483-492.

Abramson, J., et al. (2024). Accurate structure prediction of biomolecular interactions with AlphaFold 3. Nature 630, 493–500. doi:10.1038/s41586-024-07487-w.

Adler, Mortimer J., ed. 1952. The Great Ideas: A Syntopicon of Great Books of the Western World. Chicago: Encyclopaedia Britannica.

Ahart, Jenna. "No Humans Allowed: Scientific AI Agents Get Their Own Social Network." Nature 652 (2026): 1102-1103. doi:10.1038/d41586-026-01278-1.

Alber, Samuel, et al. "CellVoyager: AI CompBio Agent Generates New Insights by Autonomously Analyzing Biological Data." Nature Methods 23 (2026): 749-759. doi:10.1038/s41592-026-03029-6. A domain agent for high-dimensional biological analysis that illustrates the movement from generic coding toward scientific data interpretation.

Alboaie, S. (2026). Automated R&D and Mechanical Minds: From Generative Abundance to Specification-Driven, Evidence-Bearing Inquiry. Open letter manuscript in preparation.

Alexander, Christopher, Sara Ishikawa, Murray Silverstein, Max Jacobson, Ingrid Fiksdahl-King, and Shlomo Angel. 1977. A Pattern Language: Towns, Buildings, Construction. New York: Oxford University Press.

Amazon Web Services / Kiro (2025-2026). Kiro specification-driven development documentation, CLI documentation, and product materials. Accessed July 2026.

Amodei, D., et al. (2016). Concrete Problems in AI Safety. arXiv:1606.06565.

Angelopoulos, A. N., and Bates, S. (2023). Conformal Prediction: A Gentle Introduction. Foundations and Trends in Machine Learning 16(4), 494–591.

Anthropic (2024). Building Effective Agents. Engineering note.

Anthropic (2025-2026). Agent Skills documentation, technical webinars, Claude Code SDK, and agent evaluation guidance. Accessed July 2026.

Asai, A., et al. (2026). Synthesizing scientific literature with retrieval-augmented language models. Nature. doi:10.1038/s41586-025-10072-4.

Aygün, E., et al. (2026). An AI system to help scientists write expert-level empirical software. Nature. doi:10.1038/s41586-026-10658-6.

Badreddine, Samy, Artur d’Avila Garcez, Luciano Serafini, and Michael Spranger. 2022. ‘Logic Tensor Networks’. Artificial Intelligence 303:103649. https://doi.org/10.1016/j.artint.2021.103649.

Baker, Collin F., Charles J. Fillmore, and John B. Lowe. 1998. ‘The Berkeley FrameNet Project’. In Proceedings of the 36th Annual Meeting of the Association for Computational Linguistics and 17th International Conference on Computational Linguistics, 86–90. https://doi.org/10.3115/980845.980860.

Bayley, O., Savino, E., Slattery, A., and Noël, T. (2024). Autonomous chemistry: navigating self-driving labs in chemical and material sciences. Matter 7, 2382–2398.

Beaulieu-Jones, Brett K., and Casey S.

Becker, J., Rush, N., Barnes, B., and Rein, D. (2025). Measuring the impact of early-2025 AI on experienced open-source developer productivity. METR research report.

Bendale, A., and Boult, T. E. (2016). Towards Open Set Deep Networks. Proceedings of CVPR 2016.

Betti, Arianna, and Hein van den Berg. 2014. ‘Modelling the History of Ideas’. British Journal for the History of Philosophy 22 (4): 812–35. https://doi.org/10.1080/09608788.2014.949217.

Beurer-Kellner, L., Fischer, M., and Vechev, M. (2023). Prompting is programming: A query language for large language models. Proceedings of the ACM on Programming Languages, 7(PLDI). https://doi.org/10.1145/3591300

Birhane, A., Kasirzadeh, A., Leslie, D., and Wachter, S. (2023). Science in the age of large language models. Nature Reviews Physics, 5, 277-280.

Boden, Margaret A. 2004. The Creative Mind: Myths and Mechanisms. 2nd ed. London: Routledge.

Boiko, D. A., MacKnight, R., Kline, B., and Gomes, G. (2023). Autonomous chemical research with large language models. Nature, 624, 570-578. https://doi.org/10.1038/s41586-023-06792-0

Bonney, Rick, et al. "Next Steps for Citizen Science." Science 343, no. 6178 (2014): 1436-1437. doi:10.1126/science.1251554. A concise agenda for building scientific participation through coordinated infrastructure, rigorous methods, and reciprocal relationships with contributors.

Bouman, R., et al. (2024). A benchmark of machine learning methods for anomaly detection. Journal of Machine Learning Research.

Bran, Andres M., et al. "Augmenting Large Language Models with Chemistry Tools." Nature Machine Intelligence 6 (2024): 525-535. doi:10.1038/s42256-024-00832-8.

Brunton, S. L., Proctor, J. L., and Kutz, J. N. (2016). Discovering governing equations from data by sparse identification of nonlinear dynamical systems. Proceedings of the National Academy of Sciences 113(15), 3932–3937.

Brunton, Steven L., Joshua L.

Brynjolfsson, Erik, Danielle Li, and Lindsey R.

Bu, Dechao, et al. "Empowering AI Data Scientists Using a Multi-Agent LLM Framework with Self-Evolving Capabilities for Autonomous, Tool-Aware Biomedical Data Analyses." Nature Biomedical Engineering (2026). doi:10.1038/s41551-026-01634-6.

Burger, B., Maffettone, P. M., Gusev, V. V., et al. (2020). A mobile robotic chemist. Nature, 583, 237-241.

Bush, Vannevar. 1945. ‘As We May Think’. The Atlantic Monthly 176 (1): 101–8.

Canty, R. B., et al. (2025). Science acceleration and accessibility with self-driving labs. Nature Communications 16, 3856.

Chambers, C. D. (2019). The Registered Reports revolution: Lessons in cultural reform. Significance, 16(4), 23-27.

Chambers, Christopher D. "Registered Reports: A New Publishing Initiative at Cortex." Cortex 49, no. 3 (2013): 609-610. doi:10.1016/j.cortex.2012.12.016. A foundational proposal for evaluating the importance and design of a study before its results are known.

Chan, J. S., Chowdhury, N., Jaffe, O., et al. (2024). MLE-bench: Evaluating machine learning agents on machine learning engineering. arXiv:2410.07095.

Chan, Jun Shern, et al. "MLE-Bench: Evaluating Machine Learning Agents on Machine Learning Engineering." Preprint, arXiv:2410.07095 (2024).

Chen, H., Xiong, M., Lu, Y., et al. (2025). MLR-Bench: Evaluating AI agents on open-ended machine learning research. arXiv:2505.19955.

Chen, Ziru, et al. "ScienceAgentBench: Toward Rigorous Assessment of Language Agents for Data-Driven Scientific Discovery." Preprint, arXiv:2410.05080 (2024).

Clark, Tim, Paolo N. Ciccarese, and Carole A. Goble. 2014. 'Micropublications: A Semantic Model for Claims, Evidence, Arguments and Annotations in Biomedical Communications'. Journal of Biomedical Semantics 5:28. https://doi.org/10.1186/2041-1480-5-28.

Clune, J. (2019). AI-GAs: AI-generating algorithms, an alternate paradigm for producing general artificial intelligence. arXiv:1905.10985.

Coley, C. W., Thomas, D. A., Lummiss, J. A. M., et al. (2019). A robotic platform for flow synthesis of organic compounds informed by AI planning. Science, 365, eaax1566.

Common Workflow Language Project (2024-2026). Common Workflow Language specifications and user guide.

Cory-Wright, Ryan, et al. "Evolving Scientific Discovery by Unifying Data and Background Knowledge with AI Hilbert." Nature Communications 15 (2024): 5922. doi:10.1038/s41467-024-50074-w.

Cranmer, M., et al. (2020). Discovering symbolic models from deep learning with inductive biases. Advances in Neural Information Processing Systems 33.

Craver, Carl F., and Lindley Darden. 2013. In Search of Mechanisms: Discoveries Across the Life Sciences. Chicago: University of Chicago Press.

D'Amour, A., et al. (2022). Underspecification Presents Challenges for Credibility in Modern Machine Learning. Journal of Machine Learning Research 23(226), 1–61.

Dai, T., et al. (2024). Autonomous mobile robots for exploratory synthetic chemistry. Nature 635, 890–897.

Daston, Lorraine, and Peter Galison.

Dauparas, J., et al. (2022). Robust deep learning–based protein sequence design using ProteinMPNN. Science 378(6615), 49–56.

De Lange, M., et al. (2022). A Continual Learning Survey: Defying Forgetting in Classification Tasks. IEEE Transactions on Pattern Analysis and Machine Intelligence 44(7), 3366–3385.

de Moura, L., and Bjørner, N. (2008). Z3: An efficient SMT solver. In Tools and Algorithms for the Construction and Analysis of Systems, 337-340.

Dell'Acqua, Fabrizio, et al. "Navigating the Jagged Technological Frontier: Field Experimental Evidence of the Effects of AI on Knowledge Worker Productivity and Quality." Harvard Business School Working Paper 24-013 (2023).

Douglas, Heather.

Dumas, J.-G., Pernet, C., and Sedoglavic, A. (2025). A non-commutative algorithm for multiplying 4×4 matrices using 48 non-complex multiplications. arXiv:2506.13242.

Ecoffet, A., et al. (2021). First return, then explore. Nature 590, 580–586.

Elicit (2025-2026). Systematic review, evidence extraction, and PRISMA-support documentation.

European Commission and European Research Area Forum (2026). Living guidelines on the responsible use of generative AI in research, third version, May 2026.

Fang, Z., Li, Y., Liu, F., Han, B., and Lu, J. (2024). On the Learnability of Out-of-distribution Detection. Journal of Machine Learning Research 25, 1–83.

Farquhar, S., et al. (2024). Detecting hallucinations in large language models using semantic entropy. Nature 630, 625–630.

Fauconnier, Gilles, and Mark Turner. 2002. The Way We Think: Conceptual Blending and the Mind's Hidden Complexities. New York: Basic Books.

Fawzi, A., Balog, M., Huang, A., et al. (2022). Discovering faster matrix multiplication algorithms with reinforcement learning. Nature, 610, 47-53. https://doi.org/10.1038/s41586-022-05172-4

Feurer, Matthias, et al. "Efficient and Robust Automated Machine Learning." In Advances in Neural Information Processing Systems 28 (2015).

Feyerabend, Paul.

Fillmore, Charles J. 1982. 'Frame Semantics'. In Linguistics in the Morning Calm, edited by The Linguistic Society of Korea, 111–37. Seoul: Hanshin.

Floridi, Luciano, and Josh Cowls. "A Unified Framework of Five Principles for AI in Society." Harvard Data Science Review 1, no. 1 (2019). A compact ethical framework organized around beneficence, non-maleficence, autonomy, justice, and explicability.

Fortunato, Santo, Carl T. Bergstrom, Katy Börner, James A. Evans, Dirk Helbing, Staša Milojević, Alexander M. Petersen, et al. 2018. 'Science of Science'. Science 359 (6379): eaao0185. https://doi.org/10.1126/science.aao0185.

Fortunato, Santo, et al. "Science of Science." Science 359, no. 6379 (2018): eaao0185. doi:10.1126/science.aao0185. A large review of teams, careers, citations, institutions, novelty, and the measurable dynamics of knowledge production.

Foster, Jacob G., Andrey Rzhetsky, and James A. Evans. 2015. 'Tradition and Innovation in Scientists' Research Strategies'. American Sociological Review 80 (5): 875–908. https://doi.org/10.1177/0003122415601618.

Frady, E. Paxon, Spencer J. Kent, Bruno A. Olshausen, and Friedrich T. Sommer. 2020. 'Resonator Networks, 1: An Efficient Solution for Factoring High-Dimensional, Distributed Representations of Data Structures'. Neural Computation 32 (12): 2311–31. https://doi.org/10.1162/neco_a_01331.

Frazier, P. I. (2018). A tutorial on Bayesian optimization. arXiv:1807.02811.

Fricker, Miranda.

FutureHouse (2025-2026). Platform, Aviary, PaperQA2, and scientific agent documentation.

Fyfe, Aileen, et al. A History of Scientific Journals: Publishing at the Royal Society, 1665-2015.

Gärdenfors, Peter. 2000. Conceptual Spaces: The Geometry of Thought. Cambridge, MA: MIT Press.

Ganter, Bernhard, and Rudolf Wille. 1999. Formal Concept Analysis: Mathematical Foundations. Berlin: Springer. https://doi.org/10.1007/978-3-642-59830-2.

Garcez, Artur d'Avila, and Luis C. Lamb. 2020. 'Neurosymbolic AI: The 3rd Wave'. https://arxiv.org/abs/2012.05876.

Gebru, Timnit, et al. "Datasheets for Datasets." Communications of the ACM 64, no. 12 (2021): 86-92. doi:10.1145/3458723.

Geirhos, R., et al. (2020). Shortcut learning in deep neural networks. Nature Machine Intelligence 2, 665–673.

Geng, C., Huang, S.-J., and Chen, S. (2021). Recent Advances in Open Set Recognition: A Survey. IEEE Transactions on Pattern Analysis and Machine Intelligence 43(10), 3614–3631.

Gentner, Dedre. 1983. ‘Structure-Mapping: A Theoretical Framework for Analogy’. Cognitive Science 7 (2): 155–70. https://doi.org/10.1207/s15516709cog0702_3.

Gentner, Dedre, and Arthur B. Markman. 1997. ‘Structure Mapping in Analogy and Similarity’. American Psychologist 52 (1): 45–56. https://doi.org/10.1037/0003-066X.52.1.45.

Georgiev, B., Gómez-Serrano, J., Tao, T., and Wagner, A. Z. (2025). Mathematical exploration and discovery at scale. arXiv:2511.02864.

Ghareeb, A. E., Chang, B., Mitchener, L., et al. (2026). A multi-agent system for automating scientific discovery. Nature, 655, 497-505. https://doi.org/10.1038/s41586-026-10652-y

Giannakopoulou, D., Pressburger, T., Mavridou, A., Rhein, J., Schumann, J., and Shi, N. (2020). Formal requirements elicitation with FRET. Joint Proceedings of the REFSQ 2020 Workshops, CEUR-WS Vol. 2584.

Gibbs, I., and Candès, E. (2024). Conformal Inference for Online Prediction with Arbitrary Distribution Shifts. Journal of Machine Learning Research 25.

Goodhart, C. A. E. (1975). Problems of Monetary Management: The U.K. Experience. Papers in Monetary Economics, Reserve Bank of Australia.

Google DeepMind (2025). AlphaEvolve: A Gemini-powered coding agent for designing advanced algorithms.

Google Research (2025-2026). AI Co-Scientist technical materials and scientist-in-the-loop access documentation.

Gottweis, J., Weng, W.-H., Daryin, A., et al. (2026). Accelerating scientific discovery with Co-Scientist. Nature, 655, 487-496. https://doi.org/10.1038/s41586-026-10644-y

Gottweis, Juraj, et al. "Accelerating Scientific Discovery with Co-Scientist." Nature 655 (2026): 487-496. doi:10.1038/s41586-026-10644-y. A multi-agent system for generating, criticizing, ranking, and evolving hypotheses, supported by selected biomedical laboratory validations.

Grünwald, Peter D. 2007. The Minimum Description Length Principle. Cambridge, MA: MIT Press.

Groth, Paul, Andrew Gibson, and Jan Velterop. 2010. ‘The Anatomy of a Nanopublication’. Information Services & Use 30 (1–2): 51–56. https://doi.org/10.3233/ISU-2010-0613.

Gulrajani, I., and Lopez-Paz, D. (2021). In Search of Lost Domain Generalization. Proceedings of ICLR 2021.

Guo, C., Pleiss, G., Sun, Y., and Weinberger, K. Q. (2017). On Calibration of Modern Neural Networks. Proceedings of ICML 2017.

Häse, F., Roch, L. M., and Aspuru-Guzik, A. (2019). Next-generation experimentation with self-driving laboratories. Trends in Chemistry, 1(3), 282-291.

Ha, D., and Schmidhuber, J. (2018). World Models. arXiv:1803.10122.

Haack, Susan. "Six Signs of Scientism." Logos & Episteme 3, no. 1 (2012): 75-95. doi:10.5840/logos-episteme20123151.

Hacking, Ian.

Hao, Q., Xu, F., Li, Y., et al. (2026). Artificial intelligence tools expand scientists' impact but contract science's focus. Nature, 649, 1237-1243. https://doi.org/10.1038/s41586-025-09922-y

Hendrycks, D., and Gimpel, K. (2017). A Baseline for Detecting Misclassified and Out-of-Distribution Examples in Neural Networks. Proceedings of ICLR 2017.

Hernán, M. A., and Robins, J. M. (2020). Causal Inference: What If. Chapman & Hall/CRC.

Hitzler, Pascal, and Md Kamruzzaman Sarker, eds. 2022. Neuro-Symbolic Artificial Intelligence: The State of the Art. Amsterdam: IOS Press. https://doi.org/10.3233/FAIA369.

Hubert, T., et al. (2025). Olympiad-level formal mathematical reasoning with reinforcement learning. Nature. doi:10.1038/s41586-025-09833-y.

Hultcrantz, Monica, et al. "The GRADE Working Group Clarifies the Construct of Certainty of Evidence." Journal of Clinical Epidemiology 87 (2017): 4-13. doi:10.1016/j.jclinepi.2017.05.006.

Humphreys, Paul.

Ioannidis, J. P. A. (2005). Why most published research findings are false. PLoS Medicine, 2(8), e124.

Jansen, P., Côté, M.-A., Khot, T., et al. (2024). DiscoveryWorld: A virtual environment for developing and evaluating automated scientific discovery agents. arXiv:2406.06769.

Jaradeh, Mohamad Yaser, Allard Oelen, Kheir Eddine Farfar, Markus Prinz, Jennifer D'Souza, Gábor Kismihók, Markus Stocker, and Sören Auer. 2019. 'Open Research Knowledge Graph: Next Generation Infrastructure for Semantic Scholarly Knowledge'. https://arxiv.org/abs/1901.10816.

Jasanoff, Sheila.

Jimenez, C. E., Yang, J., Wettig, A., Yao, S., Pei, K., and Press, O. (2024). SWE-bench: Can Language Models Resolve Real-World GitHub Issues? Proceedings of ICLR 2024.

Jumper, J., et al. (2021). Highly accurate protein structure prediction with AlphaFold. Nature 596, 583–589.

Kaissis, Georgios A., et al. "Secure, Privacy-Preserving and Federated Machine Learning in Medical Imaging." Nature Machine Intelligence 2 (2020): 305-311. doi:10.1038/s42256-020-0186-1.

Kanerva, Pentti. 2009. 'Hyperdimensional Computing: An Introduction to Computing in Distributed Representation with High-Dimensional Random Vectors'. Cognitive Computation 1 (2): 139–59. https://doi.org/10.1007/s12559-009-9009-8.

Karniadakis, G. E., et al. (2021). Physics-informed machine learning. Nature Reviews Physics 3, 422–440.

Kent, Spencer J., E. Paxon Frady, Friedrich T. Sommer, and Bruno A. Olshausen. 2020. 'Resonator Networks, 2: Factorization Performance and Capacity Compared to Optimization-Based Methods'. Neural Computation 32 (12): 2332–88. https://doi.org/10.1162/neco_a_01329.

Khattab, O., Singhvi, A., Maheshwari, P., et al. (2024). DSPy: Compiling declarative language model calls into self-improving pipelines. International Conference on Learning Representations. arXiv:2310.03714.

Khemakhem, I., Kingma, D. P., Monti, R. P., and Hyvärinen, A. (2020). Variational Autoencoders and Nonlinear ICA: A Unifying Framework. Proceedings of AISTATS 2020.

King, R. D., Rowland, J., Oliver, S. G., et al. (2009). The automation of science. Science, 324, 85-89. https://doi.org/10.1126/science.1165620

King, R. D., Whelan, K. E., Jones, F. M., et al. (2004). Functional genomic hypothesis generation and experimentation by a robot scientist. Nature, 427, 247-252. https://doi.org/10.1038/nature02236

King, Ross D., et al. "The Automation of Science." Science 324, no. 5923 (2009): 85-89. doi:10.1126/science.1165620.

Kirkpatrick, J., et al. (2017). Overcoming catastrophic forgetting in neural networks. Proceedings of the National Academy of Sciences 114(13), 3521–3526.

Kleyko, Denis, Dmitri A. Rachkovskij, Evgeny Osipov, and Abbas Rahimi. 2023. 'A Survey on Hyperdimensional Computing Aka Vector Symbolic Architectures, Part i: Models and Data Transformations'. ACM Computing Surveys 55 (6): 1–40. https://doi.org/10.1145/3538531.

Kluyver, T., Ragan-Kelley, B., Pérez, F., et al. (2016). Jupyter Notebooks - a publishing format for reproducible computational workflows. In Positioning and Power in Academic Publishing, 87-90. https://doi.org/10.3233/978-1-61499-649-1-87

Knorr Cetina, Karin.

Koh, P. W., et al. (2021). WILDS: A Benchmark of in-the-Wild Distribution Shifts. Proceedings of ICML 2021.

Kon, P. T. J., Liu, J., Zhu, X., et al. (2025). EXP-Bench: Can AI conduct AI research experiments? arXiv:2505.24785.

Krenn, M., Pollice, R., Guo, S. Y., et al. (2022). On scientific understanding with artificial intelligence. Nature Reviews Physics, 4, 761-769.

Kuhn, T. (2014). A survey and classification of controlled natural languages. Computational Linguistics, 40(1), 121-170.

Kuhn, T. S. (1962). The Structure of Scientific Revolutions. University of Chicago Press.

Kuhn, Thomas S.

Kuhn, Tobias, Christine Chichester, Michael Krauthammer, Núria Queralt-Rosinach, Rubén Verborgh, George Giannakopoulos, Axel-Cyrille Ngonga Ngomo, Raffaele Viglianti, and Michel Dumontier. 2018. ‘Decentralized Provenance-Aware Publishing with Nanopublications’. PeerJ Computer Science 4:e78. https://doi.org/10.7717/peerj-cs.78.

Lakatos, I. (1978). The Methodology of Scientific Research Programmes. Cambridge University Press.

Lakatos, Imre. "Falsification and the Methodology of Scientific Research Programmes." In Criticism and the Growth of Knowledge, edited by Imre Lakatos and Alan Musgrave.

Lake, B. M., and Baroni, M. (2023). Human-like systematic generalization through a meta-learning neural network. Nature 623, 115–121.

Lam, Remi, et al. "Learning Skillful Medium-Range Global Weather Forecasting." Science 382, no. 6677 (2023): 1416-1421. doi:10.1126/science.adi2336.

Laurent, J. M., Hsieh, C., Goel, A., et al. (2024). LAB-Bench: Measuring capabilities of language models for biology research. arXiv:2407.10362.

Lebo, T., Sahoo, S., McGuinness, D., et al. (2013). PROV-O: The PROV Ontology. W3C Recommendation.

Lehman, J., and Stanley, K. O. (2011). Abandoning objectives: Evolution through the search for novelty alone. Evolutionary Computation 19(2), 189–223.

Leonelli, Sabina.

Lewis, Patrick, et al. "Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks." In Advances in Neural Information Processing Systems 33 (2020).

Li, Xu, Hanzhe Tu, and Xun Han. 2026. ‘Graph2Idea: Retrieval-Augmented Scientific Idea Generation with Graph-Structured Contexts’. https://arxiv.org/abs/2606.09105.

Lieder, F., and Griffiths, T. L. (2017). Strategy selection as rational metareasoning. Psychological Review, 124, 762-794. https://doi.org/10.1037/rev0000075

Lindsay, R. K., Buchanan, B. G., Feigenbaum, E. A., and Lederberg, J. (1993). DENDRAL: A case study of the first expert system for scientific hypothesis formation. Artificial Intelligence, 61(2), 209-261. https://doi.org/10.1016/0004-3702(93)90068-M

Liu, Junqi, et al. "AutoMedBench: Towards Medical AutoResearch with Agentic AI Models." Preprint, arXiv:2606.01961 (2026). A stage-aware benchmark showing that agents more readily make medical-AI pipelines execute than validate their reliability or submit correct artifacts.

Locatello, F., et al. (2019). Challenging Common Assumptions in the Unsupervised Learning of Disentangled Representations. Proceedings of ICML 2019.

Longino, H. E. (1990). Science as Social Knowledge. Princeton University Press.

Longino, Helen E.

Lovejoy, Arthur O. 1936. The Great Chain of Being: A Study of the History of an Idea. Cambridge, MA: Harvard University Press.

Lu, C., Lu, C., Lange, R. T., et al. (2026). Towards end-to-end automation of AI research. Nature, 651, 914-919. https://doi.org/10.1038/s41586-026-10265-5

Machamer, Peter, Lindley Darden, and Carl F. Craver. 2000. 'Thinking about Mechanisms'. Philosophy of Science 67 (1): 1–25. https://doi.org/10.1086/392759.

MacLeod, B. P., et al. (2020). Self-driving laboratory for accelerated discovery of thin-film materials. Science Advances 6(20), eaaz8867.

Madaan, A., Tandon, N., Gupta, P., et al. (2023). Self-Refine: Iterative refinement with self-feedback. Advances in Neural Information Processing Systems, 36.

Madani, A., et al. (2023). Large language models generate functional protein sequences across diverse families. Nature Biotechnology 41, 1099–1106.

Mandal, I., et al. (2025). Evaluating large language model agents for automation of atomic force microscopy. Nature Communications. doi:10.1038/s41467-025-64105-7.

Manhaeve, Robin, Sebastijan Dumančić, Angelika Kimmig, Thomas Demeester, and Luc De Raedt. 2018. 'DeepProbLog: Neural Probabilistic Logic Programming'. In Advances in Neural Information Processing Systems 31, 3749–59.

Manheim, D., and Garrabrant, S. (2019). Categorizing Variants of Goodhart's Law. arXiv:1803.04585.

Merchant, A., et al. (2023). Scaling deep learning for materials discovery. Nature 624, 80–85.

Merton, Robert K. 1957. 'Priorities in Scientific Discovery: A Chapter in the Sociology of Science'. American Sociological Review 22 (6): 635–59. https://doi.org/10.2307/2089193.

Messeri, L., and Crockett, M. J. (2024). Artificial intelligence and illusions of understanding in scientific research. Nature, 627, 49-58. https://doi.org/10.1038/s41586-024-07146-0

Messeri, Lisa, and M. J.

Microsoft Research (2023-2026). Guidance: Structured and constrained generation documentation.

Mitchell, Margaret, et al. "Model Cards for Model Reporting." In Proceedings of the Conference on Fairness, Accountability, and Transparency (2019): 220-229. doi:10.1145/3287560.3287596.

Mitchell, T. M. (1980). The Need for Biases in Learning Generalizations. CBM-TR-117, Rutgers University.

Model Context Protocol (2024-2026). Protocol specification, architecture, SDKs, and security guidance.

Moreau, Luc, et al. "PROV-DM: The PROV Data Model." W3C Recommendation, 2013.

Mouret, J.-B., and Clune, J. (2015). Illuminating search spaces by mapping elites. arXiv:1504.04909.

Moxham, Noah, and Aileen Fyfe. "The Royal Society and the Prehistory of Peer Review, 1665-1965." The Historical Journal 61, no. 4 (2018): 863-889. doi:10.1017/S0018246X17000334. A corrective to the myth that modern external peer review has existed unchanged since the first scientific journals.

Munafò, M. R., Nosek, B. A., Bishop, D. V. M., et al. (2017). A manifesto for reproducible science. Nature Human Behaviour, 1, 0021. https://doi.org/10.1038/s41562-016-0021

Munafò, Marcus R., et al. "A Manifesto for Reproducible Science." Nature Human Behaviour 1 (2017): 0021. doi:10.1038/s41562-016-0021. A synthesis of reforms in study design, reporting, collaboration, incentives, and methodological training.

Newell, A., and Simon, H. A. (1956). The logic theory machine - a complex information processing system. IRE Transactions on Information Theory, 2(3), 61-79. https://doi.org/10.1109/TIT.1956.1056797

Newman, William R. "What Have We Learned from the Recent Historiography of Alchemy?" Isis 102, no. 2 (2011): 313-321. doi:10.1086/660140. A concise guide to the recovery of alchemy as a serious material and intellectual practice rather than a comic prelude to chemistry.

Nosek, B. A., Alter, G., Banks, G. C., et al. (2015). Promoting an open research culture. Science, 348, 1422-1425. https://doi.org/10.1126/science.aab2374

Nosek, B. A., Ebersole, C. R., DeHaven, A. C., and Mellor, D. T. (2018). The preregistration revolution. Proceedings of the National Academy of Sciences, 115, 2600-2606.

Nosek, Brian A., et al. "Promoting an Open Research Culture." Science 348, no. 6242 (2015): 1422-1425. doi:10.1126/science.aab2374.

Novikov, A., Vũ, N., Eisenberger, M., et al. (2025). AlphaEvolve: A coding agent for scientific and algorithmic discovery. arXiv:2506.13131.

O'Neil, Cathy.

Open Science Collaboration. "Estimating the Reproducibility of Psychological Science." Science 349, no. 6251 (2015): aac4716. doi:10.1126/science.aac4716.

OpenAI (2024). MLE-bench: Evaluating machine learning agents on machine learning engineering.

OpenAI (2025). PaperBench: Evaluating AI's ability to replicate AI research.

Oreskes, Naomi.

Outlines Project (2023-2026). Structured generation and grammar-constrained decoding documentation.

Ovadia, Y., et al. (2019). Can You Trust Your Model’s Uncertainty? Evaluating Predictive Uncertainty under Dataset Shift. Advances in Neural Information Processing Systems 32.

Page, M. J., McKenzie, J. E., Bossuyt, P. M., et al. (2021). The PRISMA 2020 statement: An updated guideline for reporting systematic reviews. BMJ, 372, n71.

Panapitiya, G., Saldanha, E., Job, H., et al. (2026). AutoLabs: Cognitive multi-agent systems with self-correction for autonomous chemical experimentation. Scientific Reports, 16, 19554. https://doi.org/10.1038/s41598-026-45593-z

Parisi, G. I., Kemker, R., Part, J. L., Kanan, C., and Wermter, S. (2019). Continual lifelong learning with neural networks: A review. Neural Networks 113, 54–71.

Park, M., Leahey, E., and Funk, R. J. (2023). Papers and patents are becoming less disruptive over time. Nature, 613, 138-144. https://doi.org/10.1038/s41586-022-05543-x

Pearl, J. (2009). Causality: Models, Reasoning, and Inference. Cambridge University Press.

Peng, R. D. (2011). Reproducible research in computational science. Science, 334, 1226-1227. https://doi.org/10.1126/science.1213847

Peters, J., Janzing, D., and Schölkopf, B. (2017). Elements of Causal Inference. MIT Press.

Plate, Tony A. 1995. ‘Holographic Reduced Representations’. IEEE Transactions on Neural Networks 6 (3): 623–41. https://doi.org/10.1109/72.377968.

Polanyi, Michael.

Popper, K. R. (1959). The Logic of Scientific Discovery. Hutchinson.

Popper, Karl R.

Porter, Benjamin, and Declan McIntosh. "The Network Structure of Scientific Revolutions, Paper Mills and Anomalous Publishing Activity." Scientific Reports 14 (2024): 29569. doi:10.1038/s41598-024-71230-8.

Priem, Jason, Heather Piwowar, and Richard Orr. 2022. 'OpenAlex: A Fully-Open Index of Scholarly Works, Authors, Venues, Institutions, and Concepts'. https://arxiv.org/abs/2205.01833.

Principe, Lawrence M.

Ríos-García, M., Alampara, N., Gupta, C., et al. (2026). AI scientists produce results without reasoning scientifically. arXiv:2604.18805. Preprint.

Radensky, Marissa, Shibani Shahid, Raymond Fok, Pao Siangliulue, Tom Hope, and Daniel S. Weld. 2024. 'Scideator: Human–LLM Scientific Idea Generation Grounded in Research-Paper Facet Recombination'. https://arxiv.org/abs/2409.14634.

Raedt, Luc De, Sebastijan Dumančić, Robin Manhaeve, and Giuseppe Marra. 2020. 'From Statistical Relational to Neuro-Symbolic Artificial Intelligence'. In Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence, 4943–50. https://doi.org/10.24963/ijcai.2020/688.

Rahmanian, Farshid, et al. "Enabling Modular Autonomous Feedback-Loops in Materials Science through Hierarchical Experimental Laboratory Automation and Orchestration." Advanced Materials Interfaces 9 (2022): 2101987. doi:10.1002/admi.202101987. HELAO provides an architecture for asynchronous instrument control and reusable experimental workflows.

Raissi, M., Perdikaris, P., and Karniadakis, G. E. (2019). Physics-informed neural networks: A deep learning framework for solving forward and inverse problems involving nonlinear partial differential equations. Journal of Computational Physics 378, 686–707.

Ram, K. (2013). Git can facilitate greater reproducibility and increased transparency in science. Source Code for Biology and Medicine, 8, 7. https://doi.org/10.1186/1751-0473-8-7

Recht, B., Roelofs, R., Schmidt, L., and Shankar, V. (2019). Do ImageNet Classifiers Generalize to ImageNet? Proceedings of ICML 2019.

Reichenbach, H. (1938). Experience and Prediction. University of Chicago Press.

Richardson, R. A. K., et al. (2025). The entities enabling scientific fraud at scale are large, resilient, and growing rapidly. Proceedings of the National Academy of Sciences, 122, e2420092122. https://doi.org/10.1073/pnas.2420092122

Rissanen, Jorma. 1978. 'Modeling by Shortest Data Description'. Automatica 14 (5): 465–71. https://doi.org/10.1016/0005-1098(78)90005-5.

RO-Crate Research Object Community (2026). RO-Crate Metadata Specification 1.3.

Romera-Paredes, B., Barekatain, M., Novikov, A., et al. (2024). Mathematical discoveries from program search with large language models. Nature, 625, 468-475. https://doi.org/10.1038/s41586-023-06924-6

Rosch, Eleanor. 1975. 'Cognitive Representations of Semantic Categories'. Journal of Experimental Psychology: General 104 (3): 192–233. https://doi.org/10.1037/0096-3445.104.3.192.

Royal Society.

Ruan, K., Wang, X., Hong, J., et al. (2026). Evaluating large language models' divergent thinking capabilities for scientific idea generation with minimal context. Nature Communications. doi:10.1038/s41467-026-70245-1.

Ruff, L., et al. (2021). A Unifying Review of Deep and Shallow Anomaly Detection. Proceedings of the IEEE 109(5), 756–795.

Runco, Mark A., and Garrett J. Jaeger. 2012. ‘The Standard Definition of Creativity’. Creativity Research Journal 24 (1): 92–96. https://doi.org/10.1080/10400419.2012.650092.

Sagawa, S., Koh, P. W., Hashimoto, T. B., and Liang, P. (2020). Distributionally Robust Neural Networks for Group Shifts. Proceedings of ICLR 2020.

Sakana AI (2024-2026). The AI Scientist technical reports, software releases, and AI Scientist-v2 materials.

Salazar-Villacis, P., et al. (2026). The ADePT framework for assessing autonomous experimentation systems. Communications Chemistry. doi:10.1038/s42004-026-01932-9.

Sarker, Md Kamruzzaman, Lu Zhou, Aaron Eberhart, and Pascal Hitzler. 2021. ‘Neuro-Symbolic Artificial Intelligence: Current Trends’. https://arxiv.org/abs/2105.05330.

Schölkopf, B., et al. (2021). Toward Causal Representation Learning. Proceedings of the IEEE 109(5), 612–634.

Schlegel, Kevin, Peer Neubert, and Peter Protzel. 2022. ‘A Comparison of Vector Symbolic Architectures’. Artificial Intelligence Review 55:4523–55. https://doi.org/10.1007/s10462-021-10110-3.

Schmidgall, Samuel, et al. "Agent Laboratory: Using LLM Agents as Research Assistants." Preprint, arXiv:2501.04227 (2025). A multi-agent pipeline for literature review, experimentation, and report writing whose evaluations also show the value of structured human feedback and the risk of automated overrating.

Schmidgall, Samuel, and Michael Moor. "AgentRxiv: Towards Collaborative Autonomous Research." Preprint, arXiv:2503.18102 (2025).

Schmidt, M., and Lipson, H. (2009). Distilling free-form natural laws from experimental data. Science, 324, 81-85.

Schopf, Tim, and Michael Färber. 2026. ‘Is This Idea Novel? An Automated Benchmark for Judgment of Research Ideas’. https://arxiv.org/abs/2603.10303.

Schulz, Kenneth F., Douglas G.

Scite (2025-2026). Smart Citations, evidence search, and MCP integration documentation.

Selected primary research, surveys and system reports cited in the text. Publication status is described in the research disclosure.

Settles, B. (2009). Active Learning Literature Survey. University of Wisconsin-Madison Computer Sciences Technical Report 1648.

Shahid, Shibani, Marissa Radensky, Raymond Fok, Pao Siangliulue, Daniel S. Weld, and Tom Hope. 2025. ‘Literature-Grounded Novelty Assessment of Scientific Ideas’. https://arxiv.org/abs/2506.22026.

Shahriari, B., Swersky, K., Wang, Z., Adams, R. P., and de Freitas, N. (2016). Taking the Human Out of the Loop: A Review of Bayesian Optimization. Proceedings of the IEEE 104(1), 148–175.

Shao, Y., Jiang, Y., Kanell, T., et al. (2024). Assisting in writing Wikipedia-like articles from scratch with large language models. Proceedings of NAACL 2024.

Shapin, Steven. A Social History of Truth: Civility and Science in Seventeenth-Century England.

Shapin, Steven, and Simon Schaffer.

Shinn, N., Cassano, F., Gopinath, A., et al. (2023). Reflexion: Language agents with verbal reinforcement learning. Advances in Neural Information Processing Systems, 36.

Si, Chenglei, Diyi Yang, and Tatsunori Hashimoto. 2024. ‘Can LLMs Generate Novel Research Ideas? A Large-Scale Human Study with 100+ NLP Researchers’. https://arxiv.org/abs/2409.04109.

Sim, Miles, et al. "ChemOS 2.0: An Orchestration Architecture for Chemical Self-Driving Laboratories." Matter 7 (2024): 2959-2977. doi:10.1016/j.matt.2024.04.022.

Simmons, Joseph P., Leif D.

Skarlinski, M. D., Cox, S., Laurent, J. M., et al. (2024). Language agents achieve superhuman synthesis of scientific knowledge. arXiv:2409.13740.

Skinner, Quentin. 1969. ‘Meaning and Understanding in the History of Ideas’. History and Theory 8 (1): 3–53. https://doi.org/10.2307/2504188.

Smaldino, P. E., and McElreath, R. (2016). The natural selection of bad science. Royal Society Open Science, 3, 160384. https://doi.org/10.1098/rsos.160384

Smolensky, Paul. 1990. ‘Tensor Product Variable Binding and the Representation of Symbolic Structures in Connectionist Systems’. Artificial Intelligence 46 (1–2): 159–216. https://doi.org/10.1016/0004-3702(90)90007-M.

Snoek, J., Larochelle, H., and Adams, R. P. (2012). Practical Bayesian optimization of machine learning algorithms. Advances in Neural Information Processing Systems, 25.

Soiland-Reyes, S., Sefton, P., Crosas, M., et al. (2022). Packaging research artefacts with RO-Crate. Data Science, 5, 97-138. https://doi.org/10.3233/DS-210053

Soiland-Reyes, Stian, et al. "Packaging Research Artefacts with RO-Crate." Data Science 5, no. 2 (2022): 97-138. doi:10.3233/DS-210053. RO-Crate treats research as a connected object of data, code, workflows, people, and machine-readable metadata rather than as a PDF alone.

Soldatova, L. N., Aubrey, W., King, R. D., and Clare, A. (2008). The EXACT description of biomedical protocols. Bioinformatics, 24(13), i295-i303. https://doi.org/10.1093/bioinformatics/btn156

Stach, E., et al. (2021). Autonomous experimentation systems for materials development: A community perspective. Matter 4, 2702–2726.

Stanley, K. O., Lehman, J., and Soros, L. (2017). Open-endedness: The last grand challenge you’ve never heard of. O’Reilly AI.

Starace, G., Jaffe, O., Sherburn, D., et al. (2025). PaperBench: Evaluating AI's ability to replicate AI research. OpenAI research report.

Starace, Giulio, et al. "PaperBench: Evaluating AI's Ability to Replicate AI Research." OpenAI technical report (2025).

Stenmark, Mikael. "What Is Scientism?" Religious Studies 33, no. 1 (1997): 15-32. doi:10.1017/S0034412596003666. A useful taxonomy of ways in which the authority or domain of science can be extended beyond what its methods justify.

Sternlicht, Noy, and Tom Hope. 2026. ‘CHIMERA: A Knowledge Base of Scientific Idea Recombinations for Research Analysis and Ideation’. In Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), 1871–1905. San Diego, California: Association for Computational Linguistics. https://doi.org/10.18653/v1/2026.acl-long.85.

Stigler, Stephen M. 1980. ‘Stigler’s Law of Eponymy’. Transactions of the New York Academy of Sciences 39 (1 Series II): 147–57. https://doi.org/10.1111/j.2164-0947.1980.tb02775.x.

Stokes, Donald E. 1997. Pasteur’s Quadrant: Basic Science and Technological Innovation. Washington, DC: Brookings Institution Press.

Sun, Y., Wang, X., Liu, Z., Miller, J., Efros, A. A., and Hardt, M. (2020). Test-Time Training with Self-Supervision for Generalization under Distribution Shifts. Proceedings of ICML 2020.

Sutton, R. S. (2019). The Bitter Lesson. Research essay.

Swanson, D. R. (1986). Fish oil, Raynaud's syndrome, and undiscovered public knowledge. Perspectives in Biology and Medicine, 30, 7-18.

Szymanski, N. J., Rendy, B., Fei, Y., et al. (2023). An autonomous laboratory for the accelerated synthesis of novel materials. Nature, 624, 86-91. https://doi.org/10.1038/s41586-023-06734-w

Szymanski, N. J., Rendy, B., Fei, Y., et al. (2026). Author Correction: An autonomous laboratory for the accelerated synthesis of inorganic materials. Nature 650, E1. doi:10.1038/s41586-025-09992-y.

Tang, X., et al. (2025). Risks of AI scientists: prioritizing safeguarding over autonomy. Nature Communications. doi:10.1038/s41467-025-63913-1.

Taori, R., et al. (2020). Measuring Robustness to Natural Distribution Shifts in Image Classification. Advances in Neural Information Processing Systems 33.

Terwilliger, Thomas C., et al. "AlphaFold Predictions Are Valuable Hypotheses and Accelerate but Do Not Replace Experimental Structure Determination." Nature Methods 21 (2024): 110-116. doi:10.1038/s41592-023-02087-4. A careful account of where predicted protein structures assist experiment and where ligands, conformations, and density still require empirical work.

Thakkar, N., Yuksekgonul, M., Silberg, J., et al. (2026). A large-scale randomized study of large language model feedback in peer review. Nature Machine Intelligence, 8, 326-336. https://doi.org/10.1038/s42256-026-01188-x

Thede, L., Roth, K., Bethge, M., Akata, Z., and Hartvigsen, T. (2025). WikiBigEdit: Understanding the Limits of Lifelong Knowledge Editing in Large Language Models. Proceedings of ICML, PMLR 267, 59326–59354.

Thede, L., Roth, K., Henaff, O. J., Bethge, M., and Akata, Z. (2025). Reflecting on the State of Rehearsal-Free Continual Learning with Pretrained Models. Proceedings of CoLLAs, PMLR 274, 1076–1093.

Tom, G., et al. (2024). Self-driving laboratories for chemistry and materials science. Chemical Reviews 124, 9633–9732.

Trehan, Dhruv, and Paras Chopra. "Why LLMs Aren't Scientists Yet: Lessons from Four Autonomous Research Attempts." Preprint, arXiv:2601.03315 (2026). A transparent failure study documenting implementation drift, memory degradation, weak scientific taste, and premature declarations of success across end-to-end attempts.

Trinh, T. H., Wu, Y., Le, Q. V., He, H., and Luong, T. (2024). Solving olympiad geometry without human demonstrations. Nature, 625, 476-482. https://doi.org/10.1038/s41586-023-06747-5

Udrescu, S.-M., and Tegmark, M. (2020). AI Feynman: A physics-inspired method for symbolic regression. Science Advances 6(16), eaay2631.

Uzzi, Brian, Satyam Mukherjee, Michael Stringer, and Ben Jones. 2013. ‘Atypical Combinations and Scientific Impact’. Science 342 (6157): 468–72. https://doi.org/10.1126/science.1240474.

van de Schoot, Rens, et al. "An Open Source Machine Learning Framework for Efficient and Transparent Systematic Reviews." Nature Machine Intelligence 3 (2021): 125-133. doi:10.1038/s42256-020-00287-7. ASReview uses active learning to prioritize screening while preserving human inclusion decisions and auditability.

Varici, B., et al. (2025). Score-based Causal Representation Learning. Journal of Machine Learning Research.

Veit, W. (2020). Model pluralism. Philosophy of the Social Sciences, 50, 91-114. https://doi.org/10.1177/0048393119894897

Vovk, V., Gammerman, A., and Shafer, G. (2005). Algorithmic Learning in a Random World. Springer.

W3C (2013). PROV-O: The PROV Ontology. W3C Recommendation.

Wang, D., et al. (2021). Tent: Fully Test-Time Adaptation by Entropy Minimization. Proceedings of ICLR 2021.

Wang, H., Fu, T., Du, Y., et al. (2023). Scientific discovery in the age of artificial intelligence. Nature, 620, 47-60.

Wang, Jian, Reinhilde Veugelers, and Paula Stephan. 2017. ‘Bias Against Novelty in Science: A Cautionary Tale for Users of Bibliometric Indicators’. Research Policy 46 (8): 1416–36. https://doi.org/10.1016/j.respol.2017.06.006.

Wang, R., Lehman, J., Clune, J., and Stanley, K. O. (2019). Paired Open-Ended Trailblazer: Endlessly Generating Increasingly Complex and Diverse Learning Environments and Their Solutions. arXiv:1901.01753.

Watson, J. L., et al. (2023). De novo design of protein structure and function with RFdiffusion. Nature 620, 1089–1100.

Weisberg, Michael, and Ryan Muldoon. "Epistemic Landscapes and the Division of Cognitive Labor." Philosophy of Science 76, no. 2 (2009): 225-252. A model of why communities benefit when some researchers explore unfashionable regions instead of converging too quickly on the same promising peak.

Wells, H. G. 1938. World Brain. London: Methuen.

Wilkinson, M. D., Dumontier, M., Aalbersberg, I. J., et al. (2016). The FAIR Guiding Principles for scientific data management and stewardship. Scientific Data, 3, 160018.

Wittgenstein, Ludwig. 1953. Philosophical Investigations. Oxford: Blackwell.

Wolpert, D. H., and Macready, W. G. (1997). No Free Lunch theorems for optimization. IEEE Transactions on Evolutionary Computation, 1, 67-82. https://doi.org/10.1109/4235.585893

World Wide Web Consortium. 2009. ‘SKOS Simple Knowledge Organization System Reference’. W3C Recommendation.

Wuchty, S., Jones, B. F., and Uzzi, B. (2007). The increasing dominance of teams in production of knowledge. Science, 316, 1036-1039. https://doi.org/10.1126/science.1136099

Xiong, Lei, et al. "AutoResearchBench: Benchmarking AI Agents on Complex Scientific Literature Discovery." Preprint, arXiv:2604.25256 (2026).

Xu, Wanghan, et al. "ResearchClawBench: A Benchmark for End-to-End Autonomous Scientific Research." Preprint, arXiv:2606.07591 (2026).

Yao, S., Yu, D., Zhao, J., et al. (2023). Tree of Thoughts: Deliberate problem solving with large language models. Advances in Neural Information Processing Systems, 36.

Yao, S., Zhao, J., Yu, D., et al. (2023). ReAct: Synergizing reasoning and acting in language models. International Conference on Learning Representations.

Zhang, J., Hu, S., Lu, C., Lange, R. T., and Clune, J. (2026). Darwin Gödel Machine: Open-Ended Evolution of Self-Improving Agents. Proceedings of ICLR 2026; arXiv:2505.22954.

Zhou, Yifan, Qihao Yang, Yan Li, Donggang Li, Xiru Hu, Hokin Deng, Ziyang Gong, et al. 2026. ‘Ideas Have Genomes: Benchmarking Scientific Lineage Reasoning and Lineage-Grounded Idea Generation’. https://arxiv.org/abs/2607.08758.