Cureus Journal of AI-Augmented Research
https://doi.org/10.7759/s44568-026-00351-7

Editorial

The Discovery Gap: Why Artificial Intelligence Must Learn to Change Its Mind "*Rethinking Human–AI Collaboration from Molecule to Market*"

Silvano Coletti [1, 2]

1. *Department of Engineering Sciences, Università degli Studi Guglielmo Marconi, Rome, ITA*
2. *Computational Engineering and Complex Systems, Chelonia SA, Lugano, CHE*

Received: August 06, 2026 | Review began: August 10, 2026 | Review ended: October 06, 2026 | Published: November 03, 2026

© **Copyright** 2026
This is an open access article distributed under the terms of the Creative Commons Attribution License CC-BY 4.0., which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited.

Abstract

Scientific discovery requires the revision of belief: giving up a claim when the evidence turns against it, and knowing what else must change with it. Artificial intelligence (AI) has transformed research - predicting protein structures, generating molecules, screening billions of candidates - but it does not do this. Autonomous agents traverse the discovery loop end to end without holding the beliefs that loop exists to revise. I use dynamic scientific intelligence for the class of system this requires, one that holds hypotheses as persistent objects with provenance, uncertainty and explicit dependencies. Its distinguishing condition is that a system can identify what supports each commitment and what depends on it, and so revise one without disturbing the rest. Because no architecture can revise beliefs on evidence it never receives, the problem is jointly architectural and infrastructural. Using drug development from molecule to market as the test case, I identify where evidence generated downstream should revise commitments made upstream and does not, and three decision points at which the non-delegable human act is adjudication rather than generation. I propose a revision test - an evaluation distinct in kind from any predictive benchmark - and a minimum reporting standard for claims of AI-assisted discovery. Until AI learns to change its mind well, it will remain a tool rather than an infrastructure in which discovery accumulates.

Categories: Human-AI Collaboration

Keywords: dynamic scientific intelligence, scientific discovery, hypothesis revision, belief revision, causal reasoning, human-ai collaboration, drug development, research governance

Editorial

A loop that runs but does not revise

The 2024 Nobel Prize in Chemistry, awarded for advances in computational protein design and structure prediction, recognized AlphaFold2 as a decisive advance on a fifty-year problem [1], with structure prediction accounting for part of the award [2]. Machine learning has since generated novel molecules [3] and new classes of antibiotics [4], autonomous systems plan and execute experiments [5], and agentic pipelines carry a question from literature search to a drafted manuscript [6]. Yet one pattern is suggestive: in the first systematic analysis of AI-native pipelines - a small cohort the authors acknowledge as underpowered - compounds of AI origin cleared Phase I at 80-90% [7], above historical baselines of roughly 40-65% [8], then fell back to about 40% at Phase II, indistinguishable from conventionally discovered drugs [7]. The most advanced candidate with both an AI-derived target and an AI-designed lead, rentosertib, completed a Phase IIa trial in seventy-one patients whose primary endpoint was the incidence of treatment-emergent adverse events; a forced

vital capacity gain of 98.4 mL [95% confidence interval (CI) 10.9-185.9] at the highest dose was a secondary outcome [9]. A Phase III trial was registered but had not begun recruiting at the time of writing [10], and no drug of this class has yet been approved.

These figures do not establish the thesis of this paper. The most parsimonious of the four competing explanations needs no appeal to belief revision at all: generative models are optimized on the very properties that determine Phase I success, whereas Phase II measures efficacy, a property of target biology. Portfolio concentration, cohort immaturity and sponsor confounding are equally live, and, with a cohort this size, none can be excluded.

I therefore take the clinical data as motivating illustration rather than evidence. The argument does not depend on which explanation wins. Suppose the parsimonious one is right and the field has simply not validated its targets. Someone must then decide, program by program, whether the next failure implicates the target, the molecule, the trial design or the disease model. That means locating the failed commitment within the chain of reasoning that produced the program, and that chain is nowhere represented in machine-readable form. The selection explanation therefore does not compete with the argument. It describes the situation in which the inability to attribute failure becomes expensive, since a field that must validate its targets must be able to tell which target claim just died and what rested on it.

The gap lies not in the capacity to run the discovery loop but in what happens inside it. Large language models encode knowledge implicitly in parameters, with no representation that distinguishes established facts from speculative hypotheses. Targeted revision is not impossible: localized model editing can alter a specific factual association without full retraining [11,12]. But such edits are assessed on their locality and reliability, and they degrade under sequential application and generalize unevenly to the claims that follow from what was edited [13]; retraining, meanwhile, tends to overwrite prior competence [14]. What is missing is not the ability to change a parameter but the structure that would say which belief a new observation bears on, and what else rests on it. Architecture is only half of the account. Every year, the field generates thousands of expensive refutations and absorbs few, because trial results are frequently not reported within the mandated window [15] and, when published, are rarely linked retrievably to the mechanistic claim they refute. No architecture can revise beliefs on evidence that never reaches it, so the problem is jointly architectural and infrastructural - the second a precondition for the first.

What has been tried

Table *1* sets out the antecedents and why each stopped short. The bottleneck that halted the belief-revision tradition was never its theory but the cost of populating a knowledge base by hand - precisely the cost learned representations are positioned to relieve.

Program	What it achieved	Why it did not scale
Robot scientists (ADAM, EVE) [16,17]	Closed-loop hypothesis generation, discriminating experiment selection and revision in yeast genomics and drug repurposing	Hypotheses expressed in hand-built formalisms over closed domains, without the perceptual breadth to read the literature
Nobel Turing Challenge [18]	Program for systems that autonomously make Nobel-caliber discoveries	Orthogonal criterion: autonomy, how much of the loop runs unaided, rather than revisability — a system can be fully autonomous and entirely unrevisable
AGM belief revision [19]; assumption-based TMS [20]	Minimal consistency-preserving change; dependency-directed backtracking and localized retraction	Theory was never the obstacle — the cost of populating and maintaining the knowledge base was

How to cite this article:
Coletti S (November 03, 2026) The Discovery Gap: Why Artificial Intelligence Must Learn to Change Its Mind "*Rethinking Human–AI Collaboration from Molecule to Market*". Cureus J AI Augment Res 1 : es44568-026-00351-7. DOI https://doi.org/10.7759/s44568-026-00351-7

STAHLp, incremental revision of chemical models [21,22]	Scientific discovery as belief revision with explicit assumption tracking, in chemistry	Narrow domain, hand-built representations
Bayesian experimental design [23]	Mature account of allocating effort by expected information gain	Diverges from severity [24] where it matters: an experiment that could not have refuted the hypothesis can still be highly informative
DISK and its hypothesis ontology [25]	Machine-readable hypothesis evolution with provenance and automated re-analysis triggers	Not coupled to learned, literature-scale representations
Agentic hypothesis management [26]; multi-agent exploration [27]	Auditable hypothesis handling inside agent architectures, at scale	No dependency-tracked belief state; revision is not localized

TABLE 1: Prior art and what remains unbuilt

What has not been attempted, to my knowledge, is their integration at the scale of contemporary models, instrumented so the revision process can be audited. That is a claim about engineering rather than theory: dynamic scientific intelligence (DSI) is not a new theory of belief revision, since AGM and the truth maintenance tradition supplied that. What remains unbuilt is a system holding learned, literature-scale representations in dependency-tracked form, wired to the evidence streams that would refute them, whose revisions can be inspected afterwards. Each component exists; none of the systems holding them is the same system.

What a system that changes its mind would need

I use DSI for the missing class: systems organized around an explicit, evolving representation of what is known, how strongly and on what evidence.

Because Bayesian inference, active learning and online learning all revise beliefs on evidence, the boundary needs stating precisely. The diagnostic condition for DSI is that the system can say, for any commitment it holds, which observations support it and which claims depend on it, and can therefore retract that commitment alone, propagate the consequences to its dependents, and leave the rest untouched. A Bayesian active-learning loop updates a distribution over a hypothesis space and records the posterior, not the grounds: reweighting changes every hypothesis by some amount and localizes nothing, because the provenance and dependency edges that identify what to change are not represented.

The first is treating hypotheses as first-class objects. In today's systems, a hypothesis is absorbed into the training objective, beyond interrogation; even end-to-end artificial intelligence (AI) scientist pipelines inherit this, and controlled evaluation of two such systems documents inappropriate benchmark selection, data leakage, metric misuse and post-hoc selection bias - errors detectable in the execution trace but not in the finished paper [28]. Agency runs the loop; DSI holds the commitments the loop is meant to revise. Table *2* sets out the fields required, worked through the TNIK/idiopathic-pulmonary-fibrosis hypothesis behind the rentosertib program above; the last field, refutation conditions stated in advance, is what distinguishes a hypothesis from a prediction and makes revision localizable.

Field	Description	Illustrative example

Claim	Assertion stated precisely enough to be falsified	"TNIK hyperactivation drives fibrosis in IPF via downstream TGF-β signalling"
Scope	Population, context and conditions of applicability	"Human IPF patients with elevated TNIK expression (IHC score ≥ 2)"
Provenance	Evidence supporting the claim, with retrievable source identifiers	GWAS hit, mouse knockout phenotype, transcriptomic co-expression (linked by DOI or accession)
Belief state	Current confidence estimate with uncertainty bounds	0.72 ± 0.11 (prior to Phase IIa readout)
Dependency links	Upstream claims this hypothesis rests on; downstream claims that depend on it	Target biology hypothesis → lead optimization → trial endpoint selection
Refutation conditions	Observations that trigger downward revision, specified in advance	Phase II primary endpoint failure in ≥ 2 pre-specified stratified subgroups

TABLE 2: Minimum fields for a DSI hypothesis object

IPF: idiopathic pulmonary fibrosis; TNIK: TRAF2 and NCK-interacting protein kinase; TGF-β: transforming growth factor beta; GWAS: genome-wide association study; IHC: immunohistochemistry; DSI: dynamic scientific intelligence

The second is uncertainty as the engine of exploration. This means separating irreducible measurement noise from the knowledge gaps a targeted experiment could close [29], then following the severity of the test: how probable it is that an error would have been detected had one been present [24]. In drug development, that criterion is not merely epistemic, since the experiment that most severely tests a mechanistic commitment is often the one whose negative result averts the largest downstream expenditure. The third is explicit causal structure. Prediction exploits association; intervention requires causation. A gene associated with a disease may be a biomarker, a modifier or a driver, and observational data cannot distinguish these cases in general, without further assumptions or interventional evidence [30]. Methods exist, including Bayesian causal discovery under unknown interventions [31], but they run as standalone tools rather than layers within a hypothesis-object architecture. The fourth is multimodal evidence with provenance, since genomic associations, phenotypic screens and clinical signals modify belief through distinct mechanisms [32]. Tracking which experiment supports which claim lets a system locate a disagreement rather than average it away [33].

The four are interdependent: hypothesis objects without provenance [32,33] cannot be revised for cause, uncertainty [29] has nothing to guide without an explicit hypothesis to attach it to [28], and causal structure [30] propagates nothing when an assumption fails unless dependency links record what rests on it. Revision then travels only along those links, so its cost scales with the claims actually affected rather than the size of the knowledge base - which answers the objection that continuous re-evaluation is intractable.

Three points where evidence cannot decide

Even a fully realized DSI system meets decision points at which evidence underdetermines action. My claim about them concerns accountability rather than cognition: not that a machine cannot in principle perform the act, but that responsibility for it cannot be transferred, so human authority should be designed in rather than engineered out.

Problem selection reflects values that cannot be derived from data. Across 41.3 million publications, scientists adopting AI tools published 3.02 times as many papers and received 4.84 times as many citations as non-adopters, while the collective volume of topics studied contracted by 4.63% and engagement between scientists themselves fell by 22.00%

[34]. The authors interpret that shift as tracking data availability rather than citation-seeking; on that reading, the effect is harder to correct than a misaligned objective, because infrastructure concentrating capability on data-rich questions neglects rare diseases and foundational biology however well each system is aligned.

Observation validation requires judging whether a design could realistically have refuted the result - a judgement about context, not a statistical property of the data [24], since severity depends on what else could have produced the observation.

Failure interpretation raises the hardest of the three problems. Logic alone cannot determine which element of a reasoning chain was wrong, because hypotheses face evidence only in bundles [35]: a clinical failure may implicate the target, the molecule, the trial design or the disease model. A DSI system can and should localize failure within its represented hypothesis space, which no system I am aware of does. It can also, in principle, extend that space; the fifth dimension of the revision test measures exactly this, and I do not claim the act is beyond machines. What is not delegable is judging that an extension is the right one rather than a plausible confabulation - an assessment appealing to background knowledge and to what is worth pursuing. The human contribution is adjudication, not generation.

These are questions of responsibility, which is why they reach law before benchmarks. Courts in the United States and the United Kingdom have both held that an inventor must be a natural person [36,37] - in the UK by its court of final resort, in the US by the Federal Circuit with certiorari denied - as has the EPO's Legal Board of Appeal. The record is not uniform - South Africa's registry granted a DABUS patent, and an Australian first-instance judgment to the same effect was reversed on appeal - but no court of final resort has held otherwise. Those rulings settle who is accountable, not what a machine can do, and I rely on them only for the former. Authority is in any case necessary but not sufficient. A preregistered meta-analysis of 106 studies found that human-AI combinations underperform the better partner alone on average (Hedges' g = -0.23, 95% CI -0.39 to -0.07), with losses concentrated in decision tasks and gains in creation tasks [38]. Collaboration adds value where the human contributes generative judgement and subtracts it where a human must arbitrate between their own assessment and a system's output with no shared representation of why either holds.

From molecule to market

Drug development is the clearest instance. Its loop is long, its refutations are expensive and well documented, and its regulatory structure already demands the auditability DSI would supply. AI is deployed at every stage [39], but each is optimized independently, and that independence is the defect. The commitment made at the front - this entity drives this pathology in this population - rests on largely associational evidence and is what Phase II most often refutes. When refutation arrives years later as a clinical endpoint rather than a molecular claim, little propagates back. My reading of public disclosures is that some sponsors do route clinical signals into their target models, but informally and within one organization's boundary: the causal claim carries no recorded downgrade, and downstream chemistry optimizes the target as though validated.

Partial feedback exists and deserves acknowledgement - adaptive platform trials revise allocations mid-study, and preregistered protocols sometimes state falsifiable mechanistic predictions - but the limitation is structural rather than technical: revision happens only within a trial's prespecified rules and only at the clinical stage, never reaching the platforms that made the original commitment. Post-market subgroup findings are the same, since a differential response by genotype is evidence about the causal claim made at the outset and does not flow upstream.

Two features matter more than any individual stage. The failure is directional: information moves forward efficiently and backward almost not at all. And the barriers differ in kind - the ends of the development arc are separated by infrastructure, the middle by architecture, the clinical stage by the absence of a shared representation in which human and system commitments can be compared. No single intervention closes the loop. The most tractable opening is the last, because the evidence is already generated and only the connecting representation is missing.

How to cite this article:
Coletti S (November 03, 2026) The Discovery Gap: Why Artificial Intelligence Must Learn to Change Its Mind "*Rethinking Human–AI Collaboration from Molecule to Market*". Cureus J AI Augment Res 1 : es44568-026-00351-7. DOI https://doi.org/10.7759/s44568-026-00351-7

Testing for revision, and reporting it

The more demanding criterion for scientific AI is not what a system knows but what it does when it turns out to be wrong. That furnishes an evaluation distinct in kind from any predictive benchmark: present a system holding an explicit set of commitments with an observation contradicting one of them, and measure what happens to the set. Table *3* sets out the five quantities. They decompose what is usually reported as one competence, since a system can localize well and propagate not at all, or propagate aggressively and fail on stability.

Dimension	Definition	Operationalization	Characteristic failure mode
Localization	Is the minimal contradicted subset identified?	Match the system's attributed commitment against expert-attributed ground truth	Registers the inconsistency but not its source
Minimality	Does revision extend only as far as consistency requires?	Count belief changes relative to the minimum set consistent with the new evidence	Retreat: discards large knowledge regions rather than revising selectively
Propagation	Do dependent claims update when their support is retracted?	Query downstream claims after upstream retraction; verify that the expected changes occurred.	Downstream commitments persist after upstream revision
Stability	Do claims outside the affected dependency cone remain untouched?	Query beliefs structurally unrelated to the revised commitment; verify that no spurious changes occur.	Cascades into unrelated knowledge regions (catastrophic forgetting seen from outside)
Space expansion	Can the system propose a hypothesis it did not previously hold when no existing assignment accommodates the contradiction?	Present a contradiction outside the current hypothesis space; classify the response as abductive or as best-fit confabulation	Forces existing hypotheses to fit; conflates hallucination with abduction

TABLE 3: Dimensions of the revision test

A retrospective version is buildable now, since drug development has produced many expensive refutations with documented attribution, each supplying a candidate instance: a commitment set, a refuting observation, an attribution. The BACE1 inhibitor programs in Alzheimer's disease are the obvious example, where verubecestat failed in both mild-to-moderate and prodromal disease despite target engagement and reduced amyloid production [40,41]. That case is instructive because its attribution is contested: the failures were widely read as evidence against the amyloid hypothesis itself, yet lecanemab and donanemab later produced clinical benefit through amyloid removal [42,43], which narrows the defensible attribution to BACE1 inhibition at that disease stage rather than amyloid biology as such. A benchmark instance must therefore record the attribution the field made, the attribution later evidence supported, and the interval between them.

How to cite this article:
Coletti S (November 03, 2026) The Discovery Gap: Why Artificial Intelligence Must Learn to Change Its Mind "*Rethinking Human–AI Collaboration from Molecule to Market*". Cureus J AI Augment Res 1 : es44568-026-00351-7. DOI https://doi.org/10.7759/s44568-026-00351-7

That example also shows why the retrospective form is only a starting point. Its attribution is contested and available only in hindsight, so the ground truth is expert consensus rather than fact; and the canonical failures are almost certainly in the training corpus of any current model, so a system could score well by retrieving a settled narrative instead of revising a live commitment. Retrospective cases test recall of an agreed story and belong to calibration rather than evaluation.

The evaluation that matters is prospective, and the same five dimensions support it: instantiate a commitment set from a program whose readout has not occurred, record the system's dependencies and refutation conditions in advance, and score the revision when the result arrives. Held-out variants - perturbed entity identities, or unpublished attributions from an industrial partner - recover much of the signal more cheaply. The claim that a system revises must rest on evidence it has not seen.

One caution is essential. Systems with long context windows will appear to pass a naive version of this test by updating their position fluently. That is conversational accommodation, not belief revision, and the two are separable: does the revision persist once the contradicting text leaves the context, and does weak evidence produce proportionately small revision? Genuine revision is hard to induce with weak evidence and hard to reverse with strong evidence; current systems show the reverse.

Explicit representation also changes what explainability means: it ceases to be a post-hoc approximation of an opaque function [44] and becomes a trace in which every conclusion follows back to its evidence. The regulatory implication is direct. Current guidance certifies a model for a stated context of use [45,46], and the FDA draft guidance places drug discovery outside its scope, confining itself to models producing data that support decisions on safety, effectiveness or quality [45]. That boundary is defensible where a model's output is a hypothesis tested independently downstream, but hard to hold for a system whose beliefs change as evidence accumulates. What requires certification is then the process of revision: what records are kept, what audit trail exists, who is accountable.

The same logic yields something immediately actionable. Claims of AI-assisted discovery are reported in terms of what a system found; Table *4* sets out what should be reported instead. None of it requires new technology, and together it distinguishes a system that discovered something from one that produced an output which happened to be correct - a distinction existing frameworks, which cover model validation rather than discovery claims [47], cannot express.

Item	What must be reported
Hypothesis	The claim in play, stated precisely enough to be contradicted
Evidence	The observations that changed belief in it, with retrievable provenance
Alternatives	Competing hypotheses considered, and the grounds for excluding them
Belief change	The system's commitments after the evidence arrived, and how they differed from before
Human decisions	Which judgements a person made, at which points, on what basis

TABLE 4: Minimum reporting standard for a claim of AI-assisted discovery

How to cite this article:
Coletti S (November 03, 2026) The Discovery Gap: Why Artificial Intelligence Must Learn to Change Its Mind "*Rethinking Human–AI Collaboration from Molecule to Market*". Cureus J AI Augment Res 1 : es44568-026-00351-7. DOI https://doi.org/10.7759/s44568-026-00351-7

Conclusions

Realizing DSI requires architectures combining neural breadth with explicit symbolic structure, provenance standards spanning heterogeneous sources, and revision benchmarks of the kind proposed above. It equally requires institutional change: incentives that reward knowledge revision, and governance that certifies processes rather than artefacts. The infrastructural half is more neglected and, near term, more tractable. A registry of refuted mechanistic hypotheses linked to the evidence that refuted them would be a substantial contribution requiring no advance in machine learning.

The argument is built on biomedicine, and I scope its claims accordingly. Whether the same pathology holds elsewhere is a conjecture rather than a result: closed-loop autonomous platforms in materials science [48] are suggestive, since their results revise process parameters but not the motivating hypothesis, but one example does not establish a pattern. The five dimensions are stated in terms of commitments, dependencies and evidence rather than biology, so they should transfer - a prediction whose cheapest test is a second domain.

Explicit provenance also carries risks: a system holding auditable beliefs could be used to curate evidence toward a predetermined conclusion, or to manufacture consensus by controlling provenance records. Transparency makes manipulation visible but does not prevent it - governance must address who holds write access to those records, under what oversight, and what counts as a material revision requiring disclosure.

The destination is not an autonomous machine that discovers without scientists; the three decision points above ensure it cannot be. It is a scientific substrate that learns alongside its users, holding their commitments in a form evidence can reach and revise. What the next generation of scientific AI should be built to do is not to know more, but to change its mind well.

Additional Information

Author Contributions

All authors have reviewed the final version to be published and agreed to be accountable for all aspects of the work.

Concept and design: Silvano Coletti

Acquisition, analysis, or interpretation of data: Silvano Coletti

Drafting of the manuscript: Silvano Coletti

Critical review of the manuscript for important intellectual content: Silvano Coletti

Disclosures

Conflicts of interest: In compliance with the ICMJE uniform disclosure form, all authors declare the following:

- **Payment/services info:** All authors have declared that no financial support was received from any organization for the submitted work.
- **Financial relationships:** All authors have declared that they have no financial relationships at present or within the previous three years with any organizations that might have an interest in the submitted work.
- **Other relationships:** Dr. Silvano Coletti serves as a member of the Advisory Board of the Cureus Journal of AI-Augmented Research. To maintain editorial independence and avoid any potential conflict of interest, Dr. Coletti was not involved in the editorial assessment, peer review process, or publication decision for this manuscript.

How to cite this article:
Coletti S (November 03, 2026) The Discovery Gap: Why Artificial Intelligence Must Learn to Change Its Mind "*Rethinking Human–AI Collaboration from Molecule to Market*". Cureus J AI Augment Res 1 : es44568-026-00351-7. DOI https://doi.org/10.7759/s44568-026-00351-7

AI Disclosure Statement

The author used AI-assisted tools exclusively for language editing and improving the clarity of the manuscript. The AI tool was not used to generate the scientific content. All AI-assisted outputs were reviewed and verified by the authors, who take full responsibility for the accuracy, integrity and originality of the content.

Data Availability Statements

Data sharing is not applicable. This work is theoretical; no datasets were generated or analyzed.

References

1. Jumper J, Evans R, Pritzel A, et al.: Highly accurate protein structure prediction with AlphaFold. Nature. 2021, 596:583-589. 10.1038/s41586-021-03819-2
2. The Nobel Prize in Chemistry 2024. (2024). Accessed: November 2, 2026: https://www.nobelprize.org/prizes/chemistry/2024/summary/..
3. Zhavoronkov A, Ivanenkov YA, Aliper A, et al.: Deep learning enables rapid identification of potent DDR1 kinase inhibitors. Nature Biotechnology. 2019, 37:1038-1040. 10.1038/s41587-019-0224-x
4. Wong F, Zheng EJ, Valeri JA, et al.: Discovery of a structural class of antibiotics with explainable deep learning. Nature. 2024, 626:177-185. 10.1038/s41586-023-06887-8
5. Boiko DA, MacKnight R, Kline B, Gomes G: Autonomous chemical research with large language models. Nature. 2023, 624:570-578. 10.1038/s41586-023-06792-0
6. Lu C, Lu C, Lange RT, et al.: Towards end-to-end automation of AI research. Nature. 2026, 651:914-919. 10.1038/s41586-026-10265-5
7. Jayatunga MKP, Ayers M, Bruens L, Jayanth D, Meier C: How successful are AI-discovered drugs in clinical trials? A first analysis and emerging lessons. Drug Discovery Today. 2024, 29:104009. 10.1016/j.drudis.2024.104009
8. Hay M, Thomas DW, Craighead JL, Economides C, Rosenthal J: Clinical development success rates for investigational drugs. Nature Biotechnology. 2014, 32:40-51. 10.1038/nbt.2786
9. Xu Z, Ren F, Wang P, et al.: A generative AI-discovered TNIK inhibitor for idiopathic pulmonary fibrosis: a randomized phase 2a trial. Nature Medicine. 2025, 31:2602-2610. 10.1038/s41591-025-03743-2
10. Study evaluating rentosertib (INS018_055) administered orally in patients with idiopathic pulmonary fibrosis (IPF). (2026). Accessed: October 13, 2026: https://clinicaltrials.gov/study/NCT07687459.
11. Meng K, Bau D, Andonian A, Belinkov Y: Locating and editing factual associations in GPT. Advances in Neural Information Processing Systems 35. 2022, 35:17359-17372. 10.52202/068431-1262
12. Meng K, Sharma AS, Andonian A, Belinkov Y, Bau D: Mass-editing memory in a transformer. Computation and Language. 2022, arXiv:2210.07229. 10.48550/arXiv.2210.07229
13. Yao Y, Wang P, Tian B, et al.: Editing Large Language Models: Problems, Methods, and Opportunities. Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing (EMNLP 2023). Association for Computational Linguistics, Singapore; 2023. 10222-10240. 10.18653/v1/2023.emnlp-main.632
14. Kirkpatrick J, Pascanu R, Rabinowitz N, et al.: Overcoming catastrophic forgetting in neural networks. Proceedings of the National Academy of Sciences of the United States of America. 2017, 114:3521-3526. 10.1073/pnas.1611835114
15. DeVito NJ, Bacon S, Goldacre B: Compliance with legal requirement to report clinical trial results on ClinicalTrials.gov: a cohort study. Lancet. 2020, 395:361-369. 10.1016/s0140-6736(19)33220-9
16. King RD, Whelan KE, Jones FM, et al.: Functional genomic hypothesis generation and experimentation by a robot scientist. Nature. 2004, 427:247-252. 10.1038/nature02236
17. King RD, Rowland J, Oliver SG, et al.: The automation of science. Science. 2009, 324:85-89. 10.1126/science.1165620
18. Kitano H: Nobel Turing Challenge: creating the engine for scientific discovery. npj Systems Biology and Applications. 2021, 7:29. 10.1038/s41540-021-00189-3

How to cite this article:
Coletti S (November 03, 2026) The Discovery Gap: Why Artificial Intelligence Must Learn to Change Its Mind "*Rethinking Human–AI Collaboration from Molecule to Market*". Cureus J AI Augment Res 1 : es44568-026-00351-7. DOI https://doi.org/10.7759/s44568-026-00351-7

19. Alchourrón CE, Gärdenfors P, Makinson D: On the logic of theory change: Partial meet contraction and revision functions. The Journal of Symbolic Logic. 1985, 50:510-530. 10.2307/2274239
20. de Kleer J: An assumption-based TMS. Artificial Intelligence. 1986, 28:127-162. 10.1016/0004-3702(86)90080-9
21. Rose D, Langley P: Chemical discovery as belief revision. Machine Learning. 1986, 1:423-452. 10.1007/bf00114870
22. Rose D: Discovery and revision via incremental hill climbing. Knowledge Representation and Organization in Machine Learning. Brazdil PB, Konolige K (ed): Springer, Boston, MA; 1990. 187-206. 10.1007/978-1-4613-1641-1_9
23. Rainforth T, Foster A, Ivanova DR, Bickford Smith F: Modern Bayesian experimental design. Statistical Science. 2024, 39:100-114. 10.1214/23-sts915
24. Mayo DG: Statistical Inference as Severe Testing: How to Get Beyond the Statistics Wars. Cambridge University Press, Cambridge, UK; 2018. 10.1017/9781107286184
25. Gil Y, Garijo D, Ratnakar V, et al.: Towards continuous scientific data analysis and hypothesis evolution. Proceedings of the AAAI Conference on Artificial Intelligence. 2017, 31:4406-4414. 10.1609/aaai.v31i1.11157
26. Takahara I, Mizoguchi T: Toward auditable AI scientists: a hypothesis evolution protocol for LLM agents. Artificial Intelligence. 2026, arXiv:2607.09195. 10.48550/arXiv.2607.09195
27. Ghafarollahi A, Buehler MJ: SciAgents: automating scientific discovery through bioinspired multi-agent intelligent graph reasoning. Advanced Materials. 2025, 37:2413523. 10.1002/adma.202413523
28. Luo Z, Kasirzadeh A, Shah NB: The more you automate, the less you see: hidden pitfalls of AI scientist systems. Artificial Intelligence. 2025, arXiv:2509.08713. 10.48550/arXiv.2509.08713
29. Abdar M, Pourpanah F, Hussain S, et al.: A review of uncertainty quantification in deep learning: Techniques, applications and challenges. Information Fusion. 2021, 76:243-297. 10.1016/j.inffus.2021.05.008
30. Schölkopf B, Locatello F, Bauer S, Ke N, Kalchbrenner N, Goyal A, Bengio Y: Toward causal representation learning. Proceedings of the IEEE. 2021, 109:612-634. 10.1109/jproc.2021.3058954
31. Hägele A, Rothfuss J, Lorch L, Somnath VR, Schölkopf B, Krause A: BaCaDI: Bayesian causal discovery with unknown interventions. Proceedings of the 26th International Conference on Artificial Intelligence and Statistics. Proceedings of Machine Learning Research, 2023. 206:1411-1436.
32. Acosta JN, Falcone GJ, Rajpurkar P, Topol EJ: Multimodal biomedical AI. Nature Medicine. 2022, 28:1773-1784. 10.1038/s41591-022-01981-2
33. Gao S, Fang A, Huang Y, et al.: Empowering biomedical discovery with AI agents. Cell. 2024, 187:6125-6151. 10.1016/j.cell.2024.09.022
34. Hao Q, Xu F, Li Y, Evans JA: Artificial intelligence tools expand scientists' impact but contract science's focus. Nature. 2026, 649:1237-1243. 10.1038/s41586-025-09922-y
35. Quine WV: Two dogmas of empiricism. The Philosophical Review. 1951, 60:20-43. 10.2307/2181906
36. Thaler v. Vidal, 43 F.4th 1207 (Fed. Cir. 2022), cert. denied (U.S. Apr. 24, 2023). (1207). Accessed: August 6, 2026: https://www.wipo.int/wipolex/en/judgments/details/2098.
37. Thaler v Comptroller-General of Patents, Designs and Trade Marks [2023] UKSC 49. (2023). Accessed: August 6, 2026: https://www.supremecourt.uk/cases/uksc-2021-0201.
38. Vaccaro M, Almaatouq A, Malone T: When combinations of humans and AI are useful: A systematic review and meta-analysis. Nature Human Behaviour. 2024, 8:2293-2303. 10.1038/s41562-024-02024-1
39. Zhang K, Yang X, Wang Y, et al.: Artificial intelligence in drug development. Nature Medicine. 2025, 31:45-59. 10.1038/s41591-024-03434-4
40. Egan MF, Kost J, Tariot PN, et al.: Randomized trial of verubecestat for mild-to-moderate Alzheimer's disease. The New England Journal of Medicine. 2018, 378:1691-1703. 10.1056/nejmoa1706441
41. Egan MF, Kost J, Voss T, et al.: Randomized trial of verubecestat for prodromal Alzheimer's disease. The New England Journal of Medicine. 2019, 380:1408-1420. 10.1056/nejmoa1812840

How to cite this article:
Coletti S (November 03, 2026) The Discovery Gap: Why Artificial Intelligence Must Learn to Change Its Mind "*Rethinking Human–AI Collaboration from Molecule to Market*". Cureus J AI Augment Res 1 : es44568-026-00351-7. DOI https://doi.org/10.7759/s44568-026-00351-7

42. van Dyck CH, Swanson CJ, Aisen P, et al.: Lecanemab in early Alzheimer's disease. The New England Journal of Medicine. 2023, 388:9-21. 10.1056/nejmoa2212948

43. Sims JR, Zimmer JA, Evans CD, et al.: Donanemab in early symptomatic Alzheimer disease: the TRAILBLAZER-ALZ 2 randomized clinical trial. JAMA. 2023, 330:512-527. 10.1001/jama.2023.13239

44. Ghassemi M, Oakden-Rayner L, Beam AL: The false hope of current approaches to explainable artificial intelligence in health care. The Lancet Digital Health. 2021, 3:e745-e750. 10.1016/s2589-7500(21)00208-9

45. Considerations for the use of artificial intelligence to support regulatory decision-making for drug and biological products. Draft guidance for industry and other interested parties; Docket No. FDA-2024-D-4689. (2025). Accessed: June 1, 2025: https://www.fda.gov/regulatory-information/search-fda-guidance-documents/considerations-use-artificial-intelligence-s....

46. European Medicines Agency. Reflection paper on the use of artificial intelligence (AI) in the medicinal product lifecycle. (2024). Accessed: July 29, 2026: https://www.ema.europa.eu/system/files/documents/scientific-guideline/reflection-paper-use-artificial-intelligence-ai....

47. Collins GS, Moons KGM, Dhiman P, et al.: TRIPOD+AI statement: updated guidance for reporting clinical prediction models that use regression or machine learning methods. BMJ. 2024, 385:e078378. 10.1136/bmj-2023-078378

48. Nikolaev P, Hooper D, Webber F, et al.: Autonomy in materials research: a case study in carbon nanotube growth. npj Computational Materials. 2016, 2:16031. 10.1038/npjcompumats.2016.31

How to cite this article:
Coletti S (November 03, 2026) The Discovery Gap: Why Artificial Intelligence Must Learn to Change Its Mind "*Rethinking Human–AI Collaboration from Molecule to Market*". Cureus J AI Augment Res 1 : es44568-026-00351-7. DOI https://doi.org/10.7759/s44568-026-00351-7