Digital Discovery

ROYAL SOCIETY OF CHEMISTRY

OPINION

View Article Online
View Journal

Cite this: DOI: 10.1039/d6dd00030d

Scientists might be reaffirming the relevance of human oversight as AI lands in labs

Renan Gonçalves Leonel da Silva, [abc] Li Du [de] and Gil Eyal [*cf]

Artificial Intelligence (AI) is prompting scientists to reflect on the shifting role of human judgment, interpretation, and oversight in experimental practice. As AI increasingly assumes critical roles in scientific discovery, innovation, and academic labor, new paradoxes are emerging around the question of keeping humans in the loop. These paradoxes are not simply about whether humans should remain present, but about how they can remain meaningfully engaged with increasingly opaque AI-driven discovery systems. In this opinion, we examine how the promises of AI-augmented research infrastructures coexist with difficult questions about how to engage with automated and intelligent apparatuses without eroding human oversight, core scientific values such as safety and responsibility, or the broader societal relevance of scientists.

Received 21st January 2026
Accepted 7th June 2026

DOI: 10.1039/d6dd00030d

rsc.li/digitaldiscovery

Introduction

Lisanne Bainbridge's seminal 1983 article Ironies of Automation showed that automation often produces counter-intuitive outcomes: systems designed to remove humans from tasks can instead make human involvement more critical, precisely because operators are left with only monitoring rare and high-stakes failures, and lose the skills needed to intervene effectively when things go wrong. What seemed like a reduction of human labor actually shifted human roles toward complex oversight, situational assessment, and crisis management – roles that automation itself was supposed to eliminate.[1]

Today, similar ironies are resurfacing in the context of Digital Discovery: an emerging paradigm transforming fields such as drug discovery, materials science, and catalysis, shortening discovery timelines from years to months or even weeks.[2,3] This field refers to the use of computational tools, artificial intelligence, and data-driven approaches to accelerate the identification and development of new materials, molecules, and chemical processes. Rather than relying solely on traditional trial-and-error experimentation, digital discovery integrates machine learning, high-throughput simulations, and automated laboratories to navigate vast chemical spaces more efficiently. It bridges the gap between computational prediction and experimental validation, enabling researchers to prioritize the most promising candidates before committing laboratory resources. Digital discovery does not replace human expertise but augments it, allowing scientists to focus on higher-order interpretation and decision-making.[4]

But advanced models and automated experimentation platforms promise unprecedented speed and insight, yet they raise fundamental questions about whether humans can, or should, be sidelined from core decision points, interpretation, and judgment. As in Bainbridge's analysis, removing humans from routine tasks can inadvertently make their remaining involvement both more crucial and more difficult, especially in domains where tacit knowledge, creativity, and ethical reasoning are essential.

Similarly, Endsley (2023)[5] extends Bainbridge's insight to contemporary AI systems, showing that AI's cognitive focus produces its own set of paradoxes: the more capable and adaptive AI becomes, the harder it is for humans to understand its behavior, limitations, and biases, even as humans are expected to monitor and intervene when necessary. Endsley identifies how opaqueness and over-reliance on AI can erode human situational awareness and decision capacity, paradoxically making human oversight both more essential and more difficult (Fig. 1 illustrates this paradox, highlighting the collaborative effort of a research team overseeing an AI system).

These "ironies" foreground why debates about keeping humans in the loop are not simply about whether humans should be present, but how humans can remain meaningfully engaged with increasingly inscrutable AI-driven discovery systems. Recently, the A-Lab episode at Lawrence Berkeley National Laboratory offers a striking contemporary illustration of this tension.

[a]Department of Sociology, University of São Paulo, São Paulo, SP, Brazil
[b]Erna D. and Henry J. Leir Research Institute for Business, Technology and Society, Martin Tuchman School of Management, New Jersey Institute of Technology, Newark, NJ, USA
[c]Trust Collaboratory, INCITE, Columbia University, New York, NY, USA
[d]Faculty of Law, University of Macau, Macau, China
[e]Asia-Pacific Academy of Economics and Management, University of Macau, Macau, China
[f]Department of Sociology, Columbia University, New York, NY, USA. E-mail: ge2027@columbia.edu

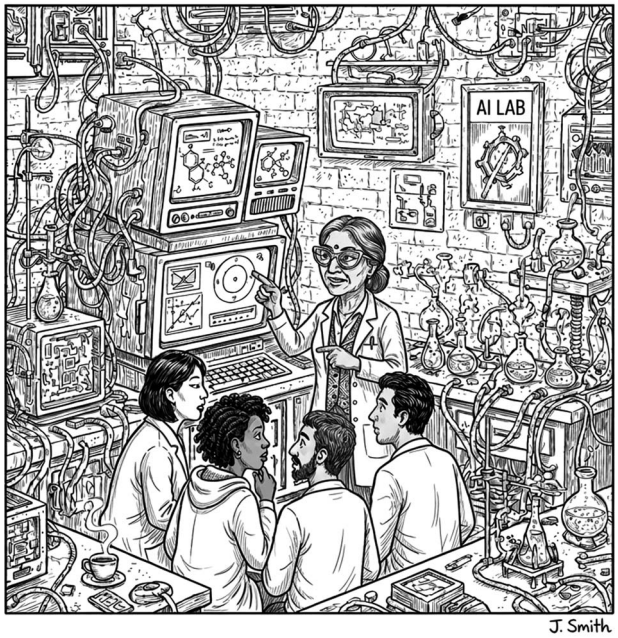

Fig. 1 The AI 'Black Box' paradox, visualized. As system opaqueness grows, human situational awareness erodes—leaving operators to manage a system they can no longer interrogate.

The A-Lab episode

In 2023, the scientific community was captivated by the A-Lab project at Lawrence Berkeley National Laboratory, powered by Google DeepMind. The project claimed to have synthesized 41 new inorganic compounds in under three weeks, publicizing a 71% success rate as proof of AI's transformative potential in materials discovery.[6] The A-Lab controversy gained significant public and scientific traction following a series of viral threads on X (formerly Twitter) by Robert Palgrave (Professor of Inorganic and Materials Chemistry at University College London, UK). The reanalysis done by Robert Palgrave and Leslie Schoop, however, revealed deep flaws: none of the compounds were truly novel, and many of the "discoveries" stemmed from reinterpretations or analytical errors.[7] The case rapidly went viral, providing material for blogs and webpages dedicated to chemistry, materials sciences, and emerging trends in digital discovery. The episode ultimately culminated in a formal correction published by *Nature*, in which the authors officially revised the status of the original paper and transparently acknowledged its limitations.[7]

Far from merely a technical mistake, the episode exposed urgent issues around oversight, transparency, and scientific responsibility in an era of autonomous experimentation. The event catalyzed a broader debate on transparency and the 'black box' nature of automated discovery. It additionally underscores the evolving role of social media as a space for real-time peer review and highlights the urgent need for institutionalized standards of responsibility in AI laboratories.[8]

This controversy marks an important inflection point. From one side, automated experiments, self-learning systems, and large data-driven models promise speed and scale previously inconceivable.[9] On the flip side, it is important to recognize the unquestionable success achieved by several tools developed for that purpose. While the A-Lab episode serves as a cautionary tale, successful AI-human collaborations demonstrate the technology's constructive potential. When AI is used to navigate vast chemical spaces while humans maintain oversight of the experimental validation, it can significantly accelerate the discovery of robust, reproducible materials.

One compelling example is the work of Szymanski *et al.* (2023),[2] who demonstrated the power of AI-human collaboration in accelerating the discovery of novel battery materials. Using an autonomous laboratory platform integrated with machine learning-guided synthesis and human expert validation, the team identified and experimentally confirmed several previously unreported inorganic compounds within a fraction of the time conventional approaches would require. Crucially, human researchers remained embedded in the validation loop, ensuring that AI-proposed candidates were subjected to rigorous experimental scrutiny before claims of discovery were advanced—a workflow that produced both speed and credibility.

A second notable case is the BELLA platform developed by Burger *et al.* (2020),[10] in which a mobile robotic chemist autonomously performed thousands of experiments to optimize the photocatalytic activity of organic semiconductor materials. The system operated continuously, navigating a large experimental parameter space far beyond what a human team could feasibly explore manually, while researchers defined the boundaries, interpreted emergent trends, and guided strategic pivots in the investigation. The resulting discoveries were independently reproducible and experimentally well-characterized, illustrating that when AI autonomy is paired with clearly defined human oversight structures, SDLs can deliver on their promise of accelerating robust and trustworthy scientific knowledge.

As algorithmic cultures take deeper root in labs, a pressing question emerges: how can the core values that have fostered scientific success survive when algorithms mediate experiment, interpretation, and validation?

Setting algorithmic cultures

These transformations reflect the maturation of computational and digital rationales in scientific practices, norms, and infrastructures. We term this phenomenon the co-production of algorithmic cultures: a broad social, technical, and ethical reconfiguration of laboratory and academic practices. Algorithmic cultures are structured by a resilient but non-linear process of encoding quantificational, probabilistic, and binary rationales into hypothesis development and experimental platforms. This mechanism affects the content and direction of values and regimes organizing contemporary scientific knowledge production and validation.

As these cultures deepen, they compel scientists to reexamine the human values that supported science's trajectory for

decades. While computing has long played a role in labs, the moment has come to shift from focusing purely on change to asking which institutional norms, values, and principles made breakthrough science possible in the first place. Among these dimensions, one stands out: increasing skepticism among scientists about AI's role in research.

Growing skepticism among scientists

A striking paradox emerges from recent data: AI use is rising even as trust in it is falling. The Wiley 2025 State of the Researcher Report finds that AI adoption among researchers leapt from 45% in 2024 to 62% in 2025, yet the percentage of scientists worried about hallucinations increased from 51% in 2024 to 64% in 2025. AI hallucinations refer to instances where an artificial intelligence system generates information that sounds confident and plausible but is factually incorrect, fabricated, or entirely disconnected from reality. This occurs because AI language models do not truly "know" facts the way humans do; instead, they predict what words and sentences are likely to follow based on patterns learned from vast amounts of text, and sometimes those predictions lead them astray in ways that are difficult to detect without independent verification.[11] Concerns over security, privacy, ethics, and transparency also climbed.[12]

Similar results have been widely published from different surveys and opinion panels administered to scientists and academic researchers between 2023 and 2025. Asked about their attitudes toward AI in their labs and academic work as a whole, scientists surveyed express concern and increasing levels of skepticism with AI tools in scientific research (especially in highly innovative domains of scientific discovery) (see Tables 1 and 2).

This shift represents less a rejection than a maturing and pragmatic engagement. In the absence of clear external guidelines, scientists' growing caution functions as an essential, ad-hoc form of risk management. As scientists gain direct experience with AI systems, they confront not just technical limitations, but certain value dilemmas provoked by integrating these systems into the process of scientific research. When researchers are confronted with AI hallucinations, namely with models confidently asserting falsehoods, they are reminded that oversight, interpretive judgment, and ethical sensibility cannot be automated away, and that caution, humility, and accountability must remain human anchors when machines err or mislead.

Thus, sustained engagement with AI as a research aid acts as a corrective to hype. Early enthusiasm assumed that more compute and data would reliably yield better models, but when practitioners observe hallucinations and opaque failures, they begin to reassert human values in the design, deployment, and oversight of these systems. The result is not rejection but recalibration: scientists emphasize not only what AI can do but how it aligns with human values and scientific norms. The Wiley survey's paradox (*i.e.*, less trust among more experienced users, Table 1) can be read as a turning point.

Persistent AI hallucinations are indeed among the key reasons why cautious behavior is adopted by senior researchers using AI for scientific purposes. In the field of chemistry, AI hallucinations take the specific and problematic form of high-stakes failure modes such as hallucinated reactivity, stoichiometry violations, and flawed stereochemical reasoning. These errors highlight exactly where human expertise—grounded in physical laws—remains a critical corrective to probabilistic models. This problem should remind us also that the progress of science depends not only on "organized skepticism," on also on trust in the integrity and expertise of other scientists.

A matter of trust

Science has historically rested on trust: confidence in peer methodology, interpretive rigor, and transparent communication. Rooted in peer review, replication, and critique, the self-correcting nature of science is undergirded by shared norms of conduct.[13,14]

The introduction of AI to scientific research, however, subjects these traditions to a stress test. Algorithms increasingly perform tasks once exclusive to skilled scientists: data analysis, experiment recommendation, hypothesis generation. In computational chemistry and materials science, automated screening, complex machine learning models, and laboratory automation are being deployed to simulate molecular behaviors, predict properties, and propose experimental paths. But speed brings distance: from experimentation, from oversight, and sometimes from the trust that grounds scientific legitimacy.

The A-Lab case illustrates this tension. AI systems often "black boxes" resistant to inspection. Scientific practice demands more than fleeting success rates: it demands interpretability, reproducibility, and clarity.[8,15] When the workings of a model are inscrutable, trust becomes brittle, and the epistemic foundation of science is threatened – with broader implications to academic research and public reputation of scientists beyond the lab's walls.[16]

AI overreliance

AI brings unique tensions into scientific reasoning. Psychological research on algorithm appreciation reveals a human tendency to overtrust machine-generated recommendations, even when human judgment might be superior.[17] AI's aura of infallibility can overshadow logical scrutiny. Meanwhile, suspicion toward opaque systems may erode confidence in discoveries, even when they are robust.[8]

Risks of AI overreliance have been recently highlighted in a preprint titled "The White Elephant in the Lab".[18] In this work, researchers active in the field of digital molecular discovery raise critical concerns about the limitations of generative models. They argue that, regardless of a model's sophistication, if the molecules it proposes cannot be synthesized, its practical value in laboratory settings is severely limited. That's why researchers have been working hard on improving those systems to allow scientists and engineers to engage with those tools as they are under development, testing and prototyping – ultimately guaranteeing proper human oversight of multiple steps of the experimentation process. In other words, meaningful engagement requires scientists to utilize computational tools that facilitate human-in-the-loop validation. For instance,

Table 1 Key findings of surveys and panels on scientists' attitudes toward AI in research labs

Year	Survey/panel	Population/scope	Key findings (attitudes toward AI in research/labs)	Trend highlights (skepticism & pragmatism)
2023	Artificial intelligence survey (SciOPS panel)	U.S. academic scientists ($n \approx 777$; valid responses 232)	Early descriptive data on perceptions and use of generative AI tools for teaching/research tasks; highlights varied comfort and uptake. (SciOPS)	Captures baseline mixed attitudes; a segment remains hesitant or non-users
2023	Nature & related research reporting on ~1600 scientists	Global researchers (nature survey reports)	~30% of surveyed scientists reported using GenAI for writing, literature reviews, and grant tasks; concern about ethical dimensions noted. (Springer)	Reflects early ethical qualms despite adoption; indicates conditions on acceptable use
2024	Generative AI usage by researchers (arXiv, Dorta-González *et al.*)	Broad researcher sample drawn from various workplaces	Examines how demographics (gender, career stage) and barriers influence AI uptake in research workflows. (arXiv)	Highlights structural and personal barriers; not simply positive uptake
2024	Survey on GenAI in Danish universities	Danish researchers ($n \approx$ 2534)	Detailed mapping of GenAI tool use across research phases; varied views on research integrity implications. (ScienceDirect)	Indicates nuanced views: accepted for some tasks, controversial for rigorous research stages
2025	Generative AI and academic scientists in US universities (PLoS One/National survey)	U.S. academic scientists ($n =$ 232)	65% used GenAI for teaching/research; 78% cited misinformation concerns, many want institutional/governance safeguards. (PLoS)	Strong evidence of adoption with heightened caution, especially about reliability and ethical governance
2025	Social Scientists on the Role of AI in Research (arXiv)	Social science researchers ($n =$ 284 + interviews)	Increased use of AI tools but greater ethical and trust concerns (*e.g.*, black-box systems, deskilling) compared to traditional ML (arXiv)	Shows field-specific skepticism toward less transparent AI methods
2025	Elsevier's global "Researcher of the Future" survey (3000 researchers)	International researchers	58% use AI in research; but only 27% feel adequately trained and only ~23% trust AI ethics, with distinct regional skepticism. (https://www.elsevier.com)	Highlights broader concerns about governance, trust, and proper training—key markers of pragmatic attitudes

platforms like AIZynthFinder or ASKCOS provide retrosynthetic route predictions that allow human experts to vet the feasibility of AI-generated molecules, transforming the AI from a 'black box' into a collaborative partner. Despite the successful deployment of such platforms and generative models supporting new tools designed to increasingly automate decision-making in molecular design, there is a tendency to privilege algorithmic output over empirical validation and expert judgment. It underscores how excessive trust in AI-driven predictions can obscure fundamental chemical constraints and marginalize human expertise. This overreliance is often unintentionally encouraged by a policy landscape that rewards AI-driven productivity while underfunding the meticulous, time-consuming work of experimental validation and ethical scrutiny. The risk of such funding bias is that it will accelerate a process of deskilling observed also in other expert domains when AI systems are integrated.

In the context of design, prototyping and deployment of self-driving labs (SDLs), deskilling refers to the gradual erosion of hands-on experimental expertise among researchers and scientists, especially those at the beginning of their careers, as increasingly automated systems take over the physical and cognitive tasks traditionally performed by humans in the laboratory. As robotic platforms, AI-driven decision-making, and automated workflows handle more of the experimental process—sample preparation, instrument operation, data collection, parameter optimization—experienced scientists may still draw on direct, tacit knowledge built through years of manual experimentation, but young scientists, coming of age into a world of SDLs, may not be able to develop these skills. They will lack the tacit knowledge of instrument behavior (*e.g.*, how a pipette "feels" when something is off), troubleshooting intuition, contextual judgment about when an automated result should be trusted or questioned, and physical intuition about

Table 2 Patterns across surveys and panels (2023–2025)

Major pattern	Description
Rapid adoption coupled with uneven confidence	Across multiple surveys, a majority of researchers report using AI for research-related activities (*e.g.*, writing, data analysis, literature review). At the same time, many express unease about validity, reliability, and oversight. For instance, in the 2025 *PLoS One* survey, while 65% reported using generative AI, 78% identified misinformation as a primary concern, illustrating adoption without full trust
Ethical and epistemic concerns are prominent	Surveys and qualitative studies—particularly among social scientists—highlight ethical and epistemic worries, including automation bias, deskilling, opacity of black-box models, and challenges to scientific accountability. These concerns are often sharper than those associated with earlier statistical or rule-based tools and motivate calls for stronger governance, transparency, and critical human mediation in laboratories
Divergent attitudes by career stage, discipline, and region	Attitudes toward AI are heterogeneous rather than uniform. Demographic analyses show that senior researchers, early-career scholars, and researchers with differing computational expertise adopt AI at different paces and express varying degrees of skepticism. Disciplinary cultures and regional research infrastructures further shape how AI is evaluated and trusted
Growing demand for governance and training	Across national and institutional contexts, researchers consistently report insufficient training and low confidence in existing governance frameworks. This gap contributes to pragmatic caution: skepticism is driven less by resistance to innovation and more by awareness of methodological, ethical, and organizational risks associated with unregulated AI use
Ethical conditions shape acceptable use	Large-scale surveys, including Nature's survey of more than 5000 academics, show strong support for disclosure requirements and ethical boundaries regarding AI use in research. While many accept AI assistance for drafting or exploratory tasks, consensus weakens for higher-stakes activities such as peer review, authorship attribution, or evaluative decision-making

materials, reagents, and equipment quirks. Over time, even experienced researchers will begin to lose these skills, per the adage "use it or lose it".

The threat of deskilling is acute even when AI handles seemingly routine tasks, because their expert execution often relies on strategic chemical knowledge that remains tacit. Capabilities such as convergent synthesis planning, protecting group strategies, and stereochemical control require a level of nuanced judgment and 'chemical intuition' that current algorithmic systems cannot replicate.

In this delicate balance, the danger is twofold: overreliance can embed systematic errors; deskilling can prevent these errors from being recognized until it is too late; yet excessive skepticism will likely deter innovation. In this sense, "The White Elephant in the Lab" exemplifies how the opacity and abstraction of AI systems can inadvertently erode the epistemic foundations of scientific inquiry, transforming tools meant to assist discovery into sources of uncertainty and misplaced confidence.

The central question becomes not whether to trust AI but how to integrate AI so that it complements, rather than supplants, human judgment, dissent, and revision within scientific communities.

Safety and operational boundaries

Digital discovery must also account for the physical constraints of AI in the lab. Enthusiasm for integrating AI in lab work often ignores robotic limitations and hardware tolerances, but scientists and SDLs engineers are aware of those limitations as of now. In fact, and as a globally engaged community, SDLs scientists have been sharing experiences and collaborate intensively to inform colleagues about issues they are likely to deal with, how far they can go and what tools are already available to solve such robotic limitations.

Furthermore, ensuring chemical safety in self-driving labs (SDLs) is being pursued as a top-priority because of the reasons raised in this opinion, such as the risk to scientists' reputation due to discredited or non-reproducible/non-replicable discoveries. Recent frameworks like Safe-SDL and monitoring tools like Chemist Eye establish necessary safety boundaries, reminding us that responsible AI integration is as much about physical risk mitigation as it is about data integrity.

Chemical safety in AI-driven laboratories represents a critical and increasingly well-defined dimension of responsible SDL development. Leong *et al.* (2025)[19] provide a foundational overview of safety considerations specific to self-driving laboratories, outlining strategies for steering autonomous systems toward safe operational practice – complemented by Munguia-Galeano *et al.* (2025)'s[20] Chemist Eye: a real-time safety monitoring tool designed to detect and flag hazardous conditions within SDL environments. At the hardware level, Longley *et al.* (2026)[21] present RobInHood, a robotic chemist platform engineered to operate within a fume hood, directly addressing the

containment and ventilation challenges inherent to automated chemical synthesis. Finally, Zhang *et al.* (2026)[22] propose Safe-SDL, a framework that embeds explicit safety boundaries into AI-driven experimental workflows, ensuring that autonomous decision-making does not exceed acceptable chemical or operational risk thresholds. Together, these contributions offer a multi-layered view of safety in SDLs, bridging ethical oversight with practical risk mitigation across hardware design, real-time monitoring, and algorithmic constraint.

Results outpace review

As automated processes accelerate discovery, interpretation, and publication, a question emerges: What becomes of trust among scientists? Is faith in the system alone sufficient, or must new forms of scrutiny and accountability emerge to catch subtler errors, misconduct, or misinterpretations?

In the A-Lab aftermath, integrity was restored not by better algorithms but by communal critique and expert reanalysis.[7] This outcome vindicates the self-correcting ideal of science but also underscores the need for heightened vigilance as research grows ever faster and more complex.

Trust in AI-enhanced science is fundamentally social and collective. Ethical responsibility demands more than reliable code; it requires accountability across teams, institutions, and scholarly communities. Transparent correction, peer challenge, and the willingness to contest error are virtues that have long defined responsible science, but are we doing enough to preserve them?

Reaffirming the critical role of human oversight

As algorithms assume greater experimental and analytical roles, scientists confront an existential question: If machines can design experiments, analyze results, and even propose hypotheses, what uniquely human value remains?

This question concerns not merely professional survival but the identity and purpose of science itself. The A-Lab episode offers a clear answer: human expertise remains indispensable for validating, contextualizing, and critically evaluating outputs that AI cannot fully explain or defend. Palgrave and Schoop's reanalysis depended on domain knowledge, skeptical inquiry, and nuanced interpretive judgment. Those are capacities no current AI system matches.

The challenge is not to resist algorithms wholesale but to redefine scientific expertise in a complementary relationship with them, knowing where human judgment should lead and where algorithmic power should be harnessed responsibly.

Cultivating disciplined skepticism

Progress in science has long depended on organized skepticism, *i.e.*, a norm calling for questioning claims, demanding evidence, and subjecting theories to rigorous scrutiny.[23] Algorithmic cultures threaten this virtue in multiple ways: the pace and volume of AI-generated outputs may overwhelm traditional review; model complexity may discourage challenge; and the authority of computation may suppress dissent.

The accelerated preprint ecosystem intensifies this threat. Between 2018 and 2024, submissions to ChemRxiv grew from approximately 1200 to over 9600, while ArXiv submission rates in relevant computational science fields nearly tripled. The deluge of unreviewed content both speeds knowledge exchange and magnifies vulnerabilities in peer review.[24]

Rapid dissemination poses ethical as well as technical challenges. What are the consequences of claims bypassing vetting? How do unreviewed findings affect public trust, resource distribution, or academic careers? Retractions, reputational damage, and confusion are on the rise: often heightened by AI-driven hype.[24] The ethical burden on authors and editors calls for rigor, transparency, and humility in a system favoring bold claims and rapid output proper of the science of our times.

Governance gap of AI in the sciences

The rise of algorithmic cultures in science is not occurring in a vacuum but is actively propelled by ambitious national policies. Governments worldwide, particularly in major research powers like the United States, China, and the European Union, have launched strategic initiatives, such as the U.S. National AI Initiative, China's The Artificial Intelligence+' Initiative, and the EU's Coordinated Plan on AI—that explicitly stand AI as an engine for scientific breakthrough and economic competitiveness. These policies provide substantial funding, infrastructure, and political legitimacy for AI-enabled research, creating a powerful top-down momentum for adoption.

However, this supportive wave of policy has significantly outpaced the development of corresponding ethical and legal frameworks specifically designed for AI in scientific contexts.[25] For example, the China's Interim Measures for the Management of Generative Artificial Intelligence Services is recognized as the world first rule on GenAI, its scope explicitly excludes scientific research. While general AI ethics principles (*e.g.*, fairness, transparency, accountability) are widely adopted, their translation into concrete guidelines, standards, and oversight mechanisms for laboratory practice remains limited.[26] This creates a critical governance gap: scientists and institutions are incentivized and equipped to use AI at an unprecedented scale and speed, yet are left without clear guardrails on how to use it responsibly. The A-Lab incident is a symptomatic failure of this gap; the drive to demonstrate AI's transformative potential collided with the absence of mandated protocols for algorithmic validation, transparency, and pre-publication audit.[27,28] The result is a systemic tension where the imperative for rapid innovation risks sidelining the procedural rigor and caution that have traditionally governed high-stakes discovery.[29]

Setting algorithmic cultures in the lab

With the use of AI, reconfiguring algorithmic cultures in the lab has become vital. The lesson of A-Lab is stark: science is not merely a technical enterprise. Trust matters and lives in laboratories: through relationships of critique, transparent correction, and shared values among researchers and PIs. Technology

and innovative methods must be integrated but never allowed to override human judgment or institutional safeguards.

Although there is a general lack of ethical and regulatory frameworks to regulate AI-driven scientific research, the governance gap can be closed in other ways. The affirmation of human values can be translated into concrete practical suggestions for institutional and procedural innovations. The scientific community must adopt actionable frameworks that embed ethics and oversight into the AI-driven research lifecycle. A critical first step is the adoption of algorithmic pre-registration. Mirroring the rigor of clinical trials, high-stakes AI-driven discovery pipelines should mandate the pre-registration of model architectures, training data parameters, and validation protocols.

Simultaneously, the mechanisms of scholarly validation must evolve. Journals and preprint servers must cultivate specific reviewer competencies and implement mandatory checklists for AI-involved research. These should require authors to disclose model limitations, training data biases, and provide full code accessibility. The goal is to shift review from a passive assessment of outputs to an active scrutiny of the process of algorithmic discovery, ensuring the methodology itself is sound, transparent, and ethically conducted.

Ultimately, funding agencies and research institutions have a powerful role to play by mandating "value-by-design" principles in grants for AI-enabled science. Research proposals could be required to explicitly outline how human oversight, explainability, and ongoing ethical review are structurally embedded within the experimental workflow. This approach would require that human judgment will be a built-in, governing feature of the research system itself. By implementing these three pillars, rigorous pre-approval, evolved peer critique, and value-centric funding, the scientific community can build the necessary infrastructure to ensure that AI serves as a tool that reinforces, rather than erodes, the foundational integrity of science.

These procedural innovations do not suffice on their own. They must be supported by quantitative benchmarks that bridge ethical oversight with scientific outcomes. Metrics such as solve rates, Routescore, and SPARROW offer measurable ways to assess the synthetic accessibility, cost, and labor effort of AI-driven projects, ensuring that 'efficiency' does not come at the expense of empirical reality.

The maturing of algorithmic cultures presents not only opportunity to advance scientific research but also to institutionalize the human values of skepticism, accountability, and wisdom. By embedding these values into the process of funding, publication, and validation, the scientific community will avoid AI's potential deleterious impact on trust, while cultivating its use as a tool that, when wisely governed, reinforces the enduring relevance and integrity of science itself.

Author contributions

Renan Gonçalves Leonel da Silva; conceptualization, data curation, formal analysis, investigation, methodology project administration, resources, supervision, validation, visualization, writing – original draft, writing – review & editing. Li Du; formal analysis validation, writing – review & editing. Gil Eyal; conceptualization, data curation, formal analysis, funding acquisition, methodology project administration, resources, supervision, validation, writing – review & editing.

Conflicts of interest

There are no conflicts to declare.

Data availability

No primary research results, software or code have been included and no new data were generated or analysed as part of this review.

Acknowledgements

This work was supported by the São Paulo Research Foundation (FAPESP) under Grants No. 2025/11062-4 and 2026/07471-9. The authors also express their gratitude to the editors and reviewers for their valuable comments and suggestions, which greatly improved the manuscript.

References

1 L. Bainbridge, Ironies of automation, *Automatica*, 1983, **19**(6), 775–779, DOI: 10.1016/0005-1098(83)90046-8.
2 N. J. Szymanski, B. Rendy, Y. Fei, R. Kumar, T. He, A. Milsted, M. J. McDermott, M. Gallant, E. D. Cubuk, A. Reif, C. Gogoi, J. Jopling, A. Doyle, J. Doyle, K. A. Persson and G. Ceder, An autonomous laboratory for the accelerated synthesis of novel materials, *Digital Discovery*, 2023, **2**, 1123–1128, DOI: 10.1039/d3dd00113j.
3 A. Aspuru-Guzik and K. A. Persson, Materials acceleration platform: accelerating advanced energy materials discovery by integrating high-throughput methods with artificial intelligence, *Mission Innovation*, 2018, 1–40.
4 D. P. Tabor, L. M. Roch, S. K. Saikin, C. Kreisbeck, D. Sheberla, J. H. Montoya, S. Dwaraknath, M. Aykol, C. Ortiz, H. Tribukait, C. Amador-Bedolla, C. J. Brabec, B. Maruyama, K. A. Persson and A. Aspuru-Guzik, Accelerating the discovery of materials for clean energy in the era of smart automation, *Nat. Rev. Mater.*, 2018, **3**, 5–20, DOI: 10.1038/s41578-018-0005-z.
5 M. R. Endsley, Ironies of artificial intelligence, *Ergonomics*, 2023, **66**(11), 1656–1668, DOI: 10.1080/00140139.2023.2243404.
6 N. J. Szymanski, B. Rendy, Y. Fei, et al., Author Correction: An autonomous laboratory for the accelerated synthesis of inorganic materials, *Nature*, 2026, **650**, E1, DOI: 10.1038/s41586-025-09992-y.
7 R. Palgrave and L. Schoop, Reanalysis of autonomous laboratory materials discovery claims, *ChemRxiv*, 2023, preprint, DOI: 10.26434/chemrxiv-2023-vzf7n.
8 C. Rudin, Stop explaining black box machine learning models for high stakes decisions and use interpretable

models instead, *Nat. Mach. Intell.*, 2019, **1**(5), 206–215, DOI: 10.1038/s42256-019-0048-x.

9 K. T. Butler, D. W. Davies, H. Cartwright, O. Isayev and A. Walsh, Machine learning for molecular and materials science, *Nature*, 2018, **559**(7715), 547–555, DOI: 10.1038/s41586-018-0337-2.

10 B. Burger, P. M. Maffettone, V. V. Gusev, C. M. Aitchison, Y. Bai, X. Wang, X. Li, B. M. Alston, B. Li, R. Clowes, N. Rankin, B. Harris, R. S. Sprick and A. I. Cooper, A mobile robotic chemist, *Nature*, 2020, **583**(7815), 237–241, DOI: 10.1038/s41586-020-2442-2.

11 Z. Ji, N. Lee, R. Frieske, T. Yu, D. Su, Y. Xu, E. Ishii, Y. Bang, A. Madotto and P. Fung, Survey of hallucination in natural language generation, *ACM Comput. Surv.*, 2023, **55**(12), 1–38, DOI: 10.1145/3571730.

12 Wiley, *ExplanAItions 2025: The evolution of AI in research*, 2025, available at https://www.wiley.com/en-us/about-us/ai-resources/ai-study/key-findings/.

13 O. O'Neill, *A question of trust*, Cambridge University Press, 2002.

14 S. Shapin, *A social history of truth: civility and science in seventeenth-century England*, University of Chicago Press, 1994.

15 R. G. L. da Silva, A. Blasimme, E. Vayena and K. E. Ormond, How Do Molecular Systems Engineering Scientists Frame the Ethics of Their Research? AJOB Empirical, *Bioethics*, 2024, **15**(3), 226–235, DOI: 10.1080/23294515.2024.2302994.

16 R. G. L. da Silva and A. Blasimme, From lab to society: Fostering clinical translation of molecular systems engineering, *Bioeng. Transl. Med.*, 2023, **9**(1), e10564, DOI: 10.1002/btm2.10564.

17 J. M. Logg, J. A. Minson and D. A. Minson, Algorithm appreciation: People prefer algorithmic to human judgment, *Organ. Behav. Hum. Decis. Process.*, 2019, **151**, 90–103, DOI: 10.1016/j.obhdp.2018.12.005.

18 S. M. Papidocha, A. Burger, V. Bernales and A. Aspuru-Guzik, The elephant in the lab: synthesizability in generative small-molecule design, *ChemRxiv*, 2025, DOI: 10.26434/chemrxiv-2025-1lcpq.

19 S. X. Leong, C. E. Griesbach, R. Zhang, *et al.*, Steering towards safe self-driving laboratories, *Nat. Rev. Chem.*, 2025, **9**, 707–722, DOI: 10.1038/s41570-025-00747-x.

20 F. Munguia-Galeano, *et al.*, Chemist Eye: safety monitoring in self-driving laboratories, *arXiv*, 2025, preprint, arXiv:2508.05148, DOI: 10.48550/arXiv.2508.05148.

21 L. Longley, *et al.*, RobInHood: a robotic chemist in a fume hood, *ChemRxiv*, 2026, DOI: 10.26434/chemrxiv-2026-s2619.

22 Y. Zhang *et al.*, Safe-SDL: Safety boundaries in AI-driven laboratories, *arXiv*, 2026, preprint, arXiv:2602.15061, DOI: 10.48550/arXiv.2602.15061.

23 R. K. Merton, *The sociology of science: theoretical and empirical investigations*, University of Chicago Press, 1973.

24 J. P. Tennant, H. Crane, T. Crick, J. Davila, A. Enkhbayar, J. Havemann, D. Kwon, *et al.*). Ten hot topics around scholarly publishing, *Publications*, 2019, **7**(2), 34, DOI: 10.3390/publications7020034.

25 D. B. Resnik and M. Hosseini, The ethics of using artificial intelligence in scientific research: new guidance needed for a new tool, *AI Ethics*, 2025, **5**(2), 1499–1521, DOI: 10.1007/s43681-024-00493-8.

26 L. Floridi, J. Cowls, M. Beltrametti, R. Chatila, P. Chazerand, V. Dignum, C. Luetge, R. Madelin, U. Pagallo, R. Ribeiro, L. O. B. Santos, M. Taddeo and E. Vayena, AI4People—An ethical framework for a good AI society: Opportunities, risks, principles, and recommendations, *Minds Mach.*, 2018, **28**(4), 689–707, DOI: 10.1007/s11023-018-9482-5.

27 A. F. Winfield and M. Jirotka, Ethical governance is essential to building trust in robotics and artificial intelligence systems, *Philos. Trans. R. Soc. A*, 2018, **376**(2133), 20180085, DOI: 10.1098/rsta.2018.0085.

28 A. Adadi and M. Berrada, Peeking inside the black-box: A survey on explainable artificial intelligence (XAI), *IEEE Access*, 2018, **6**, 52138–52160, DOI: 10.1109/ACCESS.2018.2870052.

29 M. Messing, The ethics of engineering: Moral muteness, moral imagination, and reflection in interdisciplinary practice, *Sci. Eng. Ethics*, 2022, **28**, 41, DOI: 10.1007/s11948-021-00353-9.