The Mirage of Autonomous AI Scientists

Chenhao Tan [1] **Haokun Liu** [1]

Abstract

Recent efforts to build autonomous AI scientists assume that more automation leads to faster scientific progress. This view is fundamentally flawed. Science is not purely an intelligence problem but a resource allocation problem: deciding what matters among infinite possibilities with limited time, attention, and funding. These allocation decisions cannot be delegated to machines because they are inherently value-laden and require human accountability. We propose that as AI takes over production tasks, the role of scientists will shift toward selection (choosing what to pursue) and evaluation (assessing quality and validity). Technical advances must support these roles, not just automate production. We outline principles for building tools that augment human judgment, scale evaluation capacity, and incentivize wise selection over mere output volume. We conclude with calls to action for the machine learning community to shape AI development for meaningful and accountable scientific progress.

1. Introduction

Science is not inherently valuable (Kitcher, 2001; Douglas, 2009). Most species on earth assign no value to scientific discoveries: a tree does not prioritize climate research, and a bacterium does not seek to understand evolution. What we choose to study, what we consider an important problem, and what we deem a breakthrough all reflect human values, priorities, and needs (Longino, 1990; Kuhn, 1962). Without human judgment, discovery loses its meaning, its purpose, and its worth.

Yet recent efforts to build "AI scientists" often miss this fundamental point. Multiple labs are racing to develop systems that autonomously generate hypotheses, run experiments, and write papers (Lu et al., 2024). The implicit assumption is that more automation equals faster progress. But if science derives its value from human judgment, can we truly automate it away?

Consider NeurIPS submissions as a case study (Figure 1). From 2013 to 2025, submissions grew from 1,400 to over 21,000: a 15-fold increase in just 12 years. Did machine learning progress at the same rate? Can we automate scientific progress by producing more papers? And without human judgment to assess importance and validity, what makes any of these discoveries valuable? The evidence suggests that while there have been remarkable breakthroughs, much of the increased output consists of incremental work (Chu & Evans, 2021; Park et al., 2023). More production has not meant proportionally more progress (known as the "production-progress" paradox).

This pattern reveals a deeper issue with how we think about automation in science. AI excels when tasks have well-defined inputs, outputs, and success metrics. Improving algorithm efficiency is one such example: given a specification of input and output, the goal is clear and progress is measurable. Science as a whole resists this pattern. The goal of science is not to produce papers or even discoveries in isolation; it is to advance human understanding in directions that humans find meaningful. This goal cannot be specified in advance because what counts as "meaningful" depends on context, values, and the evolving state of knowledge itself.

We believe the answer to whether AI can replace human scientists is a clear "No" as long as humanity prevails. AI will not replace human scientists; its real potential lies in reshaping how science is done. As AI expands the search space and takes over routine production tasks, the role of scientists will shift toward selection and evaluation. This shift is not a retreat; instead it reveals what was always essential about the scientific role, now brought into sharper focus by the contrast with what machines can do. As Tukey put it, "Far better an approximate answer to the right question, which is often vague, than an exact answer to the wrong question, which can always be made precise" (Tukey, 1962).

To look beyond the mirage of the "autonomous AI scientist" and reimagine how science moves forward, we make the case with three key arguments. First, science is fundamentally about resource allocation, not just automation (Section 2). Second, accountability in selection and evaluation is what separates science from AI slop (Section 3).

[1]Department of Computer Science, University of Chicago, Chicago, IL, USA. Correspondence to: Chenhao Tan <chenhao@uchicago.edu>.

Preprint. February 2, 2026.

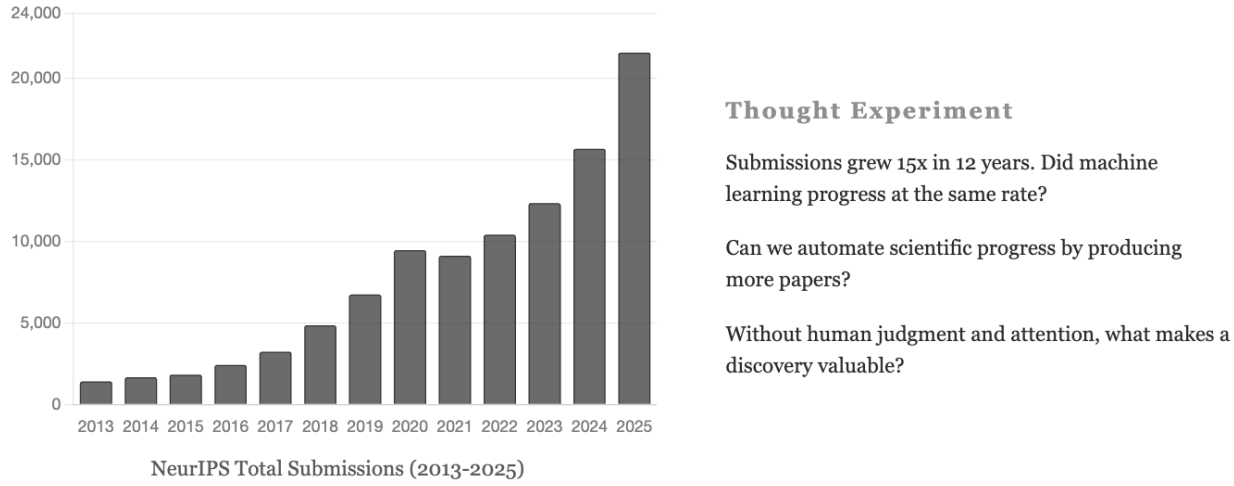

Figure 1. Is scientific progress correlated with the rate of production?

Third, technical advances must support selection and evaluation, not just production (Section 5). We then discuss the broader ecosystem changes required (Section 6) and address common counterarguments (Section 7) before concluding with calls to action for the machine learning community. Our central claim is that the bottleneck in science is not only intelligence or production capacity, but also human judgment and attention. AI can help with the former but cannot substitute for the latter.

2. Science is About Resource Allocation, Not Just Automation

To see why the idea of autonomous AI scientists is a mirage, we need to reframe science as a problem of resource allocation rather than pure intelligence. AI leaders typically portray science as an intelligence problem: build smart enough AI, and breakthroughs follow. However, this is unlikely to be the case. Even if there are genuine breakthroughs in millions of papers that AI generates, scientists may never identify and recognize them given their limited attention. Moreover, AI scientists can be costly. Computational research already consumes vast amounts of compute, but many disciplines also depend on resource-intensive real-world experiments, from biology labs to climate fieldwork. AI does not make those cheaper; instead, it risks multiplying the burden by generating ever more hypotheses that require testing. As Kapoor & Narayanan (2024) argue, this is precisely because it creates more work for the resource-constrained scientific process (again, the *production-progress paradox*).

The production-progress paradox deserves closer examination. Over the past several decades, the number of scientific publications has grown exponentially, yet measures of transformative progress have remained flat or declined. More researchers are producing more papers, but paradigm-shifting discoveries have not increased proportionally. This pattern suggests that the binding constraint on scientific progress is not the rate of paper production but something else entirely: the capacity to identify, validate, and build upon the ideas that matter.

But the real issue is not just the expense. At its core, science is a problem of resource allocation: deciding what matters among infinite possibilities with limited time, attention, and funding. Every choice reflects priorities at multiple levels: funding agencies choose which proposals to fund, scientists choose which idea to pursue, which hypothesis to test, which experiments to run, and which papers to read or write. These decisions form a nested hierarchy of resource allocation, each level constrained by the decisions above it and constraining the decisions below.

Consider the full scope of resource allocation in science. At the macro level, societies decide how much to invest in research versus other priorities, and which broad areas deserve emphasis. At the institutional level, universities and funding agencies distribute resources across fields and investigators. At the lab level, principal investigators allocate budgets and personnel across projects. At the individual level, scientists decide how to spend their hours and attention. Each of these decisions involves tradeoffs that cannot be optimized algorithmically because the objective function itself is contested. Should we prioritize basic research or applied? Incremental advances or risky bets? Problems with clear metrics or those with diffuse importance?

These choices cannot be delegated to machines. They are inherently social and value-laden. What makes a problem "important" depends on personal goals, ethical considerations, and collective priorities. A problem might be important because it affects many people, because it challenges existing theory, because it opens new methodological possibilities,

or because it resonates with cultural concerns. These criteria often conflict, and reasonable people disagree about their relative weight. Even if AI becomes effective at making some of these choices, it cannot be held accountable for those choices. Accountability requires human judgment and ownership.

The value-laden nature of scientific priorities becomes clear when we consider historical examples. The decision to prioritize cancer research over other diseases reflects judgments about suffering, fear, and political salience (Mukherjee, 2010; Proctor, 1995). The choice to fund particle physics at enormous expense reflects beliefs about fundamental knowledge and national prestige (Kevles, 1995; Weinberg, 1992). The recent surge in AI research reflects both commercial incentives and intellectual excitement. None of these allocations are objectively correct; they reflect human values that shift over time and vary across communities.

This resource allocation perspective points to two key roles: **selector** (making resource allocation decisions) and **evaluator** (gathering information to inform those decisions). As AI takes over more production tasks, these become the central responsibilities of scientists. We will need new infrastructure to support these roles. The selector role is not merely choosing among pre-defined options but actively shaping what options exist by framing problems, defining success criteria, and setting research agendas. The evaluator role is not merely checking boxes but exercising judgment about quality, significance, and credibility under uncertainty.

A Simple Model to illustrate the importance of selection. To see why selection becomes more valuable as production scales, consider a simple model. Suppose AI systems produce N scientific outputs (papers, hypotheses, experimental results), each with value v_i drawn independently from a distribution with mean μ. Scientists have limited attention: they can only read, use, and build upon K outputs, where $K \ll N$. Define welfare W as the total value of outputs that scientists actually consume.

Without effective selection, scientists sample K outputs essentially at random. Expected welfare is simply $W_{\text{random}} = K \cdot \mu$, which does not depend on N. Producing more outputs does not help if scientists cannot identify the good ones.

With effective selection and evaluation, scientists can identify and consume the top K outputs by value. The expected welfare W_{selected} equals the sum of the top K order statistics, which grows as N increases. For instance, if values follow a uniform distribution on $[0, 1]$, then $W_{\text{selected}} \approx K \cdot (1 - \frac{K}{2(N+1)})$, approaching K as $N \to \infty$ (see Appendix A for derivation).

The gap $W_{\text{selected}} - W_{\text{random}}$ grows with N: as AI produces more, the value of selection increases. This formalizes the intuition that scaling production without scaling evaluation yields diminishing returns, while investing in selection and evaluation allows welfare to grow with the expanding pool of possibilities. The bottleneck is not production but the capacity to identify what matters.

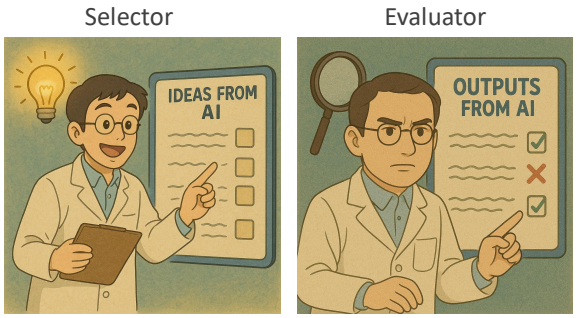

Figure 2. The evolving roles of scientists as AI takes over production tasks. Human scientists shift toward selection (choosing what to pursue) and evaluation (assessing quality and validity), while AI handles generation and execution.

3. Accountability in Selection and Evaluation is What Separates Science from AI Slop

If resource allocation is the bottleneck, then accountability is the principle that keeps those allocations meaningful. That is why selection and evaluation will become the scientist's defining roles. Next we explain these roles and the accountability requirement in detail.

The Selector Role. With AI scientists handling many production tasks, human scientists can dedicate more effort to selection throughout the research process. This includes choosing among research ideas generated by AI, deciding which hypotheses to pursue from those identified through literature, data, or simulation, and selecting implementation strategies for promising directions. A key point is that selection is not a one-time decision at the start, but a continuous process of judgment as research unfolds. Which directions show promise? Which should be abandoned? Human intuition, values, and priorities are critical in these decisions when resources are limited. This applies not only to advancing human understanding, but even to seemingly clear goals like curing cancer, as we still cannot test every possible clinical trial.

The selector role requires a form of expertise that differs from traditional scientific training. It demands broad knowledge to recognize connections across fields, taste to distinguish promising ideas from superficially attractive ones, and courage to pursue unconventional directions. While experience helps develop these capacities, history shows that junior researchers often make the most significant break-

throughs precisely because they are not bound by conventional wisdom. What matters is not seniority but the quality of deliberation about resource allocation. As selection becomes central to scientific work, all researchers must learn to contribute to this deliberation: articulating why certain directions matter, engaging with competing priorities, and building shared understanding of what constitutes important work. This is fundamentally a collective process that requires communication and collaboration.

The Evaluator Role. As AI generates more hypotheses, experimental plans, and code, scientists must rigorously evaluate these outputs. This means checking hypotheses for novelty, importance, and feasibility, detecting methodological flaws in AI-designed experiments, catching errors in AI-written code before they propagate, and assessing results before they cascade through the literature. This evaluation challenge is an instantiation of the scalable oversight problem (Amodei et al., 2016; Bowman et al., 2022): how do we maintain rigorous quality control when AI dramatically increases the volume of scientific output? We already see this challenge in the replication crisis (Open Science Collaboration, 2015). As AI accelerates generation, evaluation becomes increasingly critical.

Evaluation in science is more complex than verification in engineering. When an AI system generates a hypothesis, the question is not simply whether it is true but whether it is interesting, whether the evidence is sufficient, whether alternative explanations have been adequately considered, and whether the framing advances or obscures understanding. These judgments require contextual knowledge that extends beyond the immediate claim to encompass the state of the field, the reliability of methods, and the sociology of how scientific claims are received and built upon.

The challenge of evaluation compounds as AI-generated content proliferates. When every lab can generate thousands of hypotheses, the evaluation bottleneck tightens further. Peer review already struggles to keep pace with human-generated submissions; AI acceleration will stress the system further. This suggests that evaluation capacity may become the limiting factor in scientific progress, making investment in evaluation infrastructure as important as investment in generation capabilities.

Selection and evaluation happen at every step in the iterative process of science. In order for selection to work, scientists must "believe in" the idea. In order for evaluation to work, scientists need to take responsibility for publishing the results. You cannot hide behind "the AI said so." This accountability is what distinguishes the future of science from a world of AI-generated noise. The Virtual Lab of AI agents (**?**) is a good example, where scientists determine the goal and work with AI agents, and thorough validation culminates in a publication in Nature.

The Accountability Requirement. Accountability means that someone can be held responsible for decisions and their consequences. In science, this responsibility operates at multiple levels: researchers are accountable to their peers through peer review, to the public through the expectation that science serves societal needs, and to the historical record through the requirement that claims be replicable and honestly reported. AI systems cannot bear this responsibility because they lack the moral agency and social standing required for accountability. An AI cannot be embarrassed by a retraction, motivated by reputation concerns, or subject to professional sanctions for misconduct.

This is not merely a practical limitation but a conceptual one. Accountability requires that someone stake their credibility on a claim, which means the claim carries information about the claimant's judgment and competence. When a respected scientist endorses a finding, that endorsement provides information beyond the finding itself. It signals that someone with relevant expertise has judged the work worthy of their reputation. This signaling function cannot be replicated by an AI system, regardless of how accurate its outputs become.

4. Revisiting the Production-Progress Paradox

The emphasis on selection and evaluation could address the production-progress paradox. Progress in science comes from deep comprehension, not from producing more papers. When scientists invest effort in careful selection and think deeply about which directions matter and why, they build genuine understanding of the problem space. When they rigorously evaluate results and scrutinize methodology, assessing validity and connecting findings to broader context, they deepen their grasp of what the results actually mean.

An analogy is the "forklift at the gym" problem: if you want to build strength, automating the lifting defeats the purpose. Similarly, automating away the process of understanding defeats the purpose of science itself. Our vision avoids this trap through a specific division of labor. AI eases generation and production, expanding the search space exponentially and bringing more possibilities to examine. Humans handle judgment and accountability, deciding what matters, evaluating quality, and taking ownership of those choices. This is using the forklift to bring more weights to the gym, not to lift them for you.

The forklift analogy illuminates why naive automation can backfire. If the goal were simply to move weights, the forklift would be an unqualified improvement. But the goal is to build strength, which requires the effortful process of lifting. Similarly, if the goal of science were simply to produce

papers, AI automation would be unambiguously beneficial. But the goal is to build understanding, which requires the effortful process of grappling with ideas, evaluating evidence, and integrating findings into a coherent worldview.

This suggests a principle for beneficial AI deployment in science: automate tasks that are instrumental to understanding, not constitutive of it. Writing boilerplate code is instrumental; deciding what code to write is constitutive. Running statistical tests is instrumental; interpreting what the results mean is constitutive. Searching the literature is instrumental; synthesizing it into a novel perspective is constitutive. The challenge lies in drawing this boundary correctly, which itself requires human judgment about what activities contribute to genuine understanding.

The production-progress paradox also reveals a coordination problem. Individual scientists face incentives to publish more, but the collective result of everyone publishing more is not more progress but more noise. This is a classic tragedy of the commons: the resource being depleted is collective attention. AI that accelerates individual production without addressing collective evaluation will worsen this tragedy. The solution requires coordinating on new norms and infrastructure that value evaluation as much as production.

5. Technical Advances Must Support Selection and Evaluation, Not Just Production

Current AI research focuses heavily on automating production with better models, faster inference, and more autonomous systems. But if scientists' essential role is shifting to selection and evaluation, we also need tools and systems that help scientists perform these new roles effectively. The advances must achieve three goals:

- **Augment selection.** Tools should enhance scientists' ability to select, while keeping humans in decision-making roles.
- **Scale up evaluation.** As AI eases production, infrastructure must scale up human evaluation capabilities to match.
- **Incentivize wise selection over mere production.** Create systems that reward good judgment and careful evaluation, not just output volume.

These goals represent a significant reorientation of AI development for science. Current benchmarks and metrics emphasize generation: how many papers can a system produce, how novel are its hypotheses, how quickly can it run experiments. Future metrics must emphasize the quality of human-AI collaboration: how effectively do tools help scientists identify promising directions, how reliably do they flag potential problems, how well do they support the deliberative processes that lead to good judgment.

5.1. Augmenting Selection

Tools for selection should expand the space of possibilities that scientists can consider while preserving their authority to choose among them. Recent work on AI-assisted research ideation points toward promising directions (Wang et al., 2024; Baek et al., 2024; Si et al., 2024; Zhou et al., 2024). AI systems can identify surprising connections across distant literatures, surface neglected problems that combine tractability with potential impact, and generate diverse research directions rather than single recommendations. The key is designing these systems to enhance human exploration rather than constrain it to algorithmic preferences.

More ambitiously, selection tools could help scientists articulate and refine their own values and priorities. What do you actually care about? What would change your mind? AI systems that engage scientists in structured reflection about their goals may prove more valuable than systems that simply recommend directions. Tools that generate adversarial hypotheses, challenge assumptions, or surface inconvenient evidence could strengthen selection by forcing deeper engagement with alternatives. The most transformative tools may be those that help research communities deliberate collectively about priorities, aggregating diverse perspectives into shared research agendas.

However, designing selection tools that preserve human agency is difficult. Automation bias may lead scientists to over-rely on AI suggestions, gradually atrophying their own judgment. Algorithmic monocultures could homogenize research directions if many scientists use the same tools (Kleinberg & Raghavan, 2021). The serendipity that drives unexpected breakthroughs may be lost if AI systems optimize for expected value. These challenges suggest that selection tools must be designed not just for accuracy but for maintaining the diversity, independence, and active engagement that make scientific communities effective.

5.2. Scaling Evaluation

As AI-generated content proliferates, evaluation capacity must scale accordingly. This requires both better tools for individual evaluation and better infrastructure for collective evaluation. Individual tools might include automated checks for common errors, comparison with existing literature to assess novelty, and structured prompts that guide systematic assessment. Collective infrastructure might include platforms for distributed peer review, mechanisms for aggregating evaluations across many reviewers, and systems that track the reliability of sources over time.

The scaling challenge is profound. If AI increases the rate

of research production by a factor of 1000, whether papers, code, experimental results, or datasets, evaluation capacity must increase correspondingly or the system will be overwhelmed. This cannot be achieved simply by asking scientists to evaluate faster; it requires fundamentally new approaches to evaluation that leverage AI assistance while preserving human judgment at critical points.

One promising direction is tiered evaluation, where AI systems perform initial screening to filter out clearly flawed or uninteresting work, while human experts focus their attention on the most promising candidates. This approach mirrors how search engines work: algorithmic ranking handles the bulk of filtering, while human attention is reserved for the top results. The challenge is ensuring that the initial filtering does not systematically exclude valuable work that does not fit expected patterns.

5.3. Principles for Tool Design

Several principles should guide the development of tools for selection and evaluation:

Preserve Human Agency. Tools should present options and provide information, but leave final decisions to humans. This means avoiding systems that automatically select directions or filter out possibilities without human review. Even when AI recommendations are highly accurate, the process of considering and deciding builds the judgment that makes future decisions better. Removing humans from this loop may increase short-term efficiency at the cost of long-term capability.

Support Deliberation. Rather than optimizing for speed, tools should facilitate thoughtful consideration. This might include mechanisms for recording reasoning, comparing alternatives, and revisiting past decisions. Deliberation is where understanding develops; tools that rush this process sacrifice the cognitive benefits that make selection valuable. This runs counter to much current AI development, which optimizes for speed and convenience rather than depth of engagement.

Enable Collaboration. Selection and evaluation benefit from multiple perspectives. Tools should support discussion, disagreement, and collective judgment across research communities. Science is a social process, and the quality of scientific judgment depends on the quality of scientific discourse. Tools that isolate researchers may increase individual productivity while degrading collective intelligence.

Maintain Accountability. Every decision should have a clear human owner. Tools should create records of who made what choices and why, enabling both credit assignment and responsibility tracking. This serves both practical and epistemic purposes: practically, it ensures someone is responsible for errors; epistemically, it allows the community to calibrate trust based on track records. Anonymous or automated decisions undermine both functions.

6. The Broader Ecosystem

Looking forward, the future of science requires much more than better tools. It requires rethinking the entire ecosystem within which scientific work occurs. Here are some examples:

- AI systems should generate effective hypotheses from data, literature, and other computational approaches for human selection. This means moving beyond single-best recommendations to diverse portfolios of options with clear explanations of their tradeoffs.
- Automation should handle execution tasks while maintaining accountability for key decisions. This requires clear interfaces that mark where automated execution ends and human judgment begins.
- Publication mechanisms need rethinking. AI-run venues can surface new ideas and workflows, and we need to understand how they complement existing publication systems. The current model of peer-reviewed papers may need to evolve toward more dynamic formats that better reflect the iterative nature of AI-assisted science.
- Funding mechanisms should allocate resources to encourage and reward selection and evaluation capacity. This might mean funding "evaluation grants" that support rigorous assessment of existing work, or creating prizes for identifying flawed research before it propagates.
- Academic systems need reform to value contribution to understanding, creating career paths for evaluators and infrastructure builders. Currently, careers are built on production metrics; future systems should equally reward those who improve the quality of collective judgment.

Incentive reform. The most challenging aspect of the broader ecosystem is incentive reform. Current incentives strongly favor production over evaluation. Scientists are promoted based on publication counts and citation metrics, not on the quality of their judgment or their contributions to collective understanding. Funding flows to those who promise new discoveries, not to those who carefully evaluate existing claims. These incentives will not change automatically; they require deliberate reform.

Several mechanisms could help shift incentives toward evaluation. Funding agencies could require that a portion of grant budgets be allocated to replication and evaluation rather than new production. Journals could weight editorial decisions toward papers that carefully evaluate existing work rather than simply adding to the pile. Hiring committees could explicitly value demonstrated skill in evaluation

and selection, treating these as core competencies rather than service activities.

Technology can also play a role in incentive reform by making evaluation contributions more visible and measurable. Platforms that track who identified important problems early, who caught errors before publication, and who synthesized disparate findings into coherent frameworks could provide the data needed to reward these activities. Currently, these contributions are largely invisible; making them visible is a prerequisite for valuing them appropriately.

The goal of incentive reform is not to devalue production but to properly value the complementary activities of selection and evaluation. Science needs both generation and curation; the current imbalance toward generation reflects historical circumstances in which human generation capacity was the limiting factor. As AI shifts this constraint, the relative value of human evaluation and selection increases, and incentive systems should adjust accordingly.

7. Alternative Views

We address several objections to the framework presented above.

"AI will eventually develop its own values and priorities." This argument assumes that artificial general intelligence will emerge and develop autonomous goals. Even if this occurs, it does not solve the fundamental problem: science serves human purposes. An AI with its own values would be pursuing its own science for its own reasons, which would not satisfy the human need for understanding and control over nature. Moreover, the question of whose values should guide resource allocation in science is inherently a social and political question, not a technical one.

The deeper issue is that values in science are not merely inputs to an optimization process but are themselves discovered and refined through scientific practice. What counts as an important problem evolves as fields develop. An AI system that autonomously pursued science based on fixed values would miss this essential dynamism. And an AI that adapted its values based on its own criteria would be making choices that humans have not authorized and cannot oversee. Either way, the result would not be science in the sense that serves human purposes.

"Human judgment is biased and slow; AI would be more objective." Human judgment is indeed biased, but this is precisely why accountability matters. When humans make decisions, they can be questioned, challenged, and held responsible. When biases are identified, processes can be reformed. AI systems encode the biases of their training data and designers, but without the same mechanisms for accountability and correction. Furthermore, "objectivity" in science is not about removing human judgment but about subjecting that judgment to scrutiny through peer review, replication, and open debate.

The history of science shows that progress often comes from challenging rather than eliminating subjective judgment. Galileo's observations were dismissed as subjective illusions by contemporaries who had different theoretical commitments. Darwin's theory was attacked as reflecting his personal ideology. In each case, progress came not from removing human judgment but from refining it through debate and evidence. AI systems that bypass this process by claiming objectivity would short-circuit the mechanisms through which science corrects itself.

Speed is also not unambiguously desirable. Careful evaluation takes time, and rushing this process leads to errors that are costly to correct. The replication crisis demonstrates what happens when the scientific system prioritizes speed over rigor. AI that further accelerates production without correspondingly improving evaluation will worsen rather than solve this problem.

"This framework applies only to some fields." While the specific balance between selection, evaluation, and production varies across disciplines, the fundamental insight holds broadly. Even in fields where AI can fully automate experiments, humans must still decide which experiments to run and what the results mean. A drug discovered by AI still requires human judgment about whether to pursue clinical trials, how to weigh risks and benefits, and how to allocate limited healthcare resources.

Consider mathematics, sometimes seen as the field most amenable to AI automation because of its formal nature. Even here, human judgment determines which theorems are interesting, which proof strategies are elegant, and which results warrant publication. A theorem-proving AI might generate thousands of valid proofs, but mathematicians must decide which ones advance understanding. The same pattern holds in experimental sciences: even perfect automation of data collection and analysis leaves open the questions of what data to collect and what the analysis means.

"AI-generated ideas are just more noise." This concern is precisely why we emphasize the selector role. AI can surface possibilities that might otherwise be overlooked (e.g., the famous move 37 by AlphaGo), thus broadening the search space. But quantity does not equal quality: only through human selection and evaluation can those raw outputs become meaningful ideas rather than unfiltered noise. Used appropriately, AI can also make studies more replicable when applied to production tasks (Kapoor & Narayanan, 2024).

The key is that AI-generated ideas become valuable only when integrated into a process that includes human selection. An idea that no one evaluates or pursues is not really an idea at all; it is just text. The value of AI in idea generation comes from expanding the range of possibilities that humans consider, not from bypassing human consideration entirely. This is why tools for selection are essential complements to tools for generation.

"Evaluation cannot scale with AI production." This is a serious concern but not an argument against the framework; rather, it is an argument for investing in evaluation infrastructure. Current evaluation capacity is insufficient even for human-generated content, as evidenced by journal backlogs and the replication crisis. AI acceleration makes this problem more urgent but does not change its nature. The solution is to develop new approaches to evaluation that leverage AI assistance while preserving human judgment at critical points, as discussed in Section 5.

The alternative, accepting that evaluation cannot scale, leads to a world where the scientific literature is flooded with unchecked AI-generated content. This outcome would be far worse than the current situation, effectively destroying the reliability that makes scientific knowledge valuable. The difficulty of scaling evaluation is real, but the consequences of failing to do so are severe enough to warrant significant investment in solutions.

"Incentive structures will never change." Current incentives that reward publication volume are already under criticism within the scientific community. The pressures created by AI-generated content may accelerate reform by making the inadequacy of current metrics more obvious. Funding agencies, universities, and journals all have reasons to value quality over quantity, and tools that support selection and evaluation can help shift incentives by making these contributions more visible and measurable.

Historical precedent suggests that incentive structures can change when circumstances demand it. The rise of peer review, the development of citation metrics, and the growth of open access all represented significant shifts in how science is organized and rewarded. Each of these changes faced resistance but ultimately prevailed because they served genuine needs. The shift toward valuing evaluation and selection may follow a similar trajectory, driven by the recognition that current metrics fail to capture what matters as AI changes the nature of scientific production.

8. Conclusion

The mirage of the autonomous AI scientist dissolves once we recognize that science is fundamentally a resource allocation problem, not merely an intelligence problem. Generating more output does not automatically lead to more progress; what matters is choosing wisely among infinite possibilities. We have argued that selection and evaluation are the essential roles humans must retain, as these require accountability that cannot be delegated to machines. Technical development should support rather than replace human judgment: augmenting selection, scaling evaluation, and incentivizing wise choices over mere production volume.

This argument has particular relevance for the machine learning community. Machine learning is itself a science subject to the dynamics we describe: the explosion of submissions in machine learning conferences illustrates the production-progress paradox within our own field. But more importantly, the tools that ML researchers build will shape how this transformation unfolds across all of science. The choice between designing for full automation versus human-AI collaboration is not abstract; it is made concrete in every system we build, every benchmark we create, and every metric we optimize.

We close with concrete calls to action:

For ML researchers building AI for science. Design systems that augment human judgment rather than bypass it. This means building tools that present diverse options with explanations rather than single recommendations, that support deliberation rather than optimize for speed, and that maintain clear accountability for decisions. Evaluate success not by automation metrics but by whether scientists make better judgments with your tools than without them.

For the ML community as a whole. Confront the production-progress paradox in our own field. Develop better metrics that capture scientific contribution rather than mere output. Create incentives for careful evaluation and replication. Recognize that the most valuable contributions may be those that help us identify which of the thousands of papers actually matter.

For scientific institutions and funding agencies. Invest in evaluation infrastructure with the same intensity as production infrastructure. Fund replication studies, evaluation tools, and the training of scientists in selection and judgment. Reform incentive structures to reward demonstrated wisdom, not just prolific output.

The path forward requires recognizing that the bottleneck in science is human judgment and attention. AI should support the exercise of that judgment, not replace it. The autonomous AI scientist is a mirage because science without humans is not science at all. The real opportunity is human scientists working with AI, combining human judgment with machine capability to accelerate discovery while preserving accountability. Realizing this opportunity is both the challenge and the responsibility of our community.

Acknowledgements

We are grateful for valuable input from Davi Costa, Raul Castro Fernandez, James Evans, Ian Foster, Cristina Garbacea, Ari Holtzman, Xiao Liu, Hao Peng, Amit Sharma, and Ted Underwood.

Impact Statement

This paper presents work whose goal is to advance understanding of the relationship between AI and scientific practice. We argue for maintaining human accountability in science, which we believe is essential for ensuring that AI development serves human values and needs. Our framework emphasizes the importance of human judgment in resource allocation decisions, which has implications for how AI systems should be designed and deployed in scientific contexts.

References

Amodei, D., Olah, C., Steinhardt, J., Christiano, P., Schulman, J., and Mané, D. Concrete problems in AI safety. *arXiv preprint arXiv:1606.06565*, 2016.

Baek, J., Jauhar, S. K., Cucerzan, S., and Hwang, S. J. Researchagent: Iterative research idea generation over scientific literature with large language models. *arXiv preprint*, 2024.

Bowman, S. R., Hyun, J., Perez, E., Chen, E., Pettit, C., Heiner, S., Lukášik, K., Askell, A., Jones, A., Chen, A., et al. Measuring progress on scalable oversight for large language models. *arXiv preprint arXiv:2211.03540*, 2022.

Chu, J. S. and Evans, J. A. Slowed canonical progress in large fields of science. *Proceedings of the National Academy of Sciences*, 118(41):e2021636118, 2021.

Douglas, H. E. *Science, Policy, and the Value-Free Ideal*. University of Pittsburgh Press, 2009.

Kapoor, S. and Narayanan, A. Could AI slow science? Normal Tech Newsletter, 2024. URL https://www.normaltech.ai/p/could-ai-slow-science.

Kevles, D. J. *The Physicists: The History of a Scientific Community in Modern America*. Harvard University Press, 1995.

Kitcher, P. *Science, Truth, and Democracy*. Oxford University Press, 2001.

Kleinberg, J. and Raghavan, M. Algorithmic monoculture and social welfare. *Proceedings of the National Academy of Sciences*, 118(22):e2018340118, 2021.

Kuhn, T. S. *The Structure of Scientific Revolutions*. University of Chicago Press, 1962.

Longino, H. E. *Science as Social Knowledge: Values and Objectivity in Scientific Inquiry*. Princeton University Press, 1990.

Lu, C., Lu, C., Lange, R. T., Foerster, J., Clune, J., and Ha, D. The AI scientist: Towards fully automated open-ended scientific discovery. *arXiv preprint arXiv:2408.06292*, 2024.

Mukherjee, S. *The Emperor of All Maladies: A Biography of Cancer*. Scribner, 2010.

Open Science Collaboration. Estimating the reproducibility of psychological science. *Science*, 349(6251):aac4716, 2015.

Park, M., Leahey, E., and Funk, R. J. Papers and patents are becoming less disruptive over time. *Nature*, 613(7942): 138–144, 2023.

Proctor, R. N. *Cancer Wars: How Politics Shapes What We Know and Don't Know About Cancer*. Basic Books, 1995.

Si, C., Yang, D., and Hashimoto, T. Can llms generate novel research ideas? a large-scale human study with 100+ nlp researchers. *arXiv preprint*, 2024.

Tukey, J. W. The future of data analysis. *The Annals of Mathematical Statistics*, 33(1):1–67, 1962.

Wang, Q., Downey, D., Ji, H., and Hope, T. Scimon: Scientific inspiration machines optimized for novelty. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 279–299, 2024.

Weinberg, S. *Dreams of a Final Theory*. Pantheon Books, 1992.

Zhou, Y., Liu, H., Srivastava, T., Mei, H., and Tan, C. Hypothesis generation with large language models. *arXiv preprint arXiv:2404.04326*, 2024.

A. Derivation of Selection Welfare

We derive the expected welfare under selection for the uniform distribution case. Suppose N outputs have values $v_1, \ldots, v_N$ drawn independently from the uniform distribution on $[0, 1]$. Let $X_{(1)} \leq X_{(2)} \leq \cdots \leq X_{(N)}$ denote the order statistics.

For the uniform distribution on $[0, 1]$, the expected value of the k-th order statistic is:

$$\mathbb{E}[X_{(k)}] = \frac{k}{N+1} \tag{1}$$

With effective selection, scientists consume the top K outputs, i.e., the order statistics $X_{(N)}, X_{(N-1)}, \ldots, X_{(N-K+1)}$. The expected welfare is:

$$W_{\text{selected}} = \sum_{j=0}^{K-1} \mathbb{E}[X_{(N-j)}] = \sum_{j=0}^{K-1} \frac{N-j}{N+1} \tag{2}$$

$$= \frac{1}{N+1} \sum_{j=0}^{K-1} (N-j) \tag{3}$$

$$= \frac{1}{N+1} \left(KN - \sum_{j=0}^{K-1} j \right) \tag{4}$$

$$= \frac{1}{N+1} \left(KN - \frac{K(K-1)}{2} \right) \tag{5}$$

$$= \frac{K}{N+1} \left(N - \frac{K-1}{2} \right) \tag{6}$$

$$= K \cdot \frac{2N-K+1}{2(N+1)} \tag{7}$$

This can be rewritten as:

$$W_{\text{selected}} = K \cdot \left(1 - \frac{K}{2(N+1)} \right) + \frac{K}{2(N+1)} \tag{8}$$

For $K \ll N$, this simplifies to:

$$W_{\text{selected}} \approx K \cdot \left(1 - \frac{K}{2(N+1)} \right) \tag{9}$$

As $N \to \infty$, $W_{\text{selected}} \to K$, meaning perfect selection allows scientists to capture nearly the maximum possible value.

In contrast, without selection, scientists sample K outputs at random. Since $\mathbb{E}[v_i] = \frac{1}{2}$ for the uniform distribution:

$$W_{\text{random}} = K \cdot \frac{1}{2} \tag{10}$$

The welfare gap is:

$$W_{\text{selected}} - W_{\text{random}} \approx K \cdot \left(\frac{1}{2} - \frac{K}{2(N+1)} \right) = \frac{K}{2} \cdot \left(1 - \frac{K}{N+1} \right) \tag{11}$$

This gap approaches $\frac{K}{2}$ as N grows, demonstrating that the value of selection increases with the scale of production.