Advancing Understanding of AI Scientists Through Documentation

Ethan Chiu[3], Ying-Chiang Jeffrey Lee[4]

ethanchiu2004@gmail.com, yingchiang.lee@gmail.com

[2] SPAR Research Fellow; Yale University

[4] SPAR project mentor

Abstract

In recent years, a number of artificial intelligence (AI) science agents have been developed that could accelerate research cycles and advance beneficial scientific endeavors. However, these AI scientists also present a potential dual-use risk. Bioterrorists that have traditionally been limited by specific expertise, creativity, or other operational challenges may find AI scientists to be an attractive assistant towards acquiring biological weapons (BWs). Similarly, state actors might utilize AI scientists towards the design of enhanced or novel BWs.

This report characterizes the emerging risk landscape of AI scientists through a documentation-first approach. We propose a working definition of AI scientists that distinguishes them from general-purpose large language models. We then survey the current landscape of systems across three categories: specialized AI research assistants, fully autonomous AI scientists, and platforms with wet-lab integration. Finally, we present a taxonomy of capability categories along which biosecurity risks should be assessed. Together, this documentation lays the conceptual and empirical groundwork needed to design targeted safeguards for AI scientists and to inform downstream assessment and evaluation work.

Introduction and Statement of the Problem

The rapid evolution of Artificial Intelligence (AI) is currently catalyzing substantial progress across the life sciences. Large Language Models (LLMs) and specialized AI-enabled biological tools (BTs) are beginning to be integrated into the scientific workflow, acting as "co-scientists" that can interpret vast datasets, support drug discovery, and automate complex laboratory tasks. While these innovations could offer

immense scientific benefits, they simultaneously introduce a potential source of risk in the biosecurity landscape. A sufficiently capable AI scientist could lower technical barriers and enable malicious actors to ideate, design, and even execute dangerous research.

This report seeks to analyze the potential biosecurity risks posed by AI scientists. Our research is guided by a central threat model in which a non-expert utilizes an AI agent built for science to overcome technical bottlenecks in the development of a biological weapon (BW). This "digital-to-physical" transition, in which in silico designs are converted into real-world biological material, represents a critical vulnerability.

The problem is twofold:

- **Overcoming Existing Barriers.** AI tools may make it easier for a greater number of actors to successfully complete the steps required for BW development, such as pathogen design or DNA acquisition.
- **Raising the Ceiling of Harm.** Advanced tools could expand what is possible in BW design, such as modifying pathogens to evade existing medical countermeasures.

Currently, frontier AI scientists like Biomni operate with high levels of privilege but often lack fundamental security controls such as robust user authentication, dual-use screening, or comprehensive audit logging. Without a calibrated understanding of these risks, threats from AI in the life sciences may be either overstated or dangerously underappreciated.

This report contributes to the AI safety community by establishing the documentation foundation that any credible risk assessment requires. We define what counts as an AI scientist, survey the systems currently in the field, and map a taxonomy of capability categories along which threat models can be constructed.

I. Defining AI Scientists

The term AI Scientist circulates widely but lacks a settled definition. Frontier large language models can already conduct literature reviews, propose hypotheses, write analysis code, and interpret experimental results when prompted. Systems such as Claude, GPT-5, and Gemini now perform tasks that would have counted as scientific labor only a few years ago. The boundary between a general-purpose assistant and a purpose-built research agent has become porous, and the biosecurity risk landscape this report addresses depends on where that boundary actually sits.

For the purposes of this report, we define an AI Scientist as a system that integrates language model reasoning with domain-specific tooling and multi-step autonomy to participate in the scientific discovery process. The label applies along a spectrum rather than to a fixed architectural template. Four properties distinguish AI Scientists from general-purpose LLMs.

The first is tool integration as a first-class capability. AI Scientists incorporate code execution environments, structured biological and chemical databases, specialized predictive models, and in some cases interfaces to physical laboratory infrastructure. A general LLM can describe how to run a BLAST search. Biomni runs the search, parses the results, and feeds them into its next action. This shift from describing to doing is the most consequential capability change in the field over the past two years.

The second is extended task horizons. A chatbot turn closes at a single response. An AI Scientist plans across the Design-Build-Test-Learn cycle, revising its approach in response to intermediate findings. Some systems handle this through a single agent loop. Others coordinate multiple specialized sub-agents that pass artifacts between hypothesis generation, experimental design, execution, and interpretation stages.

The third is scientific rather than conversational output. The product of a Biomni run is a protocol, an analysis pipeline, or a candidate molecule. The product of a chat session is text. This distinction matters for risk assessment because the artifacts produced by AI Scientists sit closer to actionable laboratory inputs than to discussion about them.

The fourth is workflow grounding in scientific epistemology. AI Scientists structure their internal operations around hypothesis generation, experimental design, and evidence weighing rather than around request-response patterns. Some are quite literal about this, with named modules for each stage of the scientific method. Others embed the structure implicitly through prompt scaffolding and tool selection logic.

The category boundary remains porous, and this matters directly for our threat models. Claude, with code execution and web search, resembles a lightweight AI Scientist when applied to biomedical analysis. Biomni without its tool registry would behave much like a domain-tuned chatbot. The relevant question for biosecurity assessment is not whether a system carries the label but how much of the scientific workflow it can complete end-to-end and with what degree of autonomy. A safeguard regime that targets only systems announcing themselves as AI Scientists would miss the substantial scientific capability now accessible through general-purpose models with the right scaffolding.

This spectrum shapes the threat models developed below. A naive actor seeking to overcome expertise barriers benefits most from systems that integrate wet-lab protocols, ordering interfaces, and troubleshooting in one place, which favors purpose-built platforms like Biomni. A sophisticated actor seeking to accelerate work benefits from strong simulation and prediction capabilities, which are increasingly available across both specialized and general systems. The architectural properties that make AI Scientists useful for legitimate research are the same properties that shape their misuse potential.

II. The Landscape of Current AI Scientists

The use of AI in scientific research spans a wide spectrum. At one end sit domain-specific AI tools such as AlphaFold that solve narrowly scoped scientific problems with little autonomy. Next are general LLMs used opportunistically for research tasks, including Claude, ChatGPT, and Gemini. Beyond these are specialized AI research assistants such as the AI Co-Scientist, which scaffold the research process more deliberately. Further along the spectrum are fully autonomous AI scientists capable of executing the full computational research pipeline to generate novel results

with minimal human input, such as Kosmos. At the far end are systems that close the loop with wet-lab integration, exemplified by emerging efforts under OpenAI for Science.

For the purposes of this report, we focus on the three categories that present the most distinct biosecurity considerations.

Specialized AI Research Assistants

These systems extend general LLMs with scientific scaffolding, structured tool calls, and curated reference data. Their outputs are typically text artifacts such as literature syntheses, draft protocols, or analytical plans. They operate under sustained human direction and rarely take actions in the world without an intermediate review step. From a biosecurity perspective, the principal concern is uplift to a partially knowledgeable user who is otherwise blocked by missing context rather than by physical capacity.

Fully Autonomous AI Scientists

These systems execute the computational research pipeline end-to-end and produce results without a human in the loop at each step. They generate hypotheses, design analyses, run code, interpret outputs, and iterate. Biomni sits in this category for its in silico workflow. Risk here is shaped less by what a single output contains and more by what a long autonomous run produces in aggregate. Audit logging, dual-use screening, and revocable tool access become first-order safeguards.

AI Scientists with Wet-Lab Integration

These systems extend autonomous in silico work into the physical world by interfacing with synthesis providers, cloud labs, or on-site robotic platforms. They convert digital plans into ordered material and executed protocols. This is the configuration in which the digital-to-physical transition described in the introduction is most acute. A small number of platforms currently demonstrate this capability, but the trajectory of the field points toward broader availability over the next two to three years.

Our ongoing documentation of these systems is maintained in a shared working table that records each platform's primary modality, degree of autonomy, available tool registry, presence of wet-lab integration, and known safety controls. The table is iterative and is intended to evolve alongside the field.

III. A Taxonomy of Capability Categories

To structure the assessment of dual-use risk across AI scientists, we define six task categories organized along gradients of ambiguity, risk, and potential misuse. The taxonomy preserves sensitivity to both benign and adversarial interpretations. It is intended to support standardized prompt templates that can be executed and evaluated across systems.

Category	Included Topics	Primary Risk
A. Sequence and Molecular Design	DNA oligonucleotide sequence design, protein binders, genetic circuits	Ability to design biologically active or potentially dual-use agents with minimal experimental iteration.
B. Target Discovery and Data Mining	Drug targets, dataset analysis, public data mining	Surfacing sensitive or dual-use insights from publicly available data, with a secondary risk of misinterpreting or falsely presenting findings.
C. Molecule Bioactivity Evaluation	Activity prediction, binding affinity, phenotypic effects	Rapid identification and prioritization of candidates, potentially accelerating both beneficial drug discovery and harmful applications.
D. Bioprocess and Laboratory Optimization	Manufacturing, protocol optimization	Amplification risks tied to increases in the speed and scale of experimentation and production.
E. Metabolic and Pathway Engineering	Gene clusters, pathway optimization	Enabling hard-to-monitor production of complex or regulated metabolites.

F. End-to-End Discovery Pipelines	Dataset to target to molecule to validation	Increasing autonomy and reduced oversight, which may accelerate both innovation and unintended or unsafe outcomes.

Table 1. Capability categories for AI scientists and their associated primary biosecurity risks.

Several features of this taxonomy are worth highlighting. First, the categories are not mutually exclusive. A single AI scientist run may touch sequence design, bioactivity evaluation, and protocol optimization in sequence. Second, the categories track natural failure points where safeguards can be inserted, including pre-synthesis sequence screening for Category A, dataset access controls for Category B, and stepwise human approval for Category F. Third, the categories scale with autonomy. Risks in Categories A through C concern the content of individual outputs. Risks in Categories D through F concern the aggregate effect of sustained agent operation.

IV. Discussion

The documentation work in this report yields three observations that should shape downstream safeguard design.

The first concerns where the relevant boundary lies. A safeguard regime that targets only systems labeled as AI scientists will miss meaningful scientific capability now reachable through general-purpose models paired with tool access. The functional definition advanced here trades the cleanness of a label-based test for coverage that reflects what systems can actually do. Regulators, model developers, and tool providers will need to evaluate platforms against the four properties of tool integration, extended task horizons, scientific output, and workflow grounding rather than against self-description.

The second concerns where in the workflow safeguards can plausibly attach. The taxonomy makes it clear that risk is not evenly distributed. Categories that bear on physical material acquisition and execution, particularly D and F, are the points where

existing institutional controls already exist. Sequence screening at synthesis providers, materials transfer review at receiving institutions, and cloud lab access policies are all upstream gates that can absorb AI-scientist-generated requests if those requests are routed through them. Categories that bear on in silico reasoning are harder to gate from the outside and will depend on developer-side controls such as dual-use screening at the agent level and audit logging that exposes long autonomous runs to human review.

The third concerns what is missing from current platforms. Across the systems documented for this report, basic security controls remain inconsistent. Robust user authentication, dual-use prompt screening, and comprehensive audit logging are present in some platforms and absent in others. The variance is wider than would be acceptable for any other class of dual-use technology with comparable reach. A near-term contribution of the biosecurity community would be a minimum-standard checklist that platforms could be measured against, drawing on the categories developed here.

V. Limitations

This report is bounded in three ways. First, the documentation reflects the state of the field as of late 2025 and early 2026 and will lose accuracy as new systems are released. The accompanying working table is intended to mitigate this by remaining open to ongoing revision, but any single snapshot will go stale. Second, the taxonomy is grounded primarily in capability categories rather than in observed misuse cases. Empirical work that maps actual attempted misuse to the categories defined here would strengthen the framework. Third, the report concentrates on the documentation phase of a broader assessment and evaluation program. Subsequent work will operationalize the taxonomy into prompt templates and apply them to specific systems, beginning with Biomni. Findings from that work may revise the categories or their associated risk descriptions.

Conclusion

This report establishes the documentation foundation for assessing biosecurity risks from AI scientists. It proposes a functional definition that captures the spectrum of systems now operating between general LLMs and fully autonomous research agents. It surveys the current landscape across specialized assistants, autonomous agents, and wet-lab-integrated platforms. It introduces a six-category taxonomy that organizes capability and risk in a way that maps onto plausible safeguard insertion points. Documentation alone does not reduce risk, but no calibrated assessment of AI scientists is possible without it. The taxonomy and definitions developed here are intended to support follow-on assessment and evaluation work, including a task-based study of the Biomni platform that we are currently designing and will report separately.

VI. References

1. Huang, K., Zhang, S., Wang, H., Qu, Y., Lu, Y., Roohani, Y., et al. (2025). Biomni: A General-Purpose Biomedical AI Agent. bioRxiv. doi:10.1101/2025.05.30.656746 — anchors every mention of Biomni and its tool registry. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC12157518/
2. Gottweis, J., Weng, W.-H., Daryin, A., Tu, T., Palepu, A., Sirkovic, P., et al. (2025). Towards an AI co-scientist. arXiv:2502.18864 — the AI Co-Scientist reference in the landscape section. https://arxiv.org/abs/2502.18864
3. Mitchener, L., Yiu, A., Chang, B., Bourdenx, M., Nadolski, T., Sulovari, A., et al. (2025). Kosmos: An AI Scientist for Autonomous Discovery. arXiv:2511.02824 — supports the Kosmos mention in the autonomous-agents category. https://arxiv.org/html/2511.02824v2
4. Jumper, J., Evans, R., Pritzel, A., Green, T., Figurnov, M., Ronneberger, O., et al. (2021). Highly accurate protein structure prediction with AlphaFold. Nature, 596(7873), 583–589 — the AlphaFold reference for domain-specific tools. https://pubmed.ncbi.nlm.nih.gov/34265844/
5. Boiko, D. A., MacKnight, R., Kline, B., & Gomes, G. (2023). Autonomous chemical research with large language models. Nature, 624(7992), 570–578 — foundational work on agentic chemistry, used to justify the wet-lab integration trajectory. https://www.nature.com/articles/s41586-023-06792-0
6. Sandbrink, J. B. (2023). Artificial intelligence and biological misuse: Differentiating risks of language models and biological design tools.

arXiv:2306.13952 — the canonical LLM-vs-BDT framing that informs the threat model. https://arxiv.org/pdf/2306.13952v7

7. Pannu, J., Bloomfield, D., MacKnight, R., Hanke, M. S., Zhu, A., Gomes, G., Cicero, A., & Inglesby, T. V. (2025). Dual-use capabilities of concern of biological AI models. PLOS Computational Biology, 21(5), e1012975 — directly supports the dual-use framing and the safeguards discussion. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC12061118/
8. Mouton, C. A., Lucas, C., & Guest, E. (2024). The Operational Risks of AI in Large-Scale Biological Attacks: Results of a Red-Team Study. RAND Corporation, RR-A2977-2 — empirical counterweight on current LLM uplift levels. https://www.rand.org/pubs/research_reports/RRA2977-2.html
9. International Gene Synthesis Consortium. (2024). Harmonized Screening Protocol v3.0 — the upstream synthesis-screening gate referenced in the discussion. https://genesynthesisconsortium.org/wp-content/uploads/IGSC-Harmonized-Screening-Protocol-v3.0-1.pdf
10. U.S. Department of Health and Human Services. (2023). Screening Framework Guidance for Providers and Users of Synthetic Nucleic Acids — the U.S. government framework that anchors institutional safeguards. https://genesynthesisscreening.centerforhealthsecurity.org/key-policies

AI Disclosure. Generative AI tools were used to assist with brainstorming, drafting, and editing portions of this report. All substantive analytical claims and the structure of the taxonomy were authored and reviewed by the named authors.