AI Scientist Agents and Biosecurity: Capabilities, Risks, and Governance for Autonomous Labs

Dmitrii Magas [1]

Abstract

The transition from conversational large language models to tool-using AI scientist agents has increased both the capabilities and the risks of AI-driven science. This position paper argues that current biosecurity debates often conflate informational assistance with executable scientific agency. That distinction matters: today's AI scientist agents are increasingly effective at structured orchestration, long-horizon planning, and interaction with robotic or software-mediated workflows, but they still struggle with tacit wet-lab manipulation, ambiguous physical feedback, and open-ended troubleshooting. As a result, they do not reduce barriers uniformly across actors. This perspective shifts near-term concern away from generic "LLMs enable bioterrorism" narratives and toward actor–infrastructure pairings that can combine agentic systems with structured execution environments, especially well-funded non-state groups and some agricultural misuse pathways. We use this argument to offer a structured critique of prevailing threat narratives, a prioritization of near-term biosecurity risks under current AI scientist capabilities, and recommendations for governing the execution layer of AI-driven science through stronger synthesis screening, better biosurveillance, and greater investment in AI-assisted defensive response.

1. Introduction

The potential for artificial intelligence to facilitate catastrophic biological events is a subject of intense contemporary debate, especially with the proliferation of large language models (LLMs) capable of providing expert-level virology assistance. However, translating digital information into physical biological threats requires very specific tacit knowledge, and bioprocess engineering remain significant bottlenecks.

Recent empirical studies support this: large-scale randomized control trial by the non-profit Active Site (Hong et al., 2026) evaluated 153 novices in a BSL-2 laboratory setting. The study found no statistically significant difference in workflow completion rates between participants using frontier LLMs and those using standard internet access. As demonstrated historically, while AI can lower informational barriers, the "hands"—the physical execution of biological protocols—remain the hardest part. The same study nevertheless found modest uplift in intermediate progress, including faster cell-culture completion and higher rates of advancement through procedural steps, suggesting that stronger agentic systems or AI tightly coupled to laboratory robotics could erode exactly these execution bottlenecks over time.

Thus, participants struggled with long-horizon protocol execution, identifying the correct reagents and equipment from noisy inventories, recovering from mistakes, and acquiring tacit skills such as aseptic technique, manual dexterity, and troubleshooting in the presence of ambiguous physical feedback. We argue that these bottlenecks are precisely the things the current generation of scientific AI agents is designed to address, and that with better harnesses they may outperform web-based LLM use on some parts of the workflow while still falling short on embodied laboratory competence.

Our thesis is that the biosecurity impact of AI scientist agents depends less on generic model capability than on whether an actor can connect those agents to a reliable execution layer. We state this contribution as three claims:

1. **Capability claim:** AI scientist agents primarily improve the orchestration layer of planning, tool use, protocol generation, instrument control, and structured error handling–while tacit wet-lab competence, grounded physical perception, and ambiguous troubleshooting remain binding bottlenecks.

2. **Risk-prioritization claim:** Because these bottlenecks are unevenly distributed across actors, near-term biose-

[1]Independent Researcher, Berlin, Germany. Correspondence to: Dmitrii Magas <dmitrii@eamag.me>.

Accepted by ICML 2026 AI for Science Workshop.

curity concern should shift away from undifferentiated "LLM enables bioterrorism" narratives and toward actor–infrastructure pairs that can supply structured execution, especially well-funded non-state groups and some agricultural misuse pathways.

3. **Governance claim:** Effective governance should therefore focus not only on model access, but on the surrounding execution ecosystem: synthesis screening, cloud-lab and automation auditability, agricultural biosurveillance, and investment in AI-assisted defensive response.

2. What AI Scientist Agents Improve, and What They Still Do Not

In this paper, we use *AI scientist agent* to mean an AI system that goes beyond conversational assistance by autonomously or semi-autonomously coordinating multi-step scientific work through external tools, software, instruments, or robotic platforms. Concretely, these systems typically combine some subset of: (1) long-horizon planning, (2) tool or API invocation, (3) interaction with scientific software or instruments, (4) iterative adaptation based on intermediate results or error signals, and (5) partial delegation of execution rather than only recommendation.

2.1. Long-Horizon Execution Favors Structured Interfaces

The strongest current evidence for "AI scientist" agents is not that they replace wet-lab experts, but that they outperform standalone chat LLMs when experimental work is mediated through structured interfaces. Coscientist and ChemCrow show that tool-using systems can search documentation, invoke code, plan multistep syntheses, and interact with cloud-lab or robotic execution layers; AutoLabs similarly shows that multi-agent decomposition and guided self-correction improve robot-ready protocol generation relative to weaker baselines (Boiko et al., 2023; Bran et al., 2023; Panapitiya et al., 2025). By contrast, a broader evaluation of open-source autonomous research frameworks found that systems without strong execution scaffolds still fail to robustly complete end-to-end research tasks, indicating that long-horizon competence is highly conditional on tool access and structured feedback (Agrawal et al., 2026).

The best evidence for actual long-horizon *physical* execution comes from ORGANA: it executes multi-step chemistry workflows, including a 19-step plan and a larger electrochemistry workflow of 114 steps over roughly 130 minutes, while also reducing human interaction time relative to manual experimentation; CLAIRify similarly demonstrates multistep chemistry execution using constrained task-and-motion planning with online visual feedback (Darvish et al., 2024; Yoshikawa, 2024). The comparison is therefore layered rather than binary: agentic LLM systems improve coordination and procedural persistence relative to chat-only models, while embodied robotics remains the more credible route to sustained physical execution.

2.2. Grounding Reagents and Equipment Remains an Evidence Gap

The largest unresolved gap is not abstract reasoning but grounded laboratory disambiguation. Across the surveyed literature, there is little direct benchmarking of noisy inventory identification: synonymy, partial labels, missing consumables, packaging variation, and semi-structured bench layouts are rarely evaluated explicitly (Darvish et al., 2024; Leong et al., 2025; Mandal et al., 2025). ORGANA provides the closest demonstration, but even there the system grounds perceived objects partly through user interaction, reflecting the difficulty of operating in scenes that include transparent labware and ambiguous object state (Darvish et al., 2024). More generally, the self-driving-lab safety literature frames multimodal perception, active information gathering, and uncertainty-aware decisions as enabling technologies, not solved capabilities (Leong et al., 2025).

This matters because many wet-lab errors begin as labeling or state-estimation failures rather than high-level planning failures. A system that can write a plausible protocol but cannot reliably determine whether a reagent has already been diluted, whether two nearly identical tubes are interchangeable, or whether a vessel contains dissolved versus undissolved material is still dependent on human supervision. Current evidence therefore supports a weaker claim: AI agents can help structure action once the world is cleanly represented, but they do not yet autonomously produce that clean representation under normal laboratory conditions.

2.3. Error Recovery Is Better in Software Than in the Physical World

Mistake recovery is the area where agentic systems most clearly improve on static automation, but the gains are uneven. ChemCrow can revise procedures when a robotic platform returns structured warnings such as insufficient solvent or invalid actions, and multi-agent instrument-control studies similarly report better recovery when systems can consult memory, parse explicit error states, or escalate to a human checker (Bran et al., 2023; Mandal et al., 2025). AutoLabs shows the same pattern on protocol generation: iterative self-correction and reasoning improve numerical accuracy and step ordering, but the evaluation remains concentrated on generating better machine-readable procedures rather than recovering from off-nominal wet-lab execution (Panapitiya et al., 2025).

Physical-world recovery remains substantially harder. OR-

GANA can detect issues and alert users, with low miss rates in injected-error studies but still some false positives, and it explicitly relies on human troubleshooting when the run deviates from expectation (Darvish et al., 2024). CLAIRify closes the loop with online visual feedback and constrained replanning, yet the reported gains still fall short of open-ended recovery in cluttered, ambiguous laboratory environments (Yoshikawa, 2024). The practical takeaway is that agentic systems are currently strongest when failure is surfaced as structured software state; once recovery requires interpreting ambiguous physical feedback, the human operator remains the fallback layer.

2.4. Tacit Skill Remains the Hard Barrier

This distinction is especially important for biosecurity because tacit wet-lab skill remains the decisive bottleneck. In regulated aseptic processing, robotics can reduce contamination by removing direct human interventions and can improve traceability through detailed logging and standardized handling. But these gains come with substantial constraints around particle emission, cleanability, airflow disruption, decontamination, and validation; this is not the same thing as a system having learned expert aseptic technique in the open-ended sense relevant to biological weaponization (Tanzini et al., 2023).

More broadly, the robotics literature continues to emphasize that dexterity and ambiguous-feedback troubleshooting remain well below expert human performance. Current systems struggle with transparent-object perception, pose estimation, detecting whether material has dissolved, and manipulating filled containers without spills under tight constraints; safe performance increasingly depends on multimodal sensing, uncertainty quantification, and conservative intervention rules (Yoshikawa, 2024; Leong et al., 2025). The relevant comparison for biosecurity is therefore not whether an AI system can describe a protocol, but whether it can robustly reproduce the tacit, improvisational, and contamination-sensitive practices that make wet-lab work succeed.

2.5. Deployment Depends on Workflow Economics as Much as Capability

Many workflows are blocked less by whether they are physically automatable than by whether they are repeated often enough to justify protocol engineering. The main divide is not robot-easy versus robot-hard motions, but high-repeat versus low-repeat workflows: once a protocol recurs often enough, boxes and workcells can be worth the setup cost, whereas exploratory biology remains expensive to automate because each week perturbs the protocol again (Mahajan, 2026b). This helps explain why much of wet-lab automation remains concentrated in standardized tasks such as high-throughput screening, liquid handling, and other highly parameterized workflows.

The same heuristic also clarifies the field's directions: first focuses on translation layers that compile scientist intent into robot-ready instructions; second on hardware integration that connects boxes into workcells or cloud-lab platforms; a third on runtime intelligence that compensates for drift, calibration error, and other off-nominal conditions during execution (Mahajan, 2026b). Centralized cloud-lab models are attractive in this framework because batching work across users can raise utilization, lower per-experiment cost, and eventually justify upstream vertical integration of reagents and other inputs, although earlier cloud-lab efforts also show that automation fails to matter when logistics and debugging friction remain with the customer. The first major effects for biosecurity of AI scientist agents are more likely to appear in centralized, repetitive, or industrial settings than in arbitrary low-volume wet labs. This division aligns with the Automated Laboratory Security Tier framework, which categorizes automated facilities based on their latent physical capabilities and remote-access risk profiles (Smith et al.).

2.6. Closed-Loop Biology and Protein Language Models

In biological design, the integration of protein language models (pLMs) with active learning and automated biofoundries enables closed-loop Design–Build–Test–Learn (DBTL) cycles that can autonomously generate and evaluate new biological variants (Wang et al., 2026). Systems that couple pLMs (such as ESM-3 or ProteinMPNN) with automated liquid handling and screening assays have been shown to accelerate the search for functional protein variants, drastically reducing the number of manual design iterations required to achieve specific fitness goals (Zhang et al., 2025; Wang et al., 2026).

However, this integration does not bypass the fundamental physical constraints of wet-lab biology. Although models can design sequences in silico, functional validation still requires physical execution—expressing, purifying, and testing the designed proteins in a wet lab. Consequently, the bottleneck in biological agency is relocated further toward the physical validation layer. While closed-loop systems could in principle assist in optimizing protein designs to evade sequence-similarity screening (as highlighted by the red-teaming results of (Wittmann et al., 2025)), their real-world deployment remains tightly constrained by the physical execution layer. This relationship between digital design speed and physical execution constraints underpins the threat profiles of actors capable of supplying structured lab environments.

2.7. From Agent Capabilities to Threat Profiles

The key implication for biosecurity is that the transition from chat LLMs to AI scientist agents does not produce a uniform reduction in barriers. The systems surveyed above are strongest when tasks can be decomposed into structured digital planning, API calls, instrument control, or repetitive robot-ready procedures. They are weakest when success depends on tacit dexterity, local physical judgment, and ambiguous environmental feedback. In other words, current agents mostly strengthen the *orchestration layer* of biology, not the full stack of messy wet-lab execution.

That asymmetry changes which actors benefit: a lone actor operating in an improvised setting still faces the same hard physical bottlenecks that limited earlier bioterror efforts. By contrast, actors with the capital to build or rent automated workcells, cloud-lab capacity, or other structured execution environments can convert agentic planning into real operational throughput. Agricultural attacks also move upward in concern because the required execution is less exacting and the surrounding surveillance is weaker. This capability-to-resource mapping is the basis for the ranking in Section 3: the highest-priority threats are the ones for which AI scientist agents can be paired with an execution environment that masks their physical weaknesses.

3. Ranking Biosecurity Threat Pathways by Marginal Risk Increase from Current AI Scientist Agents

We classify potential biosecurity threats into four tiers, ranked by *near-term concern* rather than by maximum possible harm. The ranking is designed to capture a specific question: which actor classes can exploit the digital strengths of current AI scientist agents without being stopped by their remaining physical weaknesses? To make that judgment more auditable, we translate the heuristic ranking into a transparent scoring rubric using a simple multi-criteria decision analysis (MCDA).

Each threat category receives ordinal scores from 1 to 5 on four dimensions:

- **Technical Execution Barrier:** how easy it is for the actor to convert AI-assisted planning into real biological execution. A higher score means the effective barrier is lower.
- **Resource/Capital Barrier:** how difficult it is for the actor to access the required equipment, personnel, automation, and infrastructure. A higher score means the effective barrier is lower.
- **Defense Porousness:** how easy it is for the pathway to slip through current screening, surveillance, and response systems. A higher score means existing defenses are weaker or easier to evade.
- **Expected Societal/Economic Impact:** the plausible scale of near-term damage if the pathway succeeds. A higher score means greater expected harm.

This MCDA is a structured heuristic with a purpose to make the ranking's assumptions legible and debatable rather than to estimate exact probabilities or losses. The scale is concern-oriented: 5 denotes greater near-term concern, so for the two barrier dimensions a higher score means the barrier is relatively low for that pathway under current AI and automation trends. We then compute a Risk Priority Score as the unweighted average of the four criterion scores:

$$\text{Risk Priority Score} = \frac{s_1 + s_2 + s_3 + s_4}{4},$$

where s_1 through s_4 correspond to Technical Execution Barrier, Resource/Capital Barrier, Defense Porousness, and Expected Societal/Economic Impact, respectively.

This does not estimate exact event probabilities but forces the ranking to expose which assumptions are doing the work. Under this lens, nation-states still have the greatest absolute destructive capacity, but current AI scientist agents change the margin less for actors that already possess industrial resources than for actors who can exploit specific gaps in synthesis screening or biosurveillance once execution is partially automated. Table 1 summarizes this framework.

3.1. Tier 4 (Lowest Risk): State Actors

Ranking state actors lowest does *not* mean they are harmless. Rather, it means they are **the least changed by the current AI trajectory**. Historical programs show that states can solve the engineering problem, but usually only by mobilizing vast institutional capacity: the Soviet Biopreparat program eventually achieved industrial-scale production, Iraq's program remained far less effective, and the U.S. terminated its offensive program in 1969 (Smithson, 2025; Nixon, 1969). The relevant point is therefore comparative. AI does not create a qualitatively new access path for states in the way it might for smaller actors, and states still possess many alternative means of coercion and mass violence.

Under the MCDA, we assign:

- **Technical Execution Barrier = 2:** states can execute complex biological programs, but current AI does not dramatically lower that barrier for them because they already possess expertise and institutional capacity.
- **Resource/Capital Barrier = 1:** in marginal-AI terms, this pathway is not newly unlocked by reduced capital requirements; states already operate at high resource levels.

Table 1. Semi-Quantitative MCDA for Near-Term Biosecurity Threat Prioritization

Rank	Threat Category	Risk Priority Score	Why It Lands Here
4 (Lowest)	State Offensive Programs	2.00	AI adds limited marginal capability to actors that already possess industrial-scale means; the pathway is dangerous in absolute terms but not highly shifted by current AI.
3 (Low)	Individual or Small-Cell Bioterrorism	2.25	Tacit wet-lab skill and dissemination barriers remain binding, though smaller-scale harm is still conceivable if those barriers are partially cleared.
2 (High)	Agroterrorism	4.00	The technical bar is lower, surveillance is weaker, host populations are dense, and the economic downside can still be nationally severe.
1 (Highest)	Well-Funded Non-State Actors Targeting Humans	4.50	Structured private-lab or automated environments can convert agentic planning into operational throughput, while screening chokepoints are weakened by fragment assembly, sequence redesign, and eventual benchtop synthesis.

- **Defense Porousness = 1:** state-scale programs face substantial intelligence, treaty, and geopolitical scrutiny, even if those safeguards are imperfect.
- **Expected Impact = 4:** the absolute potential harm is very high.

This yields **RPS of 2**, which reflects limited *marginal* change from current AI scientist agents rather than low absolute destructive capacity.

3.2. Tier 3 (Low Risk): Individual Actors

Lone actors or very small cells remain a lower-probability pathway because the hard part is not retrieving recipes but clearing tacit wet-lab and dissemination bottlenecks without access to a structured execution environment. A useful upper-bound case is Aum Shinrikyo: despite large financial resources, dedicated facilities, and scientifically trained personnel, its biological attacks failed because of strain choice, low concentrations, fermentation failures, and poor aerosolization (Takahashi et al., 2004; Bleek, 2011). If a comparatively well-resourced non-state organization could not reliably clear these barriers, current AI scientist agents are unlikely to collapse them for a true lone actor in the near term, especially given the limited uplift seen in wet-lab execution studies (Hong et al., 2026).

Accordingly, we assign:

- **Technical Execution Barrier = 2:** AI may modestly help with planning and protocol following, but tacit wet-lab execution remains a major bottleneck.
- **Resource/Capital Barrier = 2:** lone actors still face meaningful constraints in equipment, materials, and safe operating environments.
- **Defense Porousness = 2:** a small actor can sometimes avoid attention, but procurement, synthesis, and deployment remain difficult to conceal reliably.
- **Expected Impact = 3:** plausible harm exists, but the upper bound is lower than for the top tiers.

This yields **RPS of 2.25** and places them above states in near-term concern because smaller-scale attacks require less institutional coordination, but well below the higher tiers because tacit physical execution remains the main bottleneck.

3.3. Tier 2 (High Risk): Agroterrorism

Agroterrorism ranks high because the technical bar for causing serious damage is lower, the surveillance chain is weaker, and the economic consequences can still be nationally severe. Foot-and-mouth disease and other livestock pathogens are highly contagious, modern agriculture concentrates genetically similar hosts at high density, and dissemination can be comparatively unsophisticated (Brown et al., 2022; Haley, 2019). Detection is also structurally fragile: the 2024–2025 H5N1 dairy outbreak spread for months before recognition, illustrating how large herd sizes, delayed symptom recognition, and reporting frictions can slow response (Nguyen et al., 2024). Simulation studies suggest that severe FMD agroterrorism scenarios in the United States could generate losses ranging from tens to hundreds of billions of dollars (Oladosu et al., 2013).

We therefore assign:

- **Technical Execution Barrier = 5:** the biological and operational bar is comparatively lower than for sophisticated human-targeted attacks.
- **Resource/Capital Barrier = 4:** substantial resources help, but the pathway does not require the same level of precision infrastructure as the highest-end human-targeted scenarios.

- **Defense Porousness = 4:** agricultural surveillance and reporting remain weaker and slower than many human-health monitoring systems.
- **Expected Impact = 3:** economic disruption can be severe at regional or national scale, even without the maximum casualty potential of Tier 1.

This yields **RPS of 4**, which reflects a pathway that is comparatively easy to operationalize, relatively weakly surveilled, and capable of generating major economic damage even without the full sophistication required for human-targeted attacks.

3.4. Tier 1 (Highest Risk): Well-Funded Non-State Actors Targeting Humans

A highest near-term concern is a well-funded non-state group operating a private lab or other structured automated environment. This is the actor class most able to convert agentic planning into operational throughput: the lab supplies the disciplined execution layer that current systems still need. AI and synthesis technology are weakening a defense model built around centralized providers and sequence-similarity screening. Three vulnerabilities matter most:

1. **Short-fragment assembly:** Even as federal guidance moves toward shorter screening windows, reconstructing large genomes from smaller ordered fragments remains a real pathway; the horsepox synthesis result is the clearest proof of principle (Noyce et al., 2018).
2. **AI-assisted sequence redesign:** Recent red-teaming work showed that generative protein-design tools can produce toxin variants that evade existing screening software, implying that sequence-only filters will need continual updating rather than one-time deployment (Wittmann et al., 2025).
3. **Benchtop synthesis as an emerging risk:** Current benchtop devices are still limited, but policy analyses suggest that more capable enzymatic systems could erode the centralized chokepoint further over the next decade (Langenkamp, 2024). Additionally, the rapid development of Enzymatic DNA Synthesis is driving a transition to decentralized benchtop devices. As these devices become capable of reliably printing longer double-stranded DNA fragments (up to several kilobases) in non-expert laboratory environments, they bypass the traditional sequence-and-customer screening chokepoints historically maintained by centralized providers (Anguzu, 2025). This decentralized synthesis capability allows actors to manufacture genetic material on-site with minimal oversight, significantly eroding the centralized screening model.

The main implication is therefore a weakening preventative architecture, not proof that fully *de novo* pathogen design is already routine.

Under the MCDA, we assign:

- **Technical Execution Barrier = 4:** this actor class can pair AI with disciplined execution environments, automation, and specialized staff.
- **Resource/Capital Barrier = 4:** the required resources are substantial but still much more accessible than state-scale programs.
- **Defense Porousness = 5:** centralized screening and synthesis chokepoints are increasingly stressed by fragment assembly, redesign, and prospective benchtop synthesis.
- **Expected Impact = 5:** successful human-targeted attacks could produce very high casualties and broad societal disruption.

This yields **RPS of 4.50**, making it the highest-priority tier because it combines access to structured execution environments with increasingly porous synthesis and screening chokepoints and the possibility of very high human impact.

3.5. Sensitivity to Capability Jumps

The MCDA framing also makes it easier to discuss discontinuous capability shifts. Suppose an AI breakthrough in robotic fluid manipulation, contamination-aware handling, and error recovery substantially reduced the tacit execution bottleneck for lone actors. If Tier 3's Technical Execution score rose from 2 to 4 and its Resource/Capital score rose from 2 to 3, its Risk Priority Score would increase from 2.25 to 3.00 even before changing any defense-evasion assumptions. If the same breakthrough also made procurement and protocol execution easier to conceal, pushing Defense Porousness from 2 to 3, Tier 3 would rise further to 3.25 and begin to overlap with today's Tier 2 profile.

The broader implication is that lone actors are ranked below agroterrorism and well-funded non-state private-lab misuse mainly because tacit physical execution and access to structured infrastructure remain binding constraints. If AI materially weakens those constraints, their priority could rise quickly. This dynamic assessment of how technological capability jumps alter the threat priority matrix aligns with broader contemporary evaluations of biosecurity's future trajectory (Mahajan, 2026a). It also suggests that agroterrorism is often overlooked in AI-biosecurity debates, despite combining lower execution barriers, weaker surveillance, and the potential for very large economic and societal disruption.

4. Do Current Defenses Match the Ranked Threats?

Section 3 ranks threats by where AI scientist agents can be paired with execution environments. The defense question is therefore whether current safeguards are strongest at those same bottlenecks. In practice, the stack is uneven: synthesis screening is the clearest upstream chokepoint for Tier 1, agricultural traceability remains fragile for Tier 2, cloud-lab governance is immature, and downstream detection often outruns response.

4.1. The Failure of Air Monitoring

The U.S. BioWatch program illustrates the gap between biosurveillance ambition and operational utility. The deployed system relies on daily filter collection and laboratory analysis that can take on the order of 36 hours, and DHS has reported more than seven million tests with 149 detections of naturally occurring agents but no confirmed attack detections (Council et al., 2011). That record does not prove that aerosol monitoring is impossible, but it does suggest that the current air-monitoring stack has provided limited real-world defensive value. For the ranking here, this makes broad air monitoring a weak anchor: it does not close Tier 1's upstream execution path and is too slow to carry downstream containment.

4.2. Wastewater Screening: Detection vs. Defense

By contrast, wastewater surveillance has shown clear value as an early-warning signal. CDC's National Wastewater Surveillance System can detect community-level trends before they are obvious in clinical data, while pilot metagenomic programs such as the Nucleic Acid Observatory and models such as METAGENE-1 point toward broader-spectrum detection of known and unknown agents (Kirby, 2021; Liu et al., 2025).

However, if the detected threat is novel, early sequencing does not automatically translate into an immediately deployable countermeasure. Even CEPI's rapid-response framework assumes substantially longer timelines—roughly 200 to 230 days—for novel pathogens without an already-prepared vaccine candidate (Kim et al., 2025). Wastewater surveillance is therefore a useful cross-cutting warning layer, but not a substitute for Tier 1 prevention or Tier 2 agricultural traceability.

4.3. Synthesis Screening Is the Main Upstream Preventive Chokepoint

If wastewater is the best current *detection* layer, nucleic-acid screening is the best current *prevention* layer. The 2024 OSTP Framework for Nucleic Acid Synthesis Screening and downstream NIH implementation policies treat synthetic nucleic acids as a control point: NIH intramural researchers may only order from providers or benchtop equipment manufacturers that attest to adherence, and providers are expected to screen sequences of concern, verify suspicious customers, report potentially illegitimate orders, retain records, and maintain cybersecurity (National Institutes of Health, Office of Intramural Research, 2024; Office of Science and Technology Policy, 2024). This does not solve the whole problem, but it does create a real chokepoint where misuse can be interrupted before wet-lab execution begins.

The weakness is that screening remains only as strong as its coverage, standardization, and enforcement. RAND's 2024 review of commercial nucleic-acid synthesis emphasizes that screening protocols vary across providers, providers do not currently have a built-in mechanism to share information about suspicious customers or split orders, and government still lacks a full conformity-assessment mechanism for screening systems (Crawford et al., 2024). This is the main Tier 1 chokepoint, but it still needs substantial strengthening. While sequence-only screening is a vital barrier when orders must go through major commercial foundries, it becomes a porous line of defense once synthesis capability is distributed directly to the laboratory benchtop.

4.4. Agricultural Defense Depends on Traceability and Preparedness, Not Just Detection

Because agroterrorism is one of the highest-ranked threats in this paper, the agricultural defense stack deserves to be treated separately. APHIS's own materials make clear that this layer has multiple components: Animal Disease Traceability is meant to establish where diseased and at-risk animals are, where they have been, and when they were there, precisely to reduce response time and the number of animals implicated in an outbreak (U.S. Department of Agriculture, Animal and Plant Health Inspection Service, 2026b); Foreign Animal Disease Preparedness and Response (FAD PReP) is the U.S. framework for coordinated preparedness and response to foreign animal disease incidents (U.S. Department of Agriculture, Animal and Plant Health Inspection Service, 2026a).

These depend heavily on noticing, reporting, tracing, and then acting quickly across distributed agricultural systems. That fragility is why agroterrorism ranks highly: Tier 2 attacks need not bypass every defense if detection, reporting, and traceability already operate with delay and uneven incentives.

4.5. Pathogen-Agnostic Indoor Air Measures Are a Distinct Mitigation Layer

Not all defenses depend on identifying the pathogen first: indoor-air measures are important precisely because they can reduce airborne transmission regardless of whether the

agent is known, novel, or only suspected. CDC and NIOSH explicitly frame improved ventilation, filtration, and germicidal ultraviolet (GUV/UVGI) as layered measures for reducing exposure to airborne pathogens (Centers for Disease Control and Prevention, National Institute for Occupational Safety and Health, 2024b;a; Centers for Disease Control and Prevention, 2025). This matters because it shifts some of the defense burden away from perfect detection and toward reducing indoor transmission opportunities in advance.

The main constraint here is not whether these measures work in principle, but whether they are deployed broadly enough to matter at population scale. For that reason, pathogen-agnostic indoor air is a cross-tier harm reducer: it does not block attempts upstream, but it still matters when prevention or early detection fails.

4.6. Cloud-Lab and Automation Governance Is Becoming Part of the Defense Stack

A final layer is emerging specifically because of AI scientist agents: governance of execution environments. RAND's 2025 assessment of cloud labs argues that remotely operated automated laboratories increase accessibility and flexibility for legitimate science, but that the same remote and automated features can also create new biosecurity vulnerabilities and could enable malicious actors if governance is weak (Lee et al., 2025). This means that cloud-lab oversight, auditability, access control, and anomaly detection should be treated as part of biosecurity defense architecture rather than as a separate technology-policy issue. If the major near-term risk comes from actors that can pair AI planning with structured execution environments, then governance must extend to the automated laboratories, service providers, and remote execution platforms that turn plans into physical workflows.

Taken together, the defenses are least settled where the ranking assigns most concern. Prevention is strongest where synthesis remains centralized, weaker where agricultural systems are distributed, and still emerging where AI agents meet automated execution platforms. This asymmetry helps explain why Tier 1 and Tier 2 outrank lone actors in near-term concern.

5. Conclusion

Biosecurity debates about AI-driven science should focus less on "AI in biology" in the abstract and more on the actor–infrastructure combinations that make AI scientist agents physically consequential. Current systems are strongest at software-mediated planning, coordination, tool use, and instrument-facing workflows, but they do not eliminate the tacit wet-lab and dissemination bottlenecks that have historically constrained small actors. The most credible near-term concerns are structural: actors that can pair agentic systems with structured execution environments, and misuse pathways such as agroterrorism that combine lower execution barriers with weaker surveillance.

This framing suggests three priorities. First, policy and evaluation should distinguish chat-style informational assistance from execution-coupled agentic systems, because the latter are what materially change real-world scientific throughput. Second, governance should expand beyond model access alone to the execution layer around autonomous labs, including synthesis screening, cloud-lab auditability, agricultural surveillance, and deployment of pathogen-agnostic defenses. Third, the research agenda should treat AI also as a defensive tool, especially for biosurveillance, metagenomic interpretation, and rapid countermeasure design.

Impact Statement

This position paper presents a framework for evaluating biosecurity risks in the context of machine learning and laboratory automation. By clarifying which actor-infrastructure combinations are most credibly shifted by current AI scientist agents, such as agroterrorism and well-funded private-lab misuse, and which remain more constrained, such as lone-actor de novo pathogen engineering, the paper aims to guide policymakers, laboratory operators, and AI researchers toward more targeted governance. Potential benefits include stronger oversight of DNA synthesis and automated laboratory infrastructure, improved biosurveillance, and greater investment in AI-assisted defensive response. Because the paper discusses misuse-relevant pathways, it also carries a risk of being misread as overstating current offensive capability; for that reason, it deliberately avoids operational detail and emphasizes uncertainty, prevention, and defensive preparedness.

References

Agrawal, S., Anadkat, H. B., Athimoolam, K. K., Bhardwaj, H., Chowdhury, T., Gao, S., Kamat, P., Makwana, V., Shariff, M. H., Badkul, A., Xie, L., and Sinitskiy, A. V. Can ai conduct autonomous scientific research? case studies on two real-world tasks. *bioRxiv*, 2026. URL https://api.semanticscholar.org/CorpusID:284532958.

Anguzu, S. Preventing biosecurity risks posed by next-gen benchtop DNA synthesizers. *Health Economics and Management Review*, 6(1):146–159, 2025. doi: 10.61093/hem.2025.1-09.

Bleek, P. C. Revisiting aum shinrikyo: New insights into the most extensive non-state biological weapons program to date, 2011.

Boiko, D. A., MacKnight, R., Kline, B. C., and Gomes, G. Autonomous chemical research with large language models. *Nature*, 624:570 – 578, 2023. URL https://api.semanticscholar.org/CorpusID:266432059.

Bran, A. M., Cox, S., White, A. D., and Schwaller, P. Chemcrow: Augmenting large-language models with chemistry tools. 2023. URL https://api.semanticscholar.org/CorpusID:271293795.

Brown, E., Nelson, N., Gubbins, S., and Colenutt, C. Airborne transmission of foot-and-mouth disease virus: A review of past and present perspectives. *Viruses*, 14, 2022. URL https://api.semanticscholar.org/CorpusID:248742173.

Centers for Disease Control and Prevention. Taking steps for cleaner air for respiratory virus prevention. https://www.cdc.gov/respiratory-viruses/prevention/air-quality.html, 2025. Accessed 2026-04-19.

Centers for Disease Control and Prevention, National Institute for Occupational Safety and Health. Improving air cleanliness. https://www.cdc.gov/niosh/ventilation/prevention/air-cleanliness.html, 2024a. Accessed 2026-04-19.

Centers for Disease Control and Prevention, National Institute for Occupational Safety and Health. About germicidal ultraviolet (guv). https://www.cdc.gov/niosh/ventilation/germicidal-ultraviolet/index.html, 2024b. Accessed 2026-04-19.

Council, N. R., on Life Sciences, B., on Chemical Sciences, B., on Health Sciences Policy, B., on Effectiveness of National Biosurveillance Systems, C., BioWatch, and the Public Health System. *BioWatch and public health surveillance: Evaluating systems for the early detection of biological threats: Abbreviated version*. National Academies Press, 2011.

Crawford, F. W., Webster, K., Epstein, G. L., Roberts, D., Fair, J., and Nevo, S. *Securing Commercial Nucleic Acid Synthesis*. RAND Corporation, Santa Monica, CA, 2024. doi: 10.7249/RRA3329-1.

Darvish, K., Skreta, M., Zhao, Y., Yoshikawa, N., Som, S., Bogdanovic, M., Cao, Y., Hao, H., Xu, H., Aspuru-Guzik, A., Garg, A., and Shkurti, F. Organa: A robotic assistant for automated chemistry experimentation and characterization. *ArXiv*, abs/2401.06949, 2024. URL https://api.semanticscholar.org/CorpusID:266998923.

Haley, M. Fields of danger: The looming threat of agroterrorism on the united states' agriculture. *Journal of Biosecurity, Biosafety, and Biodefense Law*, 10, 2019. URL https://api.semanticscholar.org/CorpusID:201060220.

Hong, S. Z., Kleinman, A., Mathiowetz, A., Howes, A., Cohen, J., Ganta, S., Letizia, A., Liao, D., Pahari, D., Roberts-Gaal, X., Righetti, L., and Torres, J. Measuring mid-2025 llm-assistance on novice performance in biology, 2026. URL https://arxiv.org/abs/2602.16703.

Kim, J., Sabet-Azad, R., Patel, D., Malin, G., Askary, S. H., and Särnefält, A. Fast-tracking vaccine manufacturing: Cepi's rapid response framework for the 100 days mission. *Vaccines*, 13(8), 2025. ISSN 2076-393X. doi: 10.3390/vaccines13080849. URL https://www.mdpi.com/2076-393X/13/8/849.

Kirby, A. E. Using wastewater surveillance data to support the covid-19 response—united states, 2020–2021. *MMWR. Morbidity and Mortality Weekly Report*, 70, 2021.

Langenkamp, M. Securing benchtop DNA synthesizers. Policy report, Institute for Progress, December 2024. URL https://progress.institute/securing-benchtop-dna-synthesizers/.

Lee, J., Persaud, B., Del Castello, B., Berke, A., and Zilgalvis, G. Documenting cloud labs and examining how remotely operated automated laboratories could enable bad actors. Technical Report PE-A3851-1, RAND Corporation, 2025. URL https://www.rand.org/pubs/perspectives/PEA3851-1.html.

Leong, S. X., Griesbach, C. E., Zhang, R., Darvish, K., Zhao, Y., Mandal, A., Zou, Y., Hao, H., Bernales, V., and Aspuru-Guzik, A. Steering towards safe self-driving laboratories. *Nature Reviews Chemistry*, 9:707 – 722, 2025. URL https://api.semanticscholar.org/CorpusID:280685714.

Liu, O., Jaghouar, S., Hagemann, J., Wang, S., Wiemels, J., Kaufman, J., and Neiswanger, W. METAGENE-1: Metagenomic foundation model for pandemic monitoring. *arXiv preprint arXiv:2501.02045*, 2025. URL https://arxiv.org/abs/2501.02045.

Mahajan, A. Reasons to be pessimistic (and optimistic) on the future of biosecurity. https://www.owlposting.com/p/reasons-to-be-pessimistic-and-optimistic, March 2026a. Owl Posting, published March 16, 2026; accessed 2026-04-19.

Mahajan, A. Heuristics for lab robotics, and where its future may go. https://www.owlposting.com/p/heuristics-for-lab-robotics-and-where, February 2026b. Owl Posting essay. Accessed: 2026-04-07.

Mandal, I., Soni, J., Zaki, M., Smedskjaer, M. M., Wondraczek, K., Wondraczek, L., Gosvami, N. N., and Krishnan, N. A. Evaluating large language model agents for automation of atomic force microscopy. *Nature Communications*, 16(1):9104, 2025.

National Institutes of Health, Office of Intramural Research. Policy on the ordering or provision of synthetic nucleic acids in the irp. https://oir.nih.gov/sourcebook/ethical-conduct/special-research-considerations/policy-ordering-or-provision-synthetic-nucleic-acids-irp, 2024. Accessed 2026-04-19.

Nguyen, T.-Q., Hutter, C. R., Markin, A., Thomas, M. N., Lantz, K., Killian, M. L., Janzen, G. M., Vijendran, S., Wagle, S., Inderski, B., Magstadt, D. R., Li, G., Diel, D. G., Frye, E. A., Dimitrov, K. M., Swinford, A. K., Thompson, A. C., Snevik, K. R., Suarez, D. L., Spackman, E., Lakin, S. M., Ahola, S. C., Johnson, K. R., Baker, A. L. V., Robbe-Austerman, S., Torchetti, M. K., and Anderson, T. K. Emergence and interstate spread of highly pathogenic avian influenza a(h5n1) in dairy cattle. *bioRxiv*, 2024. URL https://api.semanticscholar.org/CorpusID:269588393.

Nixon, R. Statement on chemical and biological defense policies and programs. *The American Presidency Project*, 1969.

Noyce, R. S., Lederman, S., and Evans, D. H. Construction of an infectious horsepox virus vaccine from chemically synthesized DNA fragments. *PloS one*, 13(1):e0188453, 2018.

Office of Science and Technology Policy. Framework for nucleic acid synthesis screening. Technical report, Executive Office of the President of the United States, 2024. URL https://www.whitehouse.gov/wp-content/uploads/2024/04/Nucleic-Acid_Synthesis_Screening_Framework.pdf.

Oladosu, G., Rose, A., and Lee, B. Economic impacts of potential foot and mouth disease agroterrorism in the usa: A general equilibrium analysis. *J. Bioterrorism Biodefense*, 12:1–9, 2013.

Panapitiya, G., Saldanha, E., Job, H., and Hess, O. Autolabs: Cognitive multi-agent systems with self-correction for autonomous chemical experimentation. *ArXiv*, abs/2509.25651, 2025. URL https://api.semanticscholar.org/CorpusID:281681326.

Smith, M. D., Sebastian Hanke, M., Moritz, R. L., Gillum, D., and Pannu, J. Automated laboratory security tiers: A framework for evaluating and mitigating biosecurity risks from latent capabilities. *Frontiers in Microbiology*, 17:1832401.

Smithson, A. E. Factors in illicit bioweapons programs: case studies of the soviet union/russia and iraq. *Frontiers in Political Science*, 7:1654084, 2025.

Takahashi, H., Keim, P., Kaufmann, A. F., Keys, C., Smith, K. L., Taniguchi, K., Inouye, S., and Kurata, T. Bacillus anthracis bioterrorism incident, Kameido, Tokyo, 1993. *Emerging Infectious Diseases*, 10(1):117, 2004.

Tanzini, A., Ruggeri, M., Bianchi, E., Valentino, C., Vigani, B., Ferrari, F., Rossi, S., Giberti, H., and Sandri, G. Robotics and aseptic processing in view of regulatory requirements. *Pharmaceutics*, 15(6):1581, 2023.

U.S. Department of Agriculture, Animal and Plant Health Inspection Service. Foreign animal disease preparedness and response. https://www.aphis.usda.gov/animal-emergencies/fadprep, 2026a. Accessed 2026-04-19.

U.S. Department of Agriculture, Animal and Plant Health Inspection Service. Animal disease traceability. https://www.aphis.usda.gov/livestock-poultry-disease/traceability, 2026b. Accessed 2026-04-19.

Wang, D., Huot, M., Zhang, Z., Jiang, K., Shakhnovich, E. I., and Esvelt, K. M. Without safeguards, AI-Biology integration risks accelerating future pandemics. *Frontiers in Microbiology*, 16:1734561, 2026. doi: 10.3389/fmicb.2025.1734561. URL https://doi.org/10.3389/fmicb.2025.1734561.

Wittmann, B. J., Alexanian, T., Bartling, C., Beal, J., Clore, A., Diggans, J., Flyangolts, K., Gemler, B. T., Mitchell, T., Murphy, S. T., et al. Strengthening nucleic acid biosecurity screening against generative protein design tools. *Science*, 390(6768):82–87, 2025.

Yoshikawa, N. *Towards automated robotic chemical experiments*. PhD thesis, University of Toronto (Canada), 2024.

Zhang, Q., Chen, W., Qin, M., Wang, Y., Pu, Z., Ding, K., et al. Integrating protein language models and automatic biofoundry for enhanced protein evolution. *Nature Communications*, 16:1553, 2025. doi: 10.1038/

s41467-025-56751-8. URL https://doi.org/10.1038/s41467-025-56751-8.