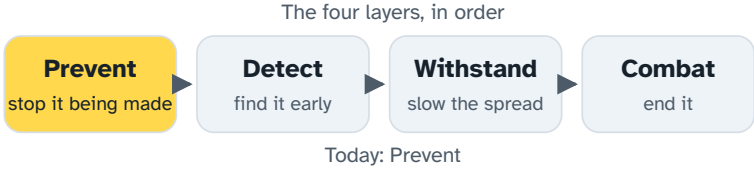
Day 2: Preventing Engineered Pandemics (Prevent)

Tuesday 27 October · discussion 3:00 to 4:50 pm New York

THE DAY MAP

How do you stop an engineered pandemic before the first person is infected?



Your deep-dive: page 2.7. DNA synthesis screening, know-your-customer checks and cloud labs. The closest thing in the course to gating something that can heal or harm. Its readings come first below; the other options follow.

30 readings here, about 235 minutes of reading. Discussion Tuesday 27 October, 3:00 to 4:50 pm New York, with Conor Spence.

2.1-1	A.I. Bots Told Scientists How to Make Biological Weapons original paper →	2 + 0 min

2.1-2	When Should We Worry About AI Being Used to Design a Pathogen? original paper →	3 + 8 min
2.1-3	To Forecast AI's Impact on Biosecurity, We Asked: Why are Attacks So Rare? original paper →	3 + 0 min
2.1-4	Horsepox Synthesis: A case of the unilateralist's curse? original paper →	3 + 8 min
2.1-5	The role of evals in the biorisk evidence hierarchy original paper →	3 + 9 min
2.2-1	Biological Weapons Convention original paper →	2 + 13 min
2.2-2	Why do we take LLMs seriously as a potential source of biorisk? original paper →	2 + 9 min
2.2-3	Evaluating Safety and Security in Biological Research original paper →	3 + 10 min
2.2-4	How to stop bioterrorists from buying dangerous DNA original paper →	2 + 7 min
2.7-1	How to Secure the DNA Supply Chain original paper →	3 + 14 min

2.7-2	Know your scientist: KYC as biosecurity infrastructure original paper →	3 + 14 min
2.7-3	No, there are not hundreds of cloud labs in biology original paper →	2 + 7 min
2.7-4	Customer Screening Resources original paper →	3 + 0 min
2.4-1	Expanding on Future Biological Weapons Convention Verification: An Interview Dr. Piers Millett original paper →	2 + 8 min
2.4-2	Tracing engineered biothreats with AI forensics: Five steps to improve attribution original paper →	2 + 8 min
2.4-3	The biosecurity benefits of genetic engineering attribution original paper →	3 + 0 min
2.4-4	Project BEACON: Using OSINT methods in BWC verification original paper →	3 + 6 min
2.4-5	Adherence to and Compliance with Arms Control, Nonproliferation, and Disarmament Agreements and Commitments (2025) original paper →	3 + 0 min

2.5-1	Confronting risks of mirror life original paper →	3 + 0 min
2.5-2	Technical Report on Mirror Bacteria: Feasibility and Risks original paper →	2 + 0 min
2.5-3	James Smith on why he quit everything to work on a biothreat nobody had heard of original paper →	3 + 0 min
2.5-4	Mirror Biology Dialogues Fund (MBDF) website original paper →	3 + 0 min
2.5-5	The Deep Lunacy that was (almost) 'DEEP VZN' original paper →	3 + 8 min
2.5-6	Behind the cancellation of a $125 million US virus hunting program original paper →	3 + 0 min
2.5-7	Recipes for Future Plagues original paper →	3 + 0 min
2.5-8	Senate Foreign Relations Committee letter on USAID's DEEP VZN program original paper →	2 + 0 min
2.5-9	Why Do Exceptionally Dangerous Gain-of-Function Experiments in Influenza? original paper →	3 + 8 min

2.6-1	Preliminary Taxonomy of AI-Bio Misuse Mitigations original paper →	2 + 9 min
2.6-2	AIxBio Research Hub original paper →	3 + 9 min
2.6-3	Deep Ignorance: Filtering Pretraining Data Builds Tamper-Resistant Safeguards into Open-Weight LLMs original paper →	3 + 0 min

WHAT THE COURSE ASKS YOU TO DO

2.3 Delay patient zero: Deep-dives

Pick 1 of the following 4 interventions to dive deeper into. We've given you some resources on the following pages as a starting point.

Aim to answer the following general questions and have a look at the deep dive pages for more detailed, intervention-specific questions: - What are the intended goals of this intervention? - What is the current state of this intervention, and to what extent is it achieving its goals? - What are the most critical holes in this intervention? - What are they key challenges to implementing a more robust version of this intervention, and how could you overcome them?

You will bring the results of your research to share in your group discussion so make sure to take notes and share great explainers or visualizations you have found.

Option 1: Deter - Reducing intent Disincentivizing the development of biological weapons.

Topics you will research: - The Biological Weapons Convention (BWC) - Genetic engineering attribution - Open Source Intelligence (OSINT)

Option 2: Deny - Regulating discovery Some knowledge becomes dangerous the moment it is created, so society needs to decide what is researched and who should be allowed to access the findings.

Topics you will research: - The de-facto ban on mirror life research - 'Dual-Use Research of Concern' (DURC), 'Gain-of-function' (GoF) research and 'Pathogens with Enhanced Pandemic Potential' (PEPP) - Virus hunting programs

Option 3: Deny - Implementing AI model-level safeguards General Purpose AI (GPAI) beats experts on knowledge tests and state-of-the art Biological AI Models (BAIM) make biology increasingly programmable. Preventing bad actors from using them to do harm while preserving their usefulness for beneficial research requires good safeguards.

Topics you will research: - Pre-training filtering - Fine-tuning + weight locking - Output filtering

Option 4: Deny - Safeguarding research infrastructure
A genome sequence is only a threat if someone can turn it into a functional virus. We can prevent this by blocking key chokepoints.

Topics you will research: - Nucleic acid synthesis - Know your customer (KYC) regulation - Cloud labs and Contract Research Organizations (CROs)

THE READINGS

2.1-1 · FURTHER · COURSE TIME 15 MIN

A.I. Bots Told Scientists How to Make Biological Weapons

Gabriel J.X. Dance · 2026

Read the original paper → · the course assigns: whole piece, from a detailed structured account of the article rather than its original text

> **Biosecurity experts hired to pressure-test chatbots before release handed the New York Times more than a dozen transcripts in which publicly available assistants walked them through acquiring genetic material, turning it into a weapon and releasing it, and the companies' reply is that plausible text is not the same as capability.**

Average summary

Summarized from a detailed structured account of the article, not its original text, so nothing here is a direct quotation.

The piece opens with Dr David Relman, a Stanford microbiologist and biosecurity expert who has advised the federal government, hired by an AI company to test its product before release. One evening last summer the chatbot explained how to modify a well-known pathogen to resist known treatments and then how to release it to maximise casualties and minimise the chance of being caught. Relman asked the Times to withhold the pathogen's name and other specifics for fear of inspiring an attack, and would not name the company. What shook him was the bot answering questions he had not thought to ask.

Relman is one of a small group of experts enlisted to vet products for catastrophic risk, and several gave the Times **more than a dozen** conversations. The reporting's claim is that the assistants did more than repeat dangerous information: they described, in bullet-pointed detail, how to buy raw genetic material, turn it into a weapon and deploy it, with some brainstorming ways to evade detection. The article withholds pathogen names and other actionable specifics at the scientists' request.

The risk framing is careful. Since **1970** there have been a few dozen fairly small biological attacks

worldwide, the article's example being the **2001** anthrax letters that killed **five** Americans. Most experts say a major catastrophe remains unlikely. The argument is that a low-probability attack could kill millions, and that AI is one of several advances widening the pool of people who could cause harm, alongside protocols now scattered across the internet, direct online sale of synthetic DNA and RNA, and the ability to split sensitive work across private labs.

Kevin Esvelt of MIT is the central source. In 2023 he asked ChatGPT to help assemble a pathogen capable of mass death, received accurate instructions, and had a colleague carry the unassembled pieces to a White House meeting. He wants the industry to restrict a wider swath of biological information and release it only to approved users.

The companies pushed back, saying the shared chats lacked the detail needed to cause harm. Anthropic's Alexandra Sanderford said there is an enormous difference between plausible-sounding text and giving someone what they would need to act.

Why it's assigned

Page 2.1 opens the Delay unit by asking what AI models can and cannot currently do, and this is the journalistic version of that question, reported from inside the companies' own pre-release testing. It is useful precisely because it holds both answers at

once: one study found ChatGPT outperformed **94 percent** of expert virologists on laboratory protocol questions, while another, backed by AI companies, found models are not especially helpful across the full range of tasks needed to build a virus from scratch. It also previews the deep dives: the model-level safeguards of option 3, and the synthesis screening of option 4, which a study in Science showed AI-generated sequence variants could slip past before the fix was implemented.

Words to know

- **Pressure-testing, or red-teaming**: hiring outside experts to attack a product before release and report what it does. It is how every transcript in this article was produced.
- **Jailbreaking**: a prompt pattern that gets a model past its own safety filters. Cassidy Nelson of the Center for Long-Term Resilience compared the current safeguards to a flimsy wooden fence.
- **Refusal threshold**: how readily a model declines a request. Anthropic says it set aggressive thresholds for biological prompts and accepts some over-refusal as a result.
- **Deprecated model version**: an older release still available after a safer one ships. The article's sharpest finding is that a refusal on a current model came with a suggestion to switch to a previous version.

- **Uplift**: how much a tool adds to what a person could already do. It is the whole disagreement: the companies argue the transcripts show information, not capability.

Numbers in this reading

- More than a dozen chatbot conversations shared with the Times by experts.
- A few dozen fairly small biological attacks worldwide since 1970, with the 2001 anthrax letters killing 5 Americans.
- Nearly 50 percent: the fall in federal biodefense budget requests last year, alongside several senior biosecurity departures from the executive branch that have not been filled.
- 94 percent of expert virologists outperformed by ChatGPT on difficult laboratory protocol questions in one study.
- 8,000 words of instructions returned to an anonymous Midwest scientist who asked a research tool for a protocol to make a virus that once caused a pandemic; the scientist said the answer was not entirely accurate but could still significantly help someone with malicious intent.
- 7 virology and biosecurity experts who reviewed the transcripts for the Times.
- Thousands of variant sequences for dangerous agents generated by AI tools in a Science study, which commercial DNA screening software could

not detect; the researchers proposed fixes, which were implemented.

- 1 arrest, in August, of a 35-year-old physician in Gujarat accused of plotting an attack, who investigators say sought advice from AI-powered search and ChatGPT.

What the author is unsure about

The article is built around a live disagreement rather than a conclusion, and it reports both sides. Most experts quoted say a major biological catastrophe has not happened and remains unlikely. Sceptics argue chatbots only surface what is already online and that making a deadly virus takes years of hands-on expertise; Gustavo Palacios, a virologist who once worked at a Department of Defense laboratory, compared viruses to the world's finest clocks and doubted an amateur could take one apart and rebuild it, while saying he is concerned about AI in the hands of experienced actors. Jens Kuhn made the related point that for skilled actors the hard part is not making an agent but turning it into a weapon. Anthropic, OpenAI and Google all said the shared chats lacked sufficient detail to enable harm. The piece discloses that the Times is suing OpenAI over copyright in model development, a claim OpenAI denies, and it notes that Esvelt is regarded in his field as something of a Cassandra. The reporting also records what it chose not to print: the scientists

supplied transcripts on condition that pathogen names and other actionable details be withheld, and Esvelt had chats he judged too dangerous to share at all.

For the discussion

- The most concrete safeguard failure in the piece is not a jailbreak but a product decision: a current model refused a biological request and pointed the user to an older version, which the company described as an intentional strategy, and the older version answered.
- The two studies cut against each other in a way worth naming, since one has a model beating 94 percent of expert virologists on protocol questions while another finds little help across the full set of tasks needed to build a virus, which is the difference between knowledge and capability.
- A question worth asking the group: Esvelt wants a wider swath of biological information restricted and released only to approved users, so who maintains that approved-user list, and does that make model access a licensing regime like the ones already used for pathogens?

2.1-2 · CORE · COURSE TIME 20 MIN

When Should We Worry About AI Being Used to Design a Pathogen?

David Manheim, Adeline Williams, Casey Aveggio, Allison Berke · 2025

Read the original paper → · the course assigns: whole piece

> **Two RAND expert panels concluded that in 2025 and the near term AI speeds up pathogen design for people who are already capable rather than doing it on its own, and that no fundamental limit in biology would stop AI from eventually designing a working pathogen.**

Average summary

RAND ran two expert **Delphi panels** to pin down what AI can and cannot do in pathogen design. One panel was experts in AI and computational biology; the other was experts in biology and bioengineering. The Delphi method, developed at RAND in the 1950s, brings experts together to show where they agree and where they keep disagreeing. The aim is not to force consensus.

The headline finding: in 2025 and the near term, AI is and will likely continue to be an assistive tool rather than an independent driver of biological design. Panel members agreed that, without the creation of intelligent systems, the ability to alter or design a pathogen without a human would not emerge within the next two years, that is, by 2027.

The AI panel named five brakes. AI's power depends on the quality of the biological data used to train it, and the data needed on pathogen functionality can be difficult, if not impossible, to collect. AI struggles with complex and long-term predictions, such as host-pathogen interactions. Experimental validation of an AI-proposed design is a significant bottleneck, because real-world testing is slow, expensive, and infeasible, if not potentially unethical. And narrow models built for one biological problem are more useful than general-purpose models such as large language models, with narrow models unlikely to be usable by non-experts.

The biology panel named five brakes of its own, all in the organism rather than the software. Transmissibility is limited physically and biologically. Environmental stability is hard to engineer and brings trade-offs, such as weaker replication. Fitness trade-offs can leave a modified pathogen weaker in unrelated ways. Engineering constraints depend on the type of pathogen, so a change that is simple in one organism can be impossible in another. And human responses, including immunization, distancing or quarantine, and medical countermeasures, can rapidly blunt the spread of even a well-designed pathogen.

Two cautions sit under all of it. The limits are interdependent and context dependent, so no single one can be relied on alone. And no strong

fundamental limit was identified: panelists agreed AI should eventually be able to design novel, viable pathogens. Experts split on the speed, some expecting slow progress and marginally increased risk, others worried that models could reach a threshold and autonomously design novel bioweapons.

The brief's advice is therefore to monitor and to mitigate what is already plausible: track four areas of progress, strengthen and globally coordinate gene synthesis screening, put identity verification, experiment prescreening and audit trails on cloud and automated labs, and keep investing in core pandemic preparedness as a backstop.

Why it's assigned

Page 2.1 opens the Delay unit by asking what AI models can and cannot currently do, how we try to measure it, and how fast that is changing. This brief is the "cannot, yet" half of that answer, built from expert elicitation rather than headlines, and it puts a date on the claim: not autonomous before 2027, on these experts' view. It also points straight at the Deny sub-layer the unit covers next, since its concrete asks are gene synthesis screening and oversight of cloud and automated labs.

In depth

What the brief is for

Over the past ten years AI tools have spread through health care in diagnostic imaging and in self-management of conditions such as diabetes. Less visibly, biology researchers now use AI to work out how the three-dimensional structures of cellular organelles and large molecules such as enzymes drive their function, work that would previously have taken years. AI also helps assess what happens when those structures are modified, with results that could save or cost lives. The brief gives one example of the stakes: AI could help show how a virus that spreads only by physical contact might be altered to spread through the air, which could help prevent that outcome or enable it. A major concern is whether AI could plausibly cause such a change itself, intentionally or unintentionally.

Governments and researchers in several countries have started setting guidelines for safe and secure use of AI in biotechnology. The authors argue those efforts run into prior questions: what counts as safe, secure and reliable AI, what AI could plausibly let scientists do, and what counts as a plausible risk today, in the near future, and further out.

The study distinguishes **AI-assisted bioengineering**, where AI makes research easier or faster, from **AI-driven bioengineering**, where AI supplies the blueprint or the instructions. The goal is to say which scenarios sit inside plausible near-future capability

and which are unlikely or impossible, in order to narrow the scope of biosecurity planning and, in the authors' phrase, improve the signal-to-noise ratio in discussions of AI in biotechnology, sometimes called AIxBio.

Fifty years of trying to control biotechnology

For at least the past 50 years policymakers have tried to control emerging biotechnology. In February 1975, biochemists and geneticists met at the International Congress on Recombinant DNA Molecules at the Asilomar Conference Center in Monterey, California, to set risk-based principles for research that inserts foreign DNA into host cells. Those principles ended a moratorium and opened the way to decades of progress on the genetic basis of disease.

Since then, periodic clashes among biotechnology, computational biology, public health, medical ethics and biosecurity policy communities have exposed conflicting cultures, values and risk perceptions around dual-use biotechnology. In the past decade groups in the United States and internationally have tried to recreate Asilomar for AI, with broader international participation and a wider set of concerns; the brief cites the National Academies of Science, Engineering, and Medicine report *The Age of AI in the Life Sciences: Benefits and Biosecurity Considerations* (2025). A major sticking point in those meetings has been the lack of clarity about natural

constraints on biological systems, which is exactly the gap this study tried to fill.

Two panels, and the five questions

Because of that gap, RAND convened two separate panels: one of AI and computational biology experts, one of biology and bioengineering experts. Both considered five questions: the key areas of uncertainty or disagreement about the limits of engineered pathogens and AI-driven bioengineering; significant unknowns in biophysics, genetic engineering and machine learning; where expert opinion converges or diverges on the plausibility of proposed limits; which conjectured limits could be verified or falsified; and which assumptions are most likely to be challenged or confirmed, and by what kinds of breakthroughs. For each question panel members were given a list of potential existing and near-term limits and judged how likely each was to stop AI from reaching a given end.

What the panels were not asked

The panels were explicitly not asked whether large language models could let untrained actors create harmful pathogens for bioterrorism. The authors say recent evidence strongly suggests they could, citing Roger Brent and T. Greg McKelvey, Jr., "Contemporary AI Foundation Models Increase Biological Weapons Risk" (arXiv, June 12, 2025), even though a RAND red-

team study only a year earlier, by Christopher A. Mouton, Caleb Lucas and Ella Guest (2024), concluded LLMs were not yet mature enough. The brief calls that reversal a head-spinning shift in thinking and reads it as a sign of the speed of AI progress.

What the AI panel agreed on

Five points. In the next few years AI could assist experts in pathogen design, for example by speeding up some tasks, but could not yet do so independently or let non-experts do so. AI's power depends on data quality, and incomplete or biased datasets lead to inaccurate predictions of how a pathogen would behave in the real world. AI struggles with complex and long-term predictions where many factors interact, such as host-pathogen interactions. Experimental validation of AI-proposed designs is a significant bottleneck, since testing is slow, expensive and infeasible, if not potentially unethical, and simulation only goes so far. And general-purpose models such as LLMs are less useful than models tailored to narrow biological problems, which are themselves unlikely to be usable by non-experts.

What the biology panel agreed on

Five near-term biological challenges. Transmissibility is limited by a pathogen's survival outside the host, its mode of transmission, and its effect on host

behaviour, since a dead or incapacitated host may be a poor vehicle. Environmental stability is hard to engineer: raising survival outside the host can reduce effective replication, and those trade-offs are hard to predict computationally. Fitness trade-offs mean a genetically modified pathogen can become weaker in unrelated ways, much like a plant or animal bred for one trait. Engineering constraints depend on the organism, so a modification that is simple in a virus may be impossible in a bacterium given current knowledge. And human responses, immunization, social distancing or quarantine, and medical countermeasures, can rapidly blunt the spread of even a well-designed pathogen.

Where the two panels met

Reading across the transcripts, the researchers found shared ground and shared uncertainty. AI is already an assistive tool and will stay one in the near term, including carrying out automated work in cloud labs, but cannot act on its own to alter or design a pathogen; panelists agreed that without the creation of intelligent systems such an ability would not emerge by 2027. AI already helps experts and malicious actors with pattern recognition in large datasets and with forming and validating hypotheses, and might automate some lab work soon, though still needing human guidance, interpretation and validation.

The risks are increasing, but experts differed on how fast. Some expected slow progress and marginally increased risk; others worried models could rapidly gain the ability to autonomously design novel bioweapons once they cross some capability threshold. The named enablers of that shift are automation, simulation, progress in clinical applications, and progress toward generally capable AI.

The limits themselves are complex, interdependent and context dependent. Experts in both panels kept pointing out how biological and technological constraints overlap and reinforce each other, which makes it hard to assess any one constraint on its own. Data quality and quantity came up in both panels as a real barrier, especially for generating novel pathogens or predicting complex functions. And no strong fundamental limit was found: panelists agreed AI should eventually be able to design novel, viable pathogens.

What they recommend

The authors read the panels as saying the worst fears, AI going rogue and autonomously designing radically novel dangerous pathogens, remain implausible at least through 2027, while progress toward autonomy could change that outlook and needs watching. AI tools are likely to keep acting as accelerators for already capable actors. Some experts expected rapid

capability gains, but no consensus emerged on direction or speed.

Governance has two immediate targets: monitoring scientific progress as it shapes the future risk landscape, and mitigating the most plausible and most actionable pathways, which involve using AI to optimize or modify existing pathogens. Monitoring should track four areas: clinical applications of bioengineering; laboratory automation that gives faster feedback for applications and for training AI; reliable simulations of biological systems that reduce the need for experimental validation or training data; and generally capable AI systems.

In the near term, experts should clarify where hard limits exist, especially where biological and AI constraints meet, which would sharpen threat models and let policymakers focus on credible misuse cases. Biological data infrastructure is called a key leverage point, since the quality, accessibility and oversight of genomic and experimental data shapes both capability and risk.

Alongside that, the brief argues for reinforcing traditional biosecurity safeguards, because AI has lowered the threshold for sophisticated misuse. Strengthening and globally coordinating gene synthesis screening can help prevent malicious or accidental creation of risky sequences, which matters more because experts agreed AI accelerates iterative

design. Cloud and automated lab platforms should sit under oversight frameworks with identity verification, experiment prescreening and audit trails, to prevent or at least detect misuse. Sustained investment in core pandemic preparedness, rapid diagnostics, scalable vaccine platforms and public health surge capacity, is the backstop that pays off even for pathogens nobody engineered.

The brief closes on culture and on the limits of voluntarism. It calls for responsible-use norms among developers and institutions, and for interdisciplinary coordination between AI practitioners, biologists and biosecurity experts, through mechanisms such as horizon scanning and iterative threat modelling. But self-governance, including predeployment review and responsible disclosure, is inherently limited in anticipating novel misuse pathways or constraining determined actors, including state-sponsored programmes. That, the authors say, is why complementary regulatory and institutional safeguards are needed as quickly as possible.

Words to know

- **Delphi panel.** A structured way of polling experts in rounds to surface agreement, persistent disagreement and the reasoning behind both. Developed at RAND in the 1950s. It matters here because the brief's findings are expert judgment gathered this way, not experimental results.

- **AI-assisted versus AI-driven bioengineering.** Assisted means AI makes the research easier or faster; driven means AI supplies the blueprint or instructions. The whole brief turns on which one we are in.
- **AIxBio.** Shorthand for AI-enabled or AI-assisted biology. Useful because it is the term the policy conversation uses.
- **Dual-use.** Research or technology that can serve both beneficial and harmful ends. The reason this field needs guidelines at all.
- **Gene synthesis screening.** Checking orders for synthetic DNA against sequences of concern before they are made and shipped. One of the brief's two concrete near-term asks.
- **Cloud lab.** A remote, largely automated laboratory that runs experiments ordered over the internet. Named both as something AI already works through and as something needing identity checks, prescreening and audit trails.
- **Fitness trade-off.** When improving one trait in an organism weakens it elsewhere. A biological brake on engineered pathogens.
- **Red team.** A group asked to attack a system or test a worst case deliberately. The 2024 RAND red-team study is the comparison point for how fast the picture changed.

Numbers in this reading

- **2025 and the near term**: the window in which AI is judged to be assistive rather than an independent driver of biological design.
- **After 2027**: when the risk landscape is expected to expand over the longer term, with experts very uncertain how fast.
- **Next two years, i.e. by 2027**: the period in which panelists agreed autonomous pathogen design would not emerge, absent the creation of intelligent systems.
- **At least the past 50 years**: how long policymakers have tried to control emerging biotechnology.
- **February 1975**: the International Congress on Recombinant DNA Molecules at Asilomar, which set risk-based principles and ended a research moratorium.
- **1950s**: when the Delphi method was developed at RAND.
- **Two panels**: one AI and computational biology, one biology and bioengineering, deliberately kept separate.
- **Five questions**, and **five agreed points from each panel**: the structure of the elicitation and of the findings.
- **Four monitoring areas**: clinical bioengineering applications, lab automation, reliable biological simulation, and generally capable AI.

- **June 12, 2025** (Brent and McKelvey arXiv paper) versus **2024** (Mouton, Lucas and Guest RAND red-team study): one year apart, opposite conclusions about whether LLMs could help untrained actors.

What the author is unsure about

A great deal, and the brief says so. Experts were very uncertain how rapidly AI capabilities would evolve, and no consensus emerged on what direction gains would take or how quickly. The Delphi method is not designed to force consensus, so persistent disagreement is a result, not a failure. The identified limits are interdependent and context dependent, which the authors flag as making any single constraint hard to assess alone. Data on pathogen functionality can be difficult, if not impossible, to collect, which limits what models can learn. Where hard limits exist is itself listed as near-term work still to be done. The reversal between the 2024 red-team finding and the 2025 evidence on LLMs is presented as a caution about how fast conclusions in this field age. And the central forward-looking claims, no autonomy by 2027, eventual capability with no fundamental barrier, are expert judgment and forecast, not measurement.

For the discussion

- The brief's two time horizons do different work: "not autonomous before 2027" is a reason to act now on screening and lab oversight, while "no fundamental

limit" is a reason not to treat today's brakes as permanent.

- Worth raising as a tension: the panels were not asked the question most people care about, whether LLMs can walk a novice through making a pathogen, because the authors treat that as already answered by evidence that appeared within a year of a RAND study concluding the opposite.
- A question for the group: if the limits really are interdependent and context dependent, what would a monitoring programme actually watch, and would the four named areas, clinical applications, lab automation, simulation, and generally capable AI, give enough warning to matter?

2.1-3 · FURTHER · COURSE TIME 6 MIN

To Forecast AI's Impact on Biosecurity, We Asked: Why are Attacks So Rare?

Abi Olvera · 2026

Read the original paper → · the course assigns: whole piece

> **Practitioners who have actually worked in labs say bioweapons are rare because they are hard to make and bad weapons to use, and near-term AI lowers some barriers without yet changing the calculus for someone who wants to cause harm.**

Average summary

Abi Olvera, of the Golden Gate Institute for AI, interviewed biosecurity professionals with decades of hands-on laboratory experience. What she found is more reassuring than the headlines: bioweapons are genuinely difficult to make, and near-term AI may not change that as much as you would think. This is part one of a four-part series. The four essays ask why bioweapons are rare, how much laboratory skill the processes really require, where in the production chain AI helps and where it does not, and why the biosecurity conversation underplays what makes successful bioweapons so rare.

Her first point is the record. Since the 1980s, only one fatal bioterrorist attack has occurred, the **2001 anthrax letters**. Other incidents, a Salmonella attack to influence a local election, a medfly release targeting California crops, and two ricin letter attempts, caused no deaths.

Her second point is that bioweapons are bad weapons. You cannot time them, because a virus spreads on its own schedule. You cannot aim them, because a virus spreads uncontrollably. Protecting your own people requires tipping your hand, since a vaccination program is hard to hide. People who want to cause harm still follow a cost-benefit logic and pick the tool most likely to work, and for almost any real-world objective that tool is a bomb, a gun, a chemical

weapon or a cyberattack: cheaper, faster to create, easier to deploy, more controllable and more predictable. She notes this logic is well studied in military strategy but less recognized in AI discussions, where researchers focus on how AI could help build a bioweapon while practitioners focus on the limiting factors.

The heart of the piece is a chart of nine factors practitioners name (it sits in an embedded chart rather than the article text): malice, resources and budget, access, sustained competence, multiple domain expertise, available options, deterrence, time and timing control, and uncontainability. On access, the chart says only tens of thousands of people in the United States have sustained access to the needed facilities and supplies, and two experts show their arithmetic in footnotes. Matt Sharkey of RAND starts from roughly 100,000 US academic biologists and filters down to 20,000 to 30,000 technically capable people, then to 6,000 to 10,000, with the private sector pushing that 1.5 to 3 times higher. Michael Montague starts broader at 400,000 to 700,000, applies three harsher cuts, and reaches 1000 in academia, then 40 to 70. They disagree on how hard "boot-up" is and whether nighttime lab use would go unnoticed.

AI does lower some barriers: deciphering research on culturing cells or dispersing pathogens, designing

modifications to a virus's genome with biodesign tools, and non-biological steps such as which shipping routes avoid regulatory oversight and customs inspection. But those same capacities also make bombs, chemical weapons and cyberattacks easier, so AI does not yet significantly change the calculus.

She is not complacent. Pathogens targeting genetic traits like ancestry or sex, new kinds of actors, and robotics and automated laboratories could all shift the math.

Why it's assigned

Page 2.1 asks what AI can and cannot currently do and how fast that is changing, and this is the practitioner's answer: the bottleneck is not knowledge alone but access, sustained competence, tacit skill and the attacker's own cost-benefit logic. For the Delay layer it matters because it says where denial still bites and why deterrence works at all, and it ends with a warning about the discourse itself, that the people who understand these risks best work in laboratories, government agencies or national security, while the advocacy and policy organizations filling that gap have incentives pushing them toward worst-case scenarios.

Words to know

- **Boot-up.** The knowledge needed to turn an assembled genome into a working, replicating virus. The step the two experts disagree most about, and the one that cuts Montague's estimate from 1000 people to 40 to 70.
- **Reverse genetics.** (Not from this reading) building a working virus starting from its genetic sequence rather than from an existing sample. Sharkey assumes only about a third of technically capable people could work this out from published protocols.
- **Molecular cloning.** A standard laboratory technique for copying and assembling DNA. Sharkey's first filter: he assumes about 75% of academic biologists can do it.
- **AI biodesign tools.** Models built to design biological sequences rather than to chat. Named as the thing that reduces the expertise needed to design genome modifications.
- **Tacit knowledge.** Skill that lives in hands and practice rather than in written protocols. The subject of the next essay in the series, and the gap she says limits how much AI can help.
- **Uncontainability.** The ninth factor: biological agents spread unpredictably, you and your allies risk infection too, and protective measures such as vaccines are prohibitively expensive.

Numbers in this reading

- **One fatal bioterrorist attack since the 1980s**: the 2001 anthrax letters.
- **Zero deaths** from the other incidents she lists: a Salmonella attack over a local election, a medfly release in California, and two ricin letter attempts.
- **Nine factors** practitioners name as making bioweapons rare.
- **Tens of thousands**: people in the United States with sustained access to the facilities and supplies, per the chart, and the figure she uses as a rough order-of-magnitude estimate.
- **Roughly 100,000**: Sharkey's starting population, about 40,000 life sciences faculty plus about 60,000 biological or biomedical doctoral students, with about 45,000 Master's students not counted because many programmes do not involve benchwork.
- **About 75%**, then **about 20 to 30%**: his filters for molecular cloning skill, then access to cell culture equipment, giving **20,000 to 30,000** technically capable people.
- **About a third**: how many of those he thinks would independently work out the reverse genetics approach, giving **6,000 to 10,000**, with the private sector pushing it **1.5 to 3 times** higher.
- **400,000 to 700,000**: Montague's broader starting population, including industry PhDs and technicians.

- **10%**, **1 in 10 labs**, **1 in 100 people**: his three cuts, for boot-up knowledge, for having the right equipment, and for unsupervised long-term access to run a multi-hundred-day project undetected.
- **1000**, then **40 to 70**: his estimate for academia, before and after boot-up difficulty.

What the author is unsure about

She flags the softest number herself: the estimate of how many people have access to facilities is back-of-the-envelope, and the two experts' bottom lines differ by two orders of magnitude, 6,000 to 10,000 against 40 to 70. She names where they disagree, on how difficult boot-up is and whether nighttime lab use would go unnoticed. She also says lab work is not the only route to a pathogen, while listing why the alternative of deliberately getting infected is hard: finding actively sick people, getting close enough, timing it around contagiousness windows, and being stuck with natural transmission rates. Her AI claim is explicitly time-limited, that AI does not *yet* significantly change the calculus. And she holds back the hardest case: bioweapons based on engineered pathogens remain unproven, their feasibility is unclear, and she defers them to the third essay.

For the discussion

- The useful reframe here is that attackers choose among weapons, so the question is not whether AI

makes bioweapons easier but whether it makes them easier faster than it makes bombs, chemical weapons and cyberattacks easier, which she says it does not yet.

- A tension worth raising: the two expert estimates of who could actually do this differ by two orders of magnitude, which means the field does not agree on the size of the population any denial or screening measure is meant to cover.
- Her closing claim is about the field itself, that the best-informed people are in labs, agencies and national security while advocacy organizations with worst-case incentives shape the public conversation, which is worth testing against the other readings on this page.

2.1-4 · CORE · COURSE TIME 15 MIN

Horsepox Synthesis: A case of the unilateralist's curse?

Gregory Lewis · 2018

Read the original paper → · the course assigns: whole piece

> **When risky research only needs one optimistic scientist to go ahead, the decision falls to the most optimistic person in the field rather than to the group, and every safety net meant to catch that, funders, regulators and journals, has the same weakness.**

Average summary

In 2016 the Canadian researcher Ryan Noyce and colleagues rebuilt **horsepox**, a virus that had gone extinct, by synthesizing a copy of its DNA from the published sequence, putting it into a cell culture alongside another poxvirus, and recovering live virus. The work was published the month before this article, in January 2018, and it was controversial straight away.

The case for it: horsepox is a close cousin of the vaccinia virus used as a smallpox vaccine, so resurrected horsepox could be used to make another and potentially better smallpox vaccine. Noyce and colleagues also argued the demonstration would inform the biosecurity conversation about smallpox. The case against: smallpox is also closely related to horsepox, so if horsepox can be made from a published sequence, someone could do the same for smallpox. Tom Inglesby, director of the Johns Hopkins Center for Health Security, is quoted on that side. Gregory Lewis puts the objection plainly: generating a risk to show it is risky is a dangerous path, and a better smallpox vaccine may be poor compensation for a higher chance of an artificial smallpox outbreak.

The article's real subject is the general problem the case illustrates, the **unilateralist's curse**, an idea proposed by Nick Bostrom and colleagues at Oxford. Imagine 100 scientists each deciding in good faith

whether synthesizing horsepox is good for humankind. For it not to happen, all 100 must independently decide against. For it to happen, one has to say yes. That asymmetry creates an action bias, and it means the decision effectively belongs to the most optimistic outlier, even when the median view is that the work should not be done.

Lewis lists what makes the curse worse: more potential actors, the passage of time, the difficulty of assessing a threat accurately, conflict of interest (the horsepox authors are co-inventors on a related patent), and a tragedy of the commons in which each scientist reasons that someone else will do it anyway.

He then shows the same weakness in each safety net. Funders review dual-use research, but the horsepox work was funded by Tonix, a private company hoping to commercialize it, and private funding and citizen science let more work bypass that net. Regulators are disjointed; work rejected in one country may be approved in another. Journals act independently, so authors can shop around, and Lewis quotes a Science editor declining the paper partly because of the administrative burden of dual-use review.

His answer is collective caution: consensus views that people follow even when they disagree, a stronger norm toward restraint, non-scientists at the table, and a "no-undercut" rule borrowed from the Australia

Group export-control regime, extended to funders, review boards and journals.

Why it's assigned

Unit 2 is the Delay layer, and this reading sits on its Deny sub-layer: denying access to the knowledge needed to build a pandemic pathogen, which here means a published sequence and a published method rather than a physical sample. The course note says the piece summarizes another controversial piece of research and publication and introduces the unilateralist's curse, and that concept is the unit's clearest answer to why information hazards are hard to govern, because deterrence and denial both fail if one well-meaning actor can act alone. Read alongside the page's AI readings, it is the pre-AI version of the same structural problem.

In depth

What was done, and why it was contested

Horsepox is a virus brought back from extinction by biotechnology. In 2016 Ryan Noyce and colleagues synthesized a copy of horsepox DNA from the virus's published DNA sequence, placed it into a cell culture alongside another poxvirus, and in this way recovered copies of live horsepox virus.

The dual-use problem is immediate. Horsepox is a close cousin of the vaccinia virus used as a smallpox

vaccine, so resurrected horsepox could be used to make another and potentially better smallpox vaccine. But smallpox itself is also closely related to horsepox, so the obvious question is whether the same route works for smallpox.

Publication of the work the month before the article set off a dispute. Noyce and colleagues argued that demonstrating feasibility would inform and advance the biosecurity conversation around smallpox. Tom Inglesby, director of the Johns Hopkins Center for Health Security, disagreed, and Lewis sums up that position in two parts: generating a risk in order to show it is risky seems a dangerous path, and the potential benefit of a better smallpox vaccine may prove poor compensation for raising the chance that a malicious actor could cause an artificial smallpox outbreak.

Beyond whether this work should have been done or published, Lewis says the case highlights a general challenge for dual-use research in biotechnology.

Curses

The unilateralist's curse was proposed by Nick Bostrom and colleagues at Oxford. Lewis sets it up with a thought experiment: 100 scientists are individually deciding whether synthesizing horsepox is a good idea. All act impartially and in good faith, and

would only do the work if they genuinely thought it was on balance good for humankind.

The structure does the damage. For the synthesis not to happen, all 100 must independently decide against it. For it to happen, one scientist who judges the benefits to outweigh the risks acts alone. That is the action bias: the work is more likely to happen when scientists decide independently than when they decide as a group.

The "curse" is the asymmetry of mistakes. If synthesis is good, a scientist who wrongly thinks it bad does little harm, because someone else who correctly thinks it good will go ahead. If synthesis is bad, a single scientist who wrongly thinks it good produces the bad outcome even though every peer correctly sees it as a mistake. So the decision is left to the most optimistic outlier. Where the most optimistic view is that research should proceed but the median view is that it should not, the research probably should not be done, and probably happens anyway.

The figure

The article refers to a table that it says makes the situation clearer, with the bottom right corner representing the unilateralist's curse, that is, the case where the work is harmful and one mistaken optimist carries it out. The table image was not captured in the saved text, so its cells are not reported here.

What makes it worse

Lewis names five aggravating factors.

The number of potential actors who can act unilaterally. If 10,000 scientists are weighing a concerning piece of research, the chance that one mistakenly thinks it a good idea is greater than with 100, or 10, or one.

Time. As time passes, the opportunity for a mistake to be made increases.

The difficulty of assessing a threat, or individuals' poor ability to assess one. The harder the judgment, or the weaker the judge, the higher the chance someone greatly underestimates a danger.

Conflict of interest. However wide the spectrum of reasonable disagreement, the spectrum of unreasonable disagreement is wider. Fame or money can colour judgment about the common good. Lewis notes that the authors of the horsepox paper are listed as co-inventors on the patent relating to their work, so they stand to gain financially from commercial applications.

Tragedy of the commons. If scientists suspect someone else might pursue dangerous work for personal gain, the implicit agreement not to do it is fragile, and the reasoning becomes: if someone is

going to do this anyway, I might as well get the benefit.

Lewis is careful about how far the case supports the theory. It is hard to say for sure whether the curse applies here: it remains unclear whether horsepox synthesis is in fact bad, and scientists seldom report the experiments they decide not to perform. He offers a circumstantial argument instead. The possibility that synthetic biology could produce dangerous pathogens, including smallpox, had been discussed openly for more than a decade. In 2014 the World Health Assembly commissioned a report on the public health implications of being able to recreate smallpox through synthetic biology. When horsepox was actually synthesized, the controversy that followed showed a real spread of opinion among reasonable people. And the researchers themselves placed their views at the optimistic end: in their paper's discussion section they noted that the sequence of variola, that is smallpox, has been known since 1993, said their studies show it is clearly accessible to current synthetic biology technology, and hoped the work would promote informed public health discussion, stimulate new evaluation of horsepox-based vaccines, and advance the capacity to rapidly produce next-generation vaccines and poxvirus-based therapeutics. Lewis reads all of this as a long-available possibility finally realized by unusually optimistic scientists.

Vulnerable governance

Several safety nets could in principle catch well-intentioned but mistaken researchers, and each has the same hole.

Funding bodies give centralized oversight, and most major funders have review mechanisms for research of dual-use concern. But not all biological research passes through them: the horsepox work was funded by Tonix, a private enterprise hoping to commercialize the result. Lewis expects the growing role of private funding, and the democratization of biology through things like citizen science, to send more work around this net.

Regulators, whether national governments, international bodies, or institutional biosafety committees at particular academic centres, are disjointed and cannot be relied on to see every danger. Lewis quotes a 2016 World Health Organization report in which the horsepox scientists themselves noted that regulators may not have fully appreciated the significance of, or potential need for, regulation or approval of steps and services involved in using commercial DNA synthesis companies, laboratory facilities, and the federal mail service to synthesize and replicate a virulent horse pathogen. The curse makes regulatory failure worse, because even if one university or country blocks the work, others need not follow. As Inglesby notes, work

considered very high risk and therefore rejected by one country may be approved by others.

Journals can decline hazardous work, as was discussed over gain-of-function influenza experiments. But journals act independently and so are open to unilateralism. Lewis quotes an editor at Science replying to a pre-submission inquiry from the horsepox authors: while recognizing the technical achievement, the journal decided the paper would not offer readers a sufficient gain of novel biological knowledge to offset the significant administrative burden the manuscript represents in terms of dual-use research of concern. His worry is that when well-resourced journals bounce dual-use work and the headache of evaluating it, it lands with publications less able to judge it accurately, or with more reckless editors. Unless journals talk to each other about concerns, authors can shop around. And even with no journal at all, self-publication and other ways of circulating work make publication review a poor last line of defence.

What can be done

The curse is greatly attenuated when individuals form a consensus view about whether a piece of research is too dangerous to allow and then follow that consensus even when they personally disagree. Lewis suggests three cultural shifts: scientists less willing to strike out alone when the downside is a public health

disaster rather than mere failure, a stronger community norm toward caution for research with population-level health consequences, and wider inclusion of non-scientific stakeholders in the decisions.

For the institutional side he borrows from the export control regime known as the Australia Group, which uses a no-undercut principle: countries are expected not to permit exports another country has rejected without consulting that country first. Similar norms could apply to funders, to regulators such as institutional biosafety committees and institutional review boards, and to journals, and such collaborations could also concentrate decisions in a smaller number of better-resourced bodies. A global governance regime is, he says, unattainable now, but steps in the right direction are not hard to identify.

Close to the precipice

Lewis is candid about the cost of restricting research: breakthroughs often come from mavericks going against consensus, and the future benefits of biotechnology will be considerable. But he argues those benefits are increasingly fragile, and closes with a cost comparison. Eradicating smallpox cost approximately $1.7 billion, took the combined efforts of thousands of people, and has saved more than 40 million lives. Horsepox synthesis required three scientists and cost $100,000.

Smallpox eradication is one of humanity's great triumphs and the costs of a synthesized smallpox outbreak would be unconscionable, yet, he writes, we edge closer to that precipice on occasional and perhaps mistaken bursts of enthusiasm and curiosity. Smallpox is not the only such risk: pathogens could be deliberately engineered as well as recreated, and further advances may reveal risks currently beyond the horizon. His conclusion is that with this much at stake the guiding principle should be collective caution rather than unilateral action, because not all discoveries justify courting global catastrophe.

Lewis thanks Joe Carlsmith, Tom Inglesby, Piers Millett, Cassidy Nelson, Carl Shulman and Andrew Snyder-Beattie for advice and criticism, noting their help does not imply agreement.

Words to know

- **Horsepox.** A poxvirus that had gone extinct and was rebuilt in 2016 from its published DNA sequence. Close relative of both the smallpox vaccine virus and smallpox itself, which is what makes it dual-use.
- **Vaccinia.** The poxvirus used as the smallpox vaccine. The reason a resurrected horsepox could have vaccine value.
- **Variola.** The virus that causes smallpox. Its sequence has been published since 1993, which is

the fact the horsepox authors use to argue the information was already out.

- **Unilateralist's curse.** The problem that when any one actor can act alone, the decision is effectively made by the most optimistic person, so mistakes toward action are much more costly than mistakes toward restraint.
- **Dual-use research of concern.** Research that could be applied either to benefit or to serious harm. The category journals, funders and regulators are supposed to review.
- **Institutional biosafety committee.** A local review body at a research institution that approves work involving biological hazards. One of the safety nets Lewis calls disjointed.
- **Australia Group.** An export-control arrangement among countries covering chemical and biological weapons-relevant goods. Lewis borrows its no-undercut principle, under which one country does not approve an export another has refused without consulting first.
- **Tragedy of the commons.** When each individual's self-interested choice degrades a shared good. Here it is the reasoning that someone else will do the risky work anyway.

Numbers in this reading

- **2016**: when Noyce and colleagues synthesized horsepox.

- **100 scientists**: the size of the thought experiment that sets up the unilateralist's curse.
- **10,000 versus 100, 10, or one**: Lewis's illustration that more potential actors means a higher chance one of them mistakenly goes ahead.
- **1993**: when the variola, that is smallpox, sequence became known, per the horsepox authors' own discussion section.
- **2014**: the World Health Assembly commissioned a report on the public health implications of recreating smallpox through synthetic biology.
- **2016**: the World Health Organization report quoting the horsepox scientists on what regulators may not have appreciated.
- **More than a decade**: how long the possibility of synthetic biology producing dangerous pathogens had been expressly discussed before the synthesis happened.
- **Approximately $1.7 billion, thousands of people, more than 40 million lives saved**: the cost, effort and payoff of eradicating smallpox.
- **Three scientists and $100,000**: the cost of synthesizing horsepox. The contrast with the eradication figures is the article's closing argument.

What the author is unsure about

Lewis is explicit that he cannot settle the central question. It is hard to say for sure whether the unilateralist's curse applies to the horsepox case: it remains unclear whether horsepox synthesis is in fact bad, and scientists seldom report experiments they decide not to perform, so the counterfactual is unobservable. What he offers instead is a circumstantial argument. He also concedes the cost of his own prescription, that restricting research has real costs because breakthroughs often come from mavericks, and that a global governance regime is unattainable now. His claims about what would help, stronger norms, no-undercut rules extended to funders and journals, are proposals, not findings.

For the discussion

- The horsepox case is a clean test of the unit's Deny layer, because what was published was a method and a sequence rather than a sample, so the thing to be denied was information that at least one journal had already decided was not worth the review burden.
- Worth raising as a tension: Lewis's remedy asks individuals to follow a consensus they believe is wrong, which is close to the opposite of how scientific judgment is usually praised, and he admits mavericks produce breakthroughs.
- A question for the group: of the three safety nets, funders, regulators and journals, which could

realistically adopt a no-undercut rule first, given that the one actor who ignores it is the whole problem?

2.1-5 · CORE · COURSE TIME 15 MIN

The role of evals in the biorisk evidence hierarchy

SecureBio and Jasper Götting · 2026

Read the original paper → · the course assigns: whole piece

> **AI evaluations are flawed proxies for real-world biorisk, but they are the best cheap, comparable and repeatable evidence available, and waiting for better evidence has a cost critics rarely count.**

Average summary

Jasper Götting, writing for SecureBio, is answering a recurring complaint: **evals** do not measure what we actually care about, so why bother? The complaint has a sharp version. We want to know whether AI increases biorisk and to convince the people who make decisions. Benchmarks do not measure the real bottlenecks a malicious actor faces. So what is needed is real-world uplift studies in genuine wet labs. An **uplift study** measures whether people do a task better with AI than without it.

Götting agrees uplift studies on biological tasks would be the strongest evidence. The argument is that evidence comes in a hierarchy, and evals hold a specific, useful place in it.

Tier 1 is theoretical argument from first principles, and he says we have been here. Pointing at the failures of past bioweapon attempts, interviewing experts, running a demonstration at the White House, running forecasting exercises, and noting AI's usefulness in fields like coding was enough to get security-minded people to take the risk seriously and fund more work. It is cheap and it was load bearing. It does not convince anyone sceptical of domain-transfer claims or anyone who needs a solid basis for a consequential decision.

Tier 2 is evals, and we are here. Evals are in silico estimators of a model capability, imperfect and noisy, but a set of many good ones can show subtle differences between models. The tier is broad: multiple-choice biology exams, open-ended questions checked against experimental ground truth, experts interrogating a model over many turns for rare knowledge, and agentic tasks in a simulated research environment. He takes the criticism seriously and lists seven known failure modes, including specification gaming, underelicitation, answer leakage, weak construct validity, sudden threshold effects,

infohazard trade-offs in keeping an eval private, and saturation.

His defence is cost and repeatability. A benchmark costs tens to hundreds of thousands of dollars to build, against millions for a well-powered uplift study, and running it on a new model is a fraction of that. He compares evals to in vitro studies in drug discovery, between pure reasoning and clinical trials. Evals can also rule capability out, with the warning that the cost of a false negative here is much, much higher than a false positive.

Tier 3 is real-world uplift studies, and we are getting there. Past evals is a cost cliff: a proper study needs many participants, time and money, and getting the design right is hard enough that a bad design produces a result that reads as "no risk" when it should read "this design cannot see the risk." In his opinion only the recent Active Site RCT counts as sufficiently large.

Tier 4 is actual AI-enabled incidents and near-misses, which he hopes we never reach, and which may not even settle the question because attributing cause after the fact is hard.

Why it's assigned

Page 2.1 asks what AI models can and cannot currently do, how we try to measure it, and how fast that is changing. This reading is the "how we try to

measure it" part, and it is honest about how weak the measurement is. For the Delay layer that matters directly: deterrence, denial and interception all depend on someone being able to say when a model has crossed a line, and this piece explains why that line is currently drawn with cheap proxies rather than real-world evidence.

In depth

The complaint, and the answer in one breath

The post opens with the questions SecureBio keeps hearing: what is the theory of change for benchmarks, and why bother if evals do not measure what we care about. His answer up front: evals beat first-principles argument, cost far less than wet lab uplift studies and repeat cheaply, and can help rule capabilities out.

The critique he is answering is about **construct validity**, whether a test measures what it claims to. Stated fully: we want to know if AI increases biorisk and, if so, convince the relevant stakeholders it can be mitigated; benchmarks do not measure the real-world bottlenecks malicious actors face; so what is needed is uplift studies measuring success in genuine wet labs. Götting concedes those would be the very strongest evidence. His counter is structural: there is a hierarchy of evidence, and evaluations occupy a specific and important position in it.

Tier 1, theoretical arguments: we have been here

Arguing from first principles was surprisingly effective at the start. The moves available: gesture at the hurdles and failure modes of past bioweapon attempts, interview experts, perform a simple demonstration at the White House, run forecasting and superforecasting exercises, and point to AI's utility in other domains such as coding. That was enough for security-minded people to take AI-enabled biorisk seriously and look for more evidence. This, he says, is how he and many others ended up in the field; the arguments justified investment in the higher tiers, and he regards the early work extremely highly. Their limit: cheap to produce, but not sufficient for anyone even slightly sceptical of domain-transfer claims, or anyone needing a solid basis for a consequential decision.

Tier 2, evals: we are here

Evals are in silico estimators of an underlying model capability, imperfect and often noisy, but many high-quality ones together can point out even subtle differences between models. A footnote sets his terms: evals are anything measuring a capability in silico, manual elicitation and benchmarks alike, and he means capability evaluations, not safeguard evaluations measuring refusal behaviour.

The tier is broad: multiple-choice biology exams, open-ended research questions validated against experimental ground truths, experts interrogating

models over many turns for rare knowledge, and complex agentic tasks in a simulated research environment. That range brings large differences in harm specificity, difficulty and construct validity: an agentic benchmark testing whether a model can re-design an actual pathogen genome is, he says, better evidence than a multiple-choice test on college virology.

He steelmans the critique: not that evals are useless, but that seven known failure modes limit what they mean.

Specification gaming. Once evals set frontier AI safety policy tiers or regulatory thresholds, developer incentives change and the gap between score and real capability can widen rather than narrow. Goodhart's law is not unique to evals, but the stakes are unusually asymmetric: a firm could suppress performance on a dangerous capability eval while the capability keeps growing.

Underelicitation. Scores are lower bounds. Scaffolding (which tools the model may use), fine-tuning on domain data, and prompt or context engineering shift performance substantially, so a motivated user gets more out of a model than a standard run does. Eval awareness and sandbagging, a model realizing it is being evaluated and concealing capability, is a further uncertainty he says is not conclusively solved.

Leakage. Scores can overestimate capability when answers leak into training data and get memorized; private holdout sets and canary strings catch some of that.

Construct validity. A textbook-biology multiple-choice test measures something, but not obviously the ability to coach someone through making a pathogen in a real lab, especially as models get better at spotting patterns in tests. Realistic tasks grounded in real bottlenecks help, but the link between eval performance and harm potential is poorly understood.

Non-linear threshold effects. Benchmark performance scales predictably, yet crossing a capability threshold can jump real-world impact, as with coding agents right after they became good enough for many real software engineering tasks. He says it is hard, maybe impossible, to know in advance where those thresholds sit for biology.

Infohazard trade-offs. A highly valid eval should be harm-specific, which may mean keeping the dataset fully private and giving up external scrutiny; private sharing with third-party auditors plus methodology write-ups softens that, but the trade-off remains.

Saturation. Benchmarks saturate and need harder successors, though they usually last a few model generations. Measurement stays unbroken while models and evals overlap, but any single eval gets

little validation before retirement. Even saturated evals, he adds, can be meaningful indicators.

His defence is cost and repeatability. Developing a benchmark costs tens to hundreds of thousands of dollars against millions for a well-powered uplift study, and running it on newly released models is a fraction of that, though possibly still significant with heavy inference-time scaling. He likens evals to in vitro studies in drug discovery, an intermediate between pure mechanistic reasoning and full clinical trials. Evals can also rule out: a properly elicited model performing badly and clearly lacking certain knowledge raises confidence that it poses no corresponding real-world risk. But real-world biorisk creates a catastrophic asymmetry, because the cost of a false negative, an erroneous rule-out, is much, much higher than the cost of a false positive.

Tier 3, real-world uplift studies: we are getting here

Past evals comes a cost cliff. Real-world evidence can be collected cheaply by recruiting ten friends and renting a lab for a weekend, but that is hardly better than anecdata. A properly powered study needs many participants and time, which means money. Done correctly, uplift randomized controlled trials are the most convincing evidence for AI's impact on real-world lab success.

"Done correctly" hinges on the number of participants, target demographic, task difficulty, harm specificity, trial duration, how the model is actually used, and the porosity of the non-LLM control group to now-ubiquitous AI integrations. Getting those right is difficult and expensive, so safety researchers might repeat it annually, not for every model release. Being slightly off produces false negatives, read as "no risk" when the honest reading is "this design can't see the risk." Even a well-designed study can be dismissed if its demographic or task type misses a stakeholder's threat model, for instance showing uplift in novices when the worry is semi-experts. He says "we are getting here" because, in his opinion, only the recent Active Site RCT qualifies as sufficiently large; the reading does not describe its design or results. Its team plans follow-ups on other threat models and methodological pitfalls, and he urges readers to watch, fund and innovate.

Between tiers 2 and 3: correlates of uplift

Suppose you ran a large uplift study, found an effect, and two months later a new frontier generation ships. Can you say anything about its real-world risk without rerunning everything? Maybe, if you collected rich qualitative data on how participants used the models. Then you can map their failures and AI-assisted successes onto benchmarks measuring the matching capability. If model A failed to guide participants

around a common methodological pitfall and model B did not, is that pattern visible in a benchmark task about the method? Is the delta in uplift effectiveness between generations the same as the delta in benchmark results? Götting calls these connections correlates of uplift, after correlates of protection in vaccinology, where easily measured immune markers estimate clinical outcomes. They matter because they give an estimator orders of magnitude cheaper, rerunnable for every new model and, crucially, for pre-release testing and risk assessment.

Tier 4, actual incidents: we hopefully never get here

He corrects himself: uplift studies are not the most conclusive evidence after all. Actual incidents or near-misses would be conclusive for AI-enabled bioterrorism risk. He hopes we never see such deliberate biological mass harm, and notes that even then it may not settle the question: counterfactual attribution, whether it would have happened anyway without AI, is very hard outside an extremely clear-cut case. His guess is that the first offenders will not be complete novices but people who were almost capable already and got a little extra help, which makes decisive proof forensically challenging; he notes the Amerithrax and Aum Shinrikyo investigations took years. (Not from this reading: Amerithrax was the FBI investigation into the 2001 anthrax letters, and Aum Shinrikyo was the Japanese

group behind the 1995 sarin attack on the Tokyo subway.) So we cannot just be reactive. Certainty may come after the fact at a catastrophically high price, and this field has to accept more uncertainty beforehand than domains like drug approval, where a false negative's cost is limited.

Not a proper tier: demos

He considered placing demos, such as individual model transcripts, between theoretical arguments and evals: they can work as an existence proof, or a convincing narrative of risk. But demos should be backed by high-quality evidence rather than replacing it, so he treats them as a way to package and communicate it.

Summary, in his own order

He does not claim evals are pristine evidence, only that they are the best readily available, and that waiting for higher-tier evidence has costs critics rarely price in. His closing bullets land on a destination: a small number of well-designed uplift studies could plausibly anchor a much larger, much cheaper eval-based monitoring regime, which he calls a highly desirable outcome. And people who dismiss evals wholesale are implicitly demanding an evidence quality this field's cost structure and release cadence cannot sustain. He thanks Seth Donoughe, Coleman

Breen, Ben Mueller, Ryan Ritterson, Cassidy Nelson and Claire Qureshi, and keeps the mistakes.

Words to know

- **Eval.** A test run on a model, in software rather than in a lab, that estimates a capability. His definition covers both manual expert interrogation and automated benchmarks, and he means capability evals, not tests of whether a model refuses.
- **Uplift study.** An experiment measuring whether people are more effective and successful at a task when they have AI help. The strongest real-world evidence in this hierarchy, and the most expensive.
- **Construct validity.** Whether a test measures the thing it claims to measure. The core of the critique he is answering.
- **Underelicitation.** Getting less out of a model than is actually in it, because of weak tooling, prompting or fine-tuning. Why eval scores are lower bounds rather than ceilings.
- **Sandbagging and eval awareness.** A model noticing it is under test and hiding what it can do. He says this is not conclusively solved.
- **Specification gaming.** Optimizing the measure instead of the thing measured, once the measure carries consequences. Goodhart's law applied to safety policy thresholds.
- **In silico.** Done in simulation or software rather than in living systems. The defining feature of the

evals tier, and the reason it is cheap.

- **Correlates of uplift.** Benchmark signals that track the real-world uplift measured in a full study, so they can stand in for it on later models. Borrowed from correlates of protection in vaccinology.

Numbers in this reading

- **Four tiers plus one non-tier**: theoretical arguments, evals, real-world uplift studies, actual incidents and near-misses, with demos deliberately left out of the ladder.
- **Tens to hundreds of thousands of dollars**: cost of developing a benchmark.
- **Millions**: cost of a well-powered uplift study. This gap is the argument for evals.
- **A fraction of that cost**: running an existing benchmark on a newly released model, though possibly still significant with heavy inference-time scaling.
- **Orders of magnitude cheaper**: what a validated correlate of uplift would give you compared with rerunning the study.
- **Ten friends and a weekend lab rental**: his sketch of a cheap real-world experiment, which he says is hardly better than anecdata.
- **Annually, not every model release**: the realistic cadence for uplift studies.

- **Two months**: the gap in his thought experiment between finishing an uplift study and the next frontier model generation arriving.
- **One study, the Active Site RCT**: the only uplift study so far he counts as sufficiently large.
- **Seven failure modes** of evals: specification gaming, underelicitation, leakage, construct validity, non-linear threshold effects, infohazard trade-offs, saturation.

What the author is unsure about

He is unusually direct about this. He does not claim evals are pristine or irrefutable. The link between eval performance and harm potential is, in his words, poorly understood. He says it is hard, maybe impossible, to know in advance where capability thresholds that cause sudden real-world jumps sit for biology. Eval awareness and sandbagging have not been conclusively solved. Rule-outs are only as good as the elicitation behind them, and a false negative here is far costlier than a false positive. Even Tier 4, a real incident, might not be conclusive because counterfactual attribution is so hard. Several of the central claims are explicitly his judgment, not findings: that only the Active Site RCT is large enough, that the first offenders will probably be people who were almost capable already, and that a small set of uplift studies could anchor a cheaper eval-based monitoring regime.

For the discussion

- The strongest line in the piece is the asymmetry: in this field a false negative is catastrophically more expensive than a false positive, which means the usual instinct to demand better evidence before acting is itself a risk choice.
- Her own ground connects here: she submitted a public comment to the U.S. Food and Drug Administration on generative AI, in her individual capacity, asking for monitoring after approval, version tracing and a way to roll back, which is close to the eval-based monitoring regime this post wants after release rather than only before it.
- A question worth asking: if a firm could suppress performance on a dangerous capability eval while the underlying capability grows, who would be in a position to notice, and does anything in the current setup give them access?

2.2-1 · CORE · COURSE TIME 30 MIN

Biological Weapons Convention

Wikipedia · 2025

Read the original paper → · the course assigns: the intro, history, summary of key articles, verification and compliance, and implementation support unit sections

> **The world's ban on biological weapons is a short, widely joined treaty with a strong moral**

> **hold and almost no machinery: no way to verify that anyone is keeping it, and a support office of a few staff.**

Average summary

The **Biological Weapons Convention** (BWC, also BTWC) bans the development, production, acquisition, transfer, stockpiling and use of biological and toxin weapons. It was opened for signature on 10 April 1972 and entered into force on 26 March 1975, after 22 states ratified it, including the three depositary governments. It was the first multilateral disarmament treaty to ban the production of a whole category of weapons of mass destruction, and it runs for unlimited duration. As of May 2025 it has 189 states party. Four states signed without ratifying, and four more have neither signed nor acceded.

The article's two-sided judgement is the thing to carry into the discussion. On one side, the BWC established a strong global norm: its preamble calls the use of such weapons repugnant to the conscience of mankind, and no state today says it has biological weapons, seeks them, or thinks using them in war would be legitimate. The biodefence expert Daniel Gerstein calls it the most important arms control treaty of the twenty-first century. On the other side, its effectiveness has been limited by thin institutional support and by the absence of any formal verification regime.

Instead of inspections, the treaty runs on annual **confidence-building measures**, politically binding reports on six topics agreed in 1986 and revised since. Participation sits slightly above 50 percent. A legally binding verification protocol was negotiated for ten years, through an expert group called VEREX and then an Ad Hoc Group that met 24 times between 1995 and 2001. In July 2001 the George W. Bush administration rejected the chairman's 210-page draft and the entire approach behind it, and the process collapsed.

What exists instead is the **Implementation Support Unit**, created by the Sixth Review Conference in 2006 and housed in the Geneva branch of the UN Office for Disarmament Affairs. It began with three full-time staff and a budget the article describes as smaller than the average McDonald's restaurant. For comparison, the Organisation for the Prohibition of Chemical Weapons has about 500 employees and the International Atomic Energy Agency around 2,600. In December 2022 states agreed to add one more full-time post, and only for 2023 to 2027.

The compliance record includes the Soviet programme, which employed up to 65,000 people, and Iraq's programme from around 1985. Article VI, the route to a UN Security Council investigation, has been used once, by Russia in 2022.

Why it's assigned

This is the shared Delay-layer reading that supplies the institutional facts behind the unit's first sub-layer, Deter. Course page 2.4 opens its deterrence option with the claim that the BWC has no verification mechanism and fewer than 5 permanent staff: this article is where that claim comes from, with the history of how the verification protocol died in 2001 and what replaced it. Read it as the answer to the page's own guiding question, why the BWC has no verification mechanism, and as the reason deterrence work now looks to attribution forensics and open-source intelligence instead.

In depth

What the treaty is and what it achieved

The full name is the Convention on the Prohibition of the Development, Production and Stockpiling of Bacteriological (Biological) and Toxin Weapons and on their Destruction. It was signed on 10 April 1972 at ceremonies in London, Moscow and Washington, D.C., and entered into force on 26 March 1975 once 22 states had ratified, including the three depositaries: the Soviet Union, the United Kingdom and the United States. There were 109 signatories. The treaty has no end date. As of May 2025, 189 states are party, with Kiribati the most recent to join. The eight states that are not party are Chad, Djibouti, Egypt, Eritrea, Haiti, Israel, Somalia and Syria; Egypt, Haiti, Somalia and Syria signed but never ratified.

The article credits the BWC with a strong global norm, visible in the preamble's language about the conscience of mankind and in the fact that no state declares that it holds or wants biological weapons or that their wartime use is legitimate. Against that, it names two limits plainly: insufficient institutional support, and no formal verification regime to monitor compliance. It also names the notable violations, the Soviet Union and Ba'athist Iraq, and notes that the Article VI complaint mechanism has been invoked once, by Russia in 2022, over what the article calls the debunked Ukraine bioweapons conspiracy theory.

History

Biological warfare goes back more than six centuries, to the Siege of Caffa in 1346 CE. International restriction began only with the **1925 Geneva Protocol**, which prohibited use but not possession or development. Several countries attached reservations about applicability and retaliation, which left it in practice a no-first-use agreement. It did not stop offensive programmes: the United States ran one from 1943 to 1969, the Soviet Union from the 1920s until at least 1992.

President Nixon ended the American programme unconditionally in 1969 through his Statement on Chemical and Biological Defense Policies and Programs. Its budget at the time was $300 million annually.

The BWC was negotiated in the Conference of the Committee on Disarmament in Geneva from 1969 to 1972, after the nuclear Non-Proliferation Treaty was concluded. Two moves made it possible: a 1968 British proposal to split chemical from biological weapons and do biological first, and the American decision in 1969 to end its own programme and back that proposal. In March 1971 the Soviet Union and its allies dropped their objection to the split and tabled their own draft; on 5 August 1971 the United States and the Soviet Union submitted identical but separate drafts.

The article records continuing calls to modernise the treaty at its Review Conferences. Filippa Lentzos and Gregory Koblentz wrote in 2016 that the crucial contemporary debates the conferences needed to take up included gain-of-function experiments, potential pandemic pathogens, CRISPR and other genome editing technologies, gene drives, and synthetic biology.

The key articles

The BWC has only 15 articles, and has been interpreted and supplemented over time by politically binding understandings reached at eight Review Conferences.

Article I is the core: never, in any circumstances, to develop, produce, stockpile, acquire or retain

biological agents or toxins of types and in quantities with no justification for prophylactic, protective or other peaceful purposes, nor the weapons, equipment or means of delivery designed to use them for hostile purposes. It bans no specific agent. It bans purposes. That approach is called the **general-purpose criterion**, and it also appears in the 1993 Chemical Weapons Convention. Its strength is that it covers every hostile use, including agents not yet invented, and accepts that biological agents are inherently dual use, since the same materials make medicines and vaccines. Its stated disadvantage is that it blurs the line between defensive and offensive research. Viruses were initially an open question, because they sit at the edge of life, but were defined as biological agents in 1969 and fall inside the treaty. Use is not named in Article I, because the Geneva Protocol already covered it, but the Fourth Review Conference in 1996 reaffirmed that use violates the BWC.

Article II requires destroying or diverting weapons and associated resources before joining. Article III bans transferring, assisting, encouraging or inducing anyone else, governments or non-state actors, to acquire them. Article IV requires each state to implement the treaty domestically, which is what lets national authorities investigate, prosecute and punish. Because no single set of measures fits every state, the choice is left to each state's discretion, depending on its legal system, size, geography, biotechnology

industry and regional economic ties. The non-governmental organisation VERTIC maintains a database of over 1,500 national laws and regulations, and launched a further database with the United Nations Institute for Disarmament Research in 2023. The United States implemented the treaty through the 1989 Biological Weapons Anti-Terrorism Act. A 2023 VERTIC report concluded that gaps persist in states' national legal frameworks.

Article V requires consultation and cooperation on disputes; procedures were agreed at the Second Review Conference in 1986 and elaborated at the Third in 1991. Two formal consultative meetings have ever happened, in 1997 at Cuba's request and in 2022 at Russia's. Article VI lets a state complain to the UN Security Council and requires cooperation with any investigation the Council launches. States are generally unwilling to use it because the Council is intensely political and its five permanent members hold a veto, including over investigations. Article VII obliges states to assist a country the Security Council finds has been endangered by a violation; it has never been invoked, though it has drawn more attention lately, partly because of evidence that terrorist organisations are interested in acquiring biological weapons. Article X protects the right to exchange materials, technology and information for peaceful purposes, and says implementing the treaty must not hamper economic or technological development; the

Seventh Review Conference in 2011 set up a database matching offers and requests of assistance.

Verification: what exists

The only regular compliance mechanism is the **confidence-building measures** system agreed at the Second Review Conference in 1986. The reports are politically binding rather than legally binding, and aim to reduce ambiguities, doubts and suspicions. After revisions at the Third, Sixth and Seventh conferences, states report annually on six items: research centres, laboratories and national biological defence programmes; outbreaks of infectious disease and similar occurrences caused by toxins; efforts to promote research results; legislation and regulations; past offensive or defensive programmes; and vaccine production facilities. One item, CBM D, was deleted in 2011. Submissions have grown, but overall participation remains slightly above 50 percent. An online platform launched in 2018, and more states now publish their reports, though many remain visible only to other states.

Verification: what failed

Unlike the chemical and nuclear regimes, the BWC has neither a compliance verification system nor an international organisation to support it. Agreement was not feasible when the treaty was written, largely because of Cold War politics but also because of a

belief that verification was unnecessary and would be hard. The American expert Jonathan B. Tucker's assessment, quoted in the article, is that this missing enforcement has undermined the treaty because it cannot prevent systematic violations.

Compliance provisions were in earlier drafts and were removed; attempts to put them back at the General Assembly in 1971 and by Sweden at the First Review Conference in 1980 both failed. After the Cold War the effort restarted. The Third Review Conference created VEREX in 1991, an expert group that met four times in 1992 and 1993 and examined 21 possible verification measures, including facility inspections and monitoring of relevant publications. A Special Conference in 1994 then created an Ad Hoc Group to negotiate a legally binding protocol. It met 24 times between 1995 and 2001 and drafted a protocol that would have created an international organisation with inspectors who could visit declared facilities routinely and investigate suspect ones. States could not agree on export controls or on how far on-site visits should reach, and by early 2001 the rolling text still had wide disagreements.

In March 2001 the Ad Hoc Group's chairman circulated a 210-page draft. At the 24th session in July 2001 the George W. Bush administration rejected both that draft and the whole approach behind it, and the negotiation collapsed. The United States argued

the protocol would not have improved compliance and would have harmed American national security and commercial interests. Analysts including Matthew Meselson and Amy Smithson criticised the decision as undermining non-proliferation and as inconsistent with the administration's own rhetoric about the biological weapons threat from Iraq and other adversaries.

Since then: at the 2019 Meeting of Experts several states pressed for resuming negotiations, but because others do not support a protocol the article reports the assessment that returning to negotiations is neither realistic nor practicable. The Biden administration appeared to reconsider the American position, with ambassador Bonnie Jenkins asking the 2021 Meeting of States Parties to create a new expert working group on strengthening implementation, transparency and assurance of compliance. In December 2022 states did establish a Working Group on strengthening the Convention, with verification and compliance among its subjects.

Accusations of non-compliance

Because biological weapons programmes are kept intensely secret, the article says it is hard to judge whether activities are legitimate defence or a violation, except in the few cases where evidence is abundant.

The Soviet Union, a party and a depositary, ran the largest, longest and most sophisticated programme, dating to the 1920s under the Red Army and expanded covertly around the time the treaty was signed under the civilian-labelled institution Biopreparat inside the Ministry of Health. It employed up to 65,000 people across several hundred facilities and weaponised the pathogens responsible for smallpox, tularemia, bubonic plague, influenza, anthrax, glanders and Marburg fever. Suspicion began with an anthrax outbreak in Sverdlovsk in 1979 that killed approximately 65 to 100 people; authorities blamed contaminated meat for years, but investigations concluded it came from an accident at a nearby military microbiology facility, and Boris Yeltsin later acknowledged that military developments were the cause. Defectors Vladimir Pasechnik and Ken Alibek strengthened Western concerns in the late 1980s, and George H. W. Bush and Margaret Thatcher raised them directly with Gorbachev. The Trilateral Agreement of 14 September 1992 between the United Kingdom, the United States and Russia was meant to establish what had happened and verify that activities had stopped. David Kelly, who took part in its visits, judged it both a significant achievement, because it produced evidence of Soviet non-compliance from 1975 to 1991, and a dramatic failure, because Russia never fully accounted for the programme it inherited. Milton Leitenberg and Raymond Zilinskas argued in 2012 that parts may continue; as of 2021 the U.S.

Department of State assesses that Russia maintains an offensive programme in violation of Articles I and II. The article also records that in February 2026 scientists at Porton Down found epibatidine in tissue samples from Alexei Navalny's body, that five European foreign ministers issued a joint statement about Moscow's means, motive and opportunity, and that the toxicologist Alastair Hay said Russian responsibility would mean a breach of the BWC as well as the Chemical Weapons Convention.

Iraq, from around 1985 under Saddam Hussein, weaponised anthrax, botulinum toxin, aflatoxin and other agents and built delivery vehicles including bombs, missile warheads and spray systems. It had signed in 1972 but ratified only in 1991, as a condition of the cease-fire ending the Gulf War. The United Nations Special Commission uncovered the programme after that war; Iraq obstructed and deceived the inspectors and admitted the programme only under heavy pressure in 1995. Iraq said it ended in 1991; many analysts believe it ran until at least 1996.

The United States has faced accusations too. Cuba used Article V in April 1997 to allege that American crop-spraying planes introduced the insect Thrips palmi in October 1996; of twelve states that submitted reports, nine found the evidence did not support the allegation and two, China and Vietnam,

called it inconclusive. On 4 September 2001 the New York Times described three American defence or intelligence projects, Project Bacchus, Project Clear Vision and Project Jefferson; the Bush administration called all three defensive, while several international law scholars argued the treaty was violated on two grounds, their exclusion from American confidence-building declarations, and Clear Vision's aim of building a working biological cluster munition against what they read as an absolute ban. A 2003 article in Politics and the Life Sciences by Milton Leitenberg, James F. Leonard and Richard Spertzel argued that parts of the American biodefence programme, particularly at the National Biodefense Analysis and Countermeasures Center, resembled elements of past offensive programmes and risked undermining the prohibitions.

Russia has repeatedly alleged that the United States operates biological weapons facilities in the Caucasus and Central Asia, especially the Richard Lugar Center in Georgia. The State Department calls those allegations groundless; Filippa Lentzos calls them unfounded and part of a disinformation campaign; the Swedish specialists Roger Roffey and Anna-Karin Tunemalm call them a Russian propaganda tool. The Article V consultative meeting Russia convened during the invasion of Ukraine reached no consensus. Going the other way, at the Fifth Review Conference in 2001 the United States charged Iran, Iraq, Libya and

North Korea, plus the signatory Syria, with covert programmes, and a 2019 State Department report raised concerns about China, Russia, North Korea and Iran, concluding that North Korea has an offensive programme in violation of Articles I and II and that Iran has not abandoned developing biological agents and toxins for offensive purposes.

The Implementation Support Unit

When the protocol effort died in 2001, it took with it the planned Organization for the Prohibition of Biological Weapons. The Sixth Review Conference in 2006 created the **Implementation Support Unit** instead, funded by states parties and housed in the Geneva branch of the UN Office for Disarmament Affairs. Its mandate is administrative support, help with national implementation, encouraging universal membership, matching assistance requests with offers, and running the confidence-building measures process.

It started with three full-time staff and a budget the article describes as smaller than the average McDonald's restaurant, and it does not compare with its chemical or nuclear counterparts: the Organisation for the Prohibition of Chemical Weapons has about 500 employees, the International Atomic Energy Agency around 2,600, and the CTBTO Preparatory Commission around 280. In December 2022 states

parties agreed to one additional full-time position, and only for the period 2023 to 2027.

Words to know

- **States Parties**: the countries that have legally joined a treaty, as opposed to those that only signed it. In the BWC, signing stopped being possible after 1975, so new members accede or succeed instead of ratifying.
- **Depositary**: the government that formally holds a treaty and receives new members' paperwork. The BWC has three, the United States, the United Kingdom and Russia, which is why joining it means depositing an instrument with one of them.
- **General-purpose criterion**: the drafting approach of banning purposes rather than listing banned agents. It is why the BWC still covers pathogens invented decades after 1972, and also why defensive and offensive research are hard to tell apart.
- **Confidence-building measures (CBMs)**: annual reports states exchange voluntarily in political rather than legal terms. They are the BWC's only routine compliance information, and only slightly more than half of states file them.
- **Verification protocol**: the legally binding add-on, negotiated from 1995 to 2001, that would have created inspectors and an organisation. Its

rejection is the single event that explains the treaty's present shape.

- **Review Conference**: the meeting held every few years where states reinterpret and supplement the treaty. Because the BWC's own text is only 15 articles, almost everything added since 1975 came from these conferences.
- **Implementation Support Unit (ISU)**: the small UN office in Geneva that administers the treaty. It is not an inspectorate and has no compliance powers.

Numbers in this reading

- 10 April 1972: the treaty opened for signature, in London, Moscow and Washington, D.C.
- 26 March 1975: entry into force, after ratification by 22 states including the three depositaries.
- 189 states party as of May 2025; 109 signatories; 4 states signed but did not ratify; 4 neither signed nor acceded.
- 15 articles in the whole treaty, supplemented by understandings from 8 Review Conferences.
- 193 parties to the Chemical Weapons Convention and 191 to the nuclear Non-Proliferation Treaty, the comparison the article uses to call BWC universality low.
- Slightly above 50 percent: the overall participation rate in annual confidence-building measures.

- 21 verification measures examined by VEREX across four meetings in 1992 and 1993.
- 24 Ad Hoc Group sessions between 1995 and 2001, and a 210-page draft protocol circulated in March 2001.
- 3 full-time ISU staff at the start, plus 1 post added in December 2022 for 2023 to 2027 only; against about 500 at the OPCW, around 2,600 at the IAEA, and around 280 at the CTBTO Preparatory Commission.
- $300 million: the annual budget of the American offensive programme when Nixon ended it in 1969.
- Up to 65,000 people in several hundred facilities: the size of the Soviet programme.
- Approximately 65 to 100 deaths in the 1979 Sverdlovsk anthrax outbreak.
- Over 1,500 national laws and regulations in VERTIC's database of domestic implementation.
- 1 Article VI complaint ever lodged, by Russia in 2022; 2 Article V consultative meetings ever held, in 1997 and 2022; Article VII never invoked.

What the author is unsure about

Being an encyclopaedia article, it reports other people's judgements rather than making its own, and it is careful about what is known. It states directly that the secrecy around biological weapons programmes makes it challenging to assess whether a country's

activities are legitimate defence or a violation, except where evidence is abundant. It frames Russia's possible continuation of the Soviet programme as an assertion by Leitenberg and Zilinskas and as an assessment by the U.S. State Department, not as established fact. It presents the claim that returning to protocol negotiations is neither realistic nor practicable as a quoted assessment, and the Biden administration's shift as something that seems to be happening. Where accusations are contested, as with Cuba's 1997 complaint and Russia's claims about Georgia and Ukraine, it gives both sides and says where no consensus was reached.

For the discussion

- The treaty's deterrent is a norm, not a mechanism: 189 states party, no state admitting it wants these weapons, and still no inspector anywhere, which is worth stating plainly before the group debates how to strengthen deterrence.
- The 2001 collapse is the hinge, and it was not a failure of technical imagination: VEREX had already identified 21 measures and a 210-page draft existed, so the obstacle was political will, which is a different problem from the one that better forensics would solve.
- Worth asking the group: if the Implementation Support Unit has four staff, who is actually supposed to notice a violation, and does the

general-purpose criterion, which bans purposes rather than agents, make that judgement harder for whoever is watching?

2.2-2 · CORE · COURSE TIME 15 MIN

Why do we take LLMs seriously as a potential source of biorisk?

Anthropic · 2025

Read the original paper → · the course assigns: whole piece

> **Anthropic answers a sceptic who asks why a general-purpose chatbot belongs in a biosecurity conversation, and its answer is that the models now match or beat specialists on dangerous biology knowledge while the material barriers that used to stop amateurs are falling.**

Average summary

This is a company explaining its own reasoning, not a neutral study. Anthropic starts from the upside: it names **Benchling**, using Claude to help researchers structure data and ask better questions, and **Biomni**, using Claude to speed up bioinformatics analysis and automate experimental design. Then it makes its central framing claim, that AI is fundamentally a dual-use technology, so identifying, measuring and mitigating misuse is part of building it responsibly.

The concrete trigger was the release of Claude Opus 4, when Anthropic activated **AI Safety Level 3 (ASL-3)** protections, including deployment measures narrowly aimed at blocking help with certain chemical, biological, radiological and nuclear tasks. The company is explicit that this was precautionary: as the model's performance on their evaluations improved, they could no longer confidently rule out that it could uplift someone with a basic STEM background.

The piece is organised as a set of sceptic's questions. Why would AI matter at all? Because terrorist groups have rapidly adopted encrypted communications, cryptocurrency and social media, and because AI is different from a search engine: the internet is cluttered with contradictory and misleading information and cannot help in real time, while a capable model can act as a guide to reliable information and a source of tacit knowledge. Why biology rather than cyber or fraud, which Anthropic also works on? Because a virus can spread far beyond its initial target, and because falling nucleic acid synthesis costs, standardised reagent kits and easy access to standard equipment have weakened what used to be a passive biodefence.

Don't experts know more? Not reliably any more. Within a year Claude went from underperforming world-class experts on a virology troubleshooting

evaluation to comfortably exceeding that baseline, and the human baseline was specialists answering questions in their own speciality.

Does that matter in practice? In controlled trials, participants had up to two days to draft a bioweapons acquisition plan; those with Claude 4 models, especially Opus 4, scored much higher and made substantially fewer critical failures than an internet-only control group. A 2024 wet lab pilot found no uplift, but everyone did surprisingly well, including the control group, which hints that tacit knowledge may be a weaker barrier than assumed. That pilot had just **8** participants, and a larger study with the Frontier Model Forum and Sentinel Bio is under way.

Why it's assigned

This is the shared Delay reading that justifies the whole unit's focus on AI. Page 2.1 promises to look at what models can and cannot currently do and how we measure it, and this piece shows a developer's own measurements and the decisions they triggered. It also sits directly under the Deny sub-layer: the deployment classifiers, the ASL-3 threshold and the nucleic acid synthesis screening it endorses are exactly the model-level and infrastructure-level safeguards that deep-dive options 3 and 4 take apart.

In depth

The frame: dual use, stated by the developer

The piece opens with the benefit side, because its argument depends on both halves being real. Anthropic says its work is animated by AI's potential to advance scientific discovery, especially in biology and medicine, and names two customers: Benchling, helping researchers structure data, ask better questions and spend more time on science, and Biomni, speeding up bioinformatics analysis and automating experimental design. Then the turn: AI is fundamentally a dual-use technology, and a key tenet of developing it responsibly is to identify, measure and mitigate the prospects for malicious actors to misuse the very capabilities that make it useful.

The decision that prompted the post was the release of Claude Opus 4, when Anthropic activated AI Safety Level 3 protections. Those included deployment measures narrowly focused on preventing the model from assisting with certain tasks related to chemical, biological, radiological and nuclear weapons development. Anthropic describes this as precautionary: improving model performance on their own evaluations meant they could no longer confidently rule out that their most advanced model could uplift people with basic STEM backgrounds who tried to develop such weapons. Because of the potential consequences, biological weapons were the major initial focus.

The authors acknowledge that this looks odd. Every safety framework released by frontier AI labs includes some reference to biorisk, even though frontier large language models are generalists rather than biology specialists, unlike foundation models such as AlphaFold, and even though plenty of other security threats could be prioritised instead. They say they understand the scepticism, and that the aim is not alarmism: discussions of AI and biorisk sit firmly in the category of low-probability, high-impact scenarios.

What AI has to do with weapons at all

Two arguments, one about continuity and one about difference. The continuity argument: in recent years terrorist groups have rapidly adopted encrypted communications, cryptocurrency and social media, and there is no reason to expect anything different with AI. People seeking weapons information already moved from physical pamphlets and manuals to internet searches, so they will query AI next.

The difference argument is the one Anthropic leans on. The internet is cluttered with contradictory and misleading information, and a website cannot provide real-time assistance when someone hits trouble in a complex, unfamiliar process. A sufficiently advanced model may act as three things at once: a guide to reliable information, a source of otherwise tacit and inaccessible knowledge about implementation, and a

research assistant that can process data and generate insight immediately.

Why biology rather than everything else

Anthropic says biological risk is not its only concern, and that it invests considerable resources in the implications for cybersecurity and in gathering intelligence about actual platform misuse for fraud, malware development and influence operations.

Two factors make biorisk especially concerning to them. First, the consequences of a successful biological attack are unusually severe, because an attack with a virus could spread far beyond the initial target in a way that is qualitatively different from weapons with more local effects. Second, improvements elsewhere in biotechnology may have lowered material barriers that previously acted as a passive biodefence: the decreasing cost of nucleic acid synthesis, standardisation of reagent kits, and easy access to standard molecular biology equipment such as PCR machines are all making material acquisition less of a bottleneck. AI then reduces the remaining barriers to information and know-how. High consequences plus increasing plausibility is the combination that makes it a priority.

Knowledge: the models are catching the specialists

Language models are trained on vast amounts of data ranging from financial models to fanfiction, so they learn something about almost everything, including protein structures, genetic engineering techniques, virology and synthetic biology. What Anthropic calls surprising is the degree to which this biological knowledge has scaled up as models improved.

They are careful about the limit: Claude and other models cannot currently do science autonomously at an expert level. But on several evaluations testing knowledge relevant to potentially dangerous aspects of biology, models are nearing and sometimes exceeding expert human performance, a pattern they report across multiple benchmarks and especially in molecular biology.

The detail that gives this its force is the baseline: the human comparison is scientists answering questions within their own speciality. So models are not only developing a breadth that human experts struggle to match, they are in some cases outpacing specialists on their own ground.

The figures

Figure 1 charts Claude's performance over time on the **VCT**, an evaluation of a model's ability to troubleshoot virology tasks, designed by SecureBio; the shape of it is the point the text makes in words, that within a year Claude moved from below the world-class expert

baseline to comfortably above it. Figure 2 reports the bioweapons acquisition uplift trial in two panels, raw scores on top and the number of critical failures participants hit underneath, so the same trial is shown both as how good the plans were and as how often they contained a fatal flaw. The saved text includes the captions and the surrounding argument but not the plotted values, so the exact axis numbers are not available here.

Applied scenarios: the uplift trials

Anthropic argues that performance on a quiz is not enough. You need evaluations involving real people that closely mirror the actual threat scenarios, and you need to measure the model against the most easily accessible alternative, which is the internet.

Their controlled trials gave participants up to two days to draft a comprehensive bioweapons acquisition plan. The control group had access to basic internet resources only. The model-assisted group also had access to Claude with safeguards removed, which Anthropic defends as necessary in limited cases so that trusted parties can assess the technology's maximum capabilities. Biodefence experts then graded the plans against a detailed rubric covering key steps of the acquisition pathway. Participants with Claude 4 models, especially Claude Opus 4, received much higher scores and produced

plans with substantially fewer critical failures than the internet-only control group.

The caveat is theirs and it is firm: text-based uplift trials are imperfect proxies for the real world, which adds tacit knowledge, materials access and actor persistence. What the trials demonstrate is the underlying plausibility that AI could give a non-expert malicious actor differentially superior assistance, not that it would work in practice.

The laboratory question, still open

Anthropic separates three things that are easy to blur: answering multiple-choice questions about virology, helping with text-based acquisition plans, and actually helping someone in a laboratory.

Two objections sit in the way. One is tacit knowledge: skills hard to put into words, such as the visual cues that tell you when a reaction has reached a particular point, plus received wisdom like the parts of a protocol nobody writes down. If those barriers defeat any novice, the uplift worry shrinks. The other is evidence: validating the concern needs large-scale evidence from an actual laboratory experiment.

In 2024 Anthropic ran a preliminary wet lab uplift trial using basic biology laboratory protocols, with some participants given a version of Claude and others only the internet. They observed no evidence of uplift. But they also noticed that all participants, including the

internet-only group, did surprisingly well on every task. Their reading is that the pilot provides no strong evidence either way about AI uplift, while suggesting that tacit knowledge may not be the bottleneck many have claimed. They flag the limits plainly: the study had 8 participants and the protocols were fairly basic.

The follow-up is a larger study co-sponsored through the **Frontier Model Forum** with the pandemic prevention non-profit Sentinel Bio, designed to test whether AI can uplift non-experts to the point of carrying out expert-level laboratory biology tasks. The control group's performance will indicate how much tacit knowledge limits non-experts, and the larger sample should give a clearer measure of uplift. Anthropic says the results will be important to its own assessment of risk, which is an admission that the assessment is not settled.

What can be done

Three headings. On **information sharing**, Anthropic says its current focus is evaluating whether models can help non-experts produce biological threats, informed by consultations with biodefence experts, while recognising this is far from the only scenario. Governments hold differentiated knowledge about threat actors' plans and intentions, and Anthropic reports benefiting from voluntary collaboration and pre-deployment testing with the US Center for AI Standards and Innovation and the UK AI Security

Institute, calling public-private partnership a promising path. It also backs transparency norms: developers publishing a Secure Development Framework, as Anthropic's Responsible Scaling Policy is, and publishing dangerous capability evaluation results.

On **deployment safeguards**, the company describes automated guards built on its constitutional classifiers research, which monitor inputs and outputs in real time and block a narrow class of potentially harmful information. These apply only to models whose potential to meaningfully uplift threat actors cannot be ruled out, which to date means the Claude Opus 4 family, again as a precaution. Separately, a Safeguards team monitors platform activity across all models for signs of misuse, including dangerous applications of biology, which informs enforcement action.

On **national action**, Anthropic welcomes biosecurity's place in the new US AI Action Plan, singling out national security evaluations through the Center for AI Standards and Innovation and the move to bolster nucleic acid synthesis screening.

The closing position is deliberately uncertain. The topic is rife with uncertainty about threat actors, about model capabilities, and about how exactly those capabilities translate into risk. What Anthropic claims is only that a growing body of evidence and the

gravity of biological threats make this something developers and policymakers must attend to.

Words to know

- **Dual-use**: useful for both legitimate and harmful ends. It is the reason safeguards here have to be narrow, since the same capability serves researchers and misusers.
- **AI Safety Level 3 (ASL-3)**: a tier in Anthropic's own internal safety framework, which triggers extra deployment protections. It matters because the trigger was not evidence of harm but the loss of confidence that harm could be ruled out.
- **Uplift**: how much a tool improves what a person could otherwise achieve alone. The whole argument turns on uplift relative to the internet, not on whether a model knows dangerous things.
- **Tacit knowledge**: skill that is hard to write down, like the visual cues that signal timing, or unwritten steps in a protocol. If it is a hard barrier, novices stay blocked; Anthropic's pilot hints it may be softer than assumed.
- **Constitutional classifiers**: the research behind Anthropic's automated guards, which watch inputs and outputs in real time and block a narrow class of content.
- **Nucleic acid synthesis screening**: checking orders for synthesised DNA or RNA before they

ship. Anthropic names it as a national measure that complements model-level controls.

- **Secure Development Framework**: a published document in which a developer commits to how it will handle dangerous capabilities. Anthropic's version is its Responsible Scaling Policy.

Numbers in this reading

- 8: participants in the 2024 preliminary wet lab uplift trial, which Anthropic itself calls very small.
- Within a year: how long Claude took to go from underperforming world-class experts on the virology troubleshooting evaluation to comfortably exceeding that baseline.
- Up to two days: the time participants were given to draft a bioweapons acquisition plan in the text-based uplift trials.
- Claude Opus 4 family: the only models to date carrying the ASL-3 deployment classifiers.

What the author is unsure about

A great deal, and the piece says so repeatedly. It places AI biorisk in the low-probability, high-impact category rather than claiming an attack is likely. It states that models cannot currently do science autonomously at expert level. It calls text-based uplift trials imperfect proxies for the real world, which involves tacit knowledge, materials access and actor persistence. It reports the wet lab pilot as providing

no strong evidence either way, and names its sample size and the simplicity of its tasks as the reasons. It says the ASL-3 decision was precautionary, resting on an inability to rule out uplift rather than on a finding of it. And it closes by naming three live uncertainties: who the threat actors are, what the models can do, and how capabilities translate into risk. The forward-looking claims about the Frontier Model Forum study are plans, not results.

For the discussion

- The decision rule here is worth naming out loud: Anthropic turned on its strictest deployment controls not because it found the model dangerous but because it could no longer prove it was not, which is a precautionary standard that most of biosecurity policy does not yet use.
- The most interesting finding is the one that cuts against the company's own worry: the wet lab pilot showed no uplift, and the control group did surprisingly well, which undercuts tacit knowledge as a reliable barrier rather than supporting AI as the threat.
- A tension worth raising: the uplift trials required giving a group access to a model with safeguards removed, so measuring this risk means temporarily creating it, and the group should ask who gets to authorise that and who sees the results.

2.2-3 · CORE · COURSE TIME 20 MIN

Evaluating Safety and Security in Biological Research

Adeline Williams, John Tarangelo, Barbara Del Castello, Rebecca Moritz, Saskia Popescu, Matt Sharkey · 2025

Read the original paper → · the course assigns: the first 7 pages, with a closer look at the framework on page 7

> **American policy on risky biological research has named the categories of concern since 2004 but never built a standard way to score any particular project, and this RAND team proposes one: a risk score and a benefit score, applied again and again across the life of a study.**

Average summary

The problem the authors set out is old. In 2004 the National Research Council's **Fink Report**, *Biotechnology Research in an Age of Terrorism*, coined **dual-use research of concern** (DURC), research that is useful for legitimate science but could be put to malicious ends, and listed seven categories of experiments of concern that the scientific community should review before anyone undertakes them. Since then, recombinant DNA technology, gene editing tools and more recently artificial intelligence have made the problem bigger, and the US government has issued policy after policy:

2012, 2014, a 2017 framework at Health and Human Services for funding decisions about research involving enhanced potential pandemic pathogens, and the 2024 DURC/PEPP policy.

On **May 5, 2025**, Executive Order 14292, "Improving the Safety and Security of Biological Research," directed that the 2024 policy be revised or replaced. The order defines dangerous gain-of-function research as research on an infectious agent or toxin with the potential to cause disease by enhancing its pathogenicity or increasing its transmissibility, and lists seven qualifying research activities in Section 8.

The authors' complaint is that all of these criteria are subjective, so judging the real consequences of a given project in a standardised way is still hard. Their answer is a pair of ordinal scoring matrices. The **risk score** estimates how bad the consequences would be if there were a biosecurity breach, meaning misuse, or a biosafety breach, and rests on five factors: human health considerations, transmissibility characteristics, medical countermeasure availability with immune recognition and ability to control, societal and environmental implications including agricultural, economic and national security concerns, and other risks. The **benefit score** estimates what the research offers science and society, across seven factors: medical countermeasure availability and development, scientific knowledge generation, disease surveillance

and preparedness, public health policy, One Health advancement, technology development, and other benefits. Each factor has two to four indicators, the indicators add up, and the totals are normalised so the two scores can be compared against each other.

They are deliberately modest about it. The scales come from published work but warrant further validation, ideally by funding agencies or government experts with more data. The tool sets no thresholds, does not weight some factors more heavily than others, and contains no criteria specific to particular pathogens. It would never be the only way to decide whether a project counts as dangerous gain-of-function research. What they recommend is that the Office of Science and Technology Policy and other agencies build a more robust version through an expert working group led by the Executive Office of the President.

Why it's assigned

This is the Delay-layer reading about who decides, and on what basis. The unit's Deny sub-layer is about denying access to dangerous knowledge, and the first point where anyone can deny anything is the moment a research project is proposed and reviewed. The piece shows the American oversight machinery from the inside, the institutional review entities, the funders, the executive orders, and shows that after twenty years it still lacks a standard scoring method.

That makes it the natural companion to deep-dive option 2 on regulating discovery, and the course page's guiding question about what process individual scientists use to judge their own risk-benefit ratio.

In depth

Where the idea of DURC came from

In 2004 the National Research Council of the National Academies of Sciences, Engineering, and Medicine published *Biotechnology Research in an Age of Terrorism*, known as the **Fink Report**, in response to growing concern about misuse of biological research. It introduced the concept of dual-use research of concern, and identified seven categories of experiments of concern that would warrant review and discussion within the scientific community before being undertaken. As the report lists them, those are experiments that would render a vaccine ineffective; confer resistance to therapeutically useful antibiotics or antiviral agents; enhance the virulence of a pathogen or render a non-pathogen virulent; increase the transmissibility of a pathogen; alter the host range of a pathogen; enable evasion of diagnostic and detection methods; or enable the weaponisation of a biological agent or toxin.

Since 2004, the authors write, advances such as recombinant DNA technology, gene editing tools and more recently artificial intelligence have amplified dual-use concerns. The US government responded with a sequence of guidance documents rather than a single rule: the 2012 United States Government Policy for Oversight of Life Sciences Dual Use Research of Concern, the 2014 policy for institutional oversight, the Department of Health and Human Services framework from 2017 for guiding funding decisions about proposed research involving enhanced potential pandemic pathogens, and the 2024 implementation guidance covering both DURC and **pathogens with enhanced pandemic potential** (PEPP).

The executive order that reset the field

On May 5, 2025, Executive Order 14292, "Improving the Safety and Security of Biological Research," directed that the 2024 DURC/PEPP policy be revised or replaced. The authors say this order will shape the future of the policy, and that an important next step is clarifying exactly what research the new policy covers.

The order's own definition of dangerous gain-of-function research is scientific research on an infectious agent or toxin with the potential to cause disease by enhancing its pathogenicity or increasing its transmissibility. Section 8 then lists seven research activities that qualify, which the authors note are

closely related to the experimental effects already named in the federal and institutional DURC policies and in the Fink Report. The order's main stated aim is safeguarding US citizens from this category of research, and its provisions include four things: revise or replace the 2024 guidance so the covered scope is clear; develop guidance to end federal funding for dangerous gain-of-function research in countries of concern; develop a strategy to govern, track and limit non-federally funded dangerous gain-of-function research inside the United States that could cause significant societal consequences; and develop a reporting mechanism through which recipients of federal research funds, and non-federally funded researchers to the extent the law permits, report such research.

The authors' diagnosis is that even with all of this, evaluating the true consequences of a project from subjective criteria alone, in a standardised way, remains challenging. Section 4 of the order, "Secure Future Research Through Commonsense Frameworks," is the hook they hang their proposal on.

The two scores

They define ordinal scoring matrices, meaning ranked bands rather than precise measurements.

The **risk score** estimates how impactful the consequences of a research project would be in the

event of a biosecurity breach, which they gloss as misuse, or a biosafety breach. Built on existing US government DURC/PEPP policy and the risk assessment literature, it rests on five factors: human health considerations; transmissibility characteristics; medical countermeasure availability, immune recognition and ability to control; societal and environmental implications, including agricultural, economic and national security concerns; and other risks.

The **benefit score** estimates the value the research may bring to the scientific community and to society, and rests on seven factors: medical countermeasure availability and development; scientific knowledge generation; disease surveillance and preparedness; public health policy; One Health advancement; technology development; and other benefits.

For each risk and benefit factor the authors defined two to four indicators, which are added up to produce the overall score. The scores are then normalised and compared relative to each other, so that a project's risk and benefit can be set side by side despite the two scales having different numbers of factors. The scoring scales were developed from published works, and the authors state directly that they warrant further validation, especially by funding agencies or other government experts who may hold more extensive data. An annex with a companion risk-benefit

calculator and scoring sheet is published separately at www.rand.org/t/PEA4125-2.

Where the tool would sit in the existing system

The proposal is explicitly additive, not a replacement. It could supplement the National Institutes of Health's Tools for the Identification, Assessment, Management and Responsible Communication of Dual Use Research of Concern, the 2014 companion guide of institutional best practice.

The existing DURC policy requires institutions to form an **institutional review entity** (IRE), which writes institutional policy and reviews potential DURC research from the grant proposal stage through to manuscript review. The authors see the scoring system being used throughout that life cycle by IREs working with principal investigators, subject-matter experts, the DURC/PEPP committee within Health and Human Services, funding agencies, and what they call a future bioresponsibility agency designed specifically for research oversight. Those bodies could use the scores, and how the scores relate to each other, as one input in deciding whether research should be funded, how much risk mitigation to require or amend, and whether alternative experimental approaches should be considered instead.

They also identify a concrete gap the tool would fill. The 2024 DURC/PEPP policy required researchers to

submit an assessment of their own research at the proposal stage, but to the authors' knowledge no publicly available standardised template or process exists for those initial assessments.

Their sketch of how it would run: a researcher and a funder each complete an assessment at proposal stage; if the project is funded, both assessments go to the institution's IRE for review; a risk mitigation plan is then developed, revised if needed, and approved alongside the funder's terms and conditions. As data accumulate, the IRE periodically re-evaluates. If the balance begins to shift toward unacceptable risk, the institution would be required to notify the funder, with the specific triggers written explicitly into the funding terms. The examples the authors give of such triggers are increased virus replication in laboratory or animal systems, a change in transmission dynamics, and incident notifications. A final assessment could also happen before publication or data release.

They argue the framework would help people already inside the process: researchers could use it during project design to anticipate how funders or review entities will see their work, to explore lower-risk alternative approaches, and to propose mitigations early. They add a practical benefit, that it could streamline review and reduce the delays caused by rounds of revision passing back and forth between investigators and funders. Although it was developed

in the US system, they say it could be adapted for other countries.

The figure

Figure 1, on page 7, is the whole proposal on one page, and it reads left to right. On the left, the risk factors are each tagged with their role in a hazard-vulnerability-consequence framework: human health considerations as consequence, transmissibility characteristics as hazard, impacts on medical countermeasures and immunity as vulnerability, and societal and environmental implications as consequence. These feed a single risk score. Below them the benefit factors, medical countermeasure availability and development, scientific knowledge generation, disease surveillance and preparedness, public health policy, One Health advancement and technology development, feed a benefit score. Three notes beside the scores carry the mechanics: each risk factor has multiple indicators, each indicator has a standardised scoring scheme, and scores are normalised to account for the different number of risk and benefit factors. On the right, the two scores flow into a research life cycle with four stages, and the figure lists what the scores decide at each one. At grant proposal and experimental design: designate the work as DURC/PEPP, consider safer research methods, propose risk mitigation strategies, make funding decisions. At data generation and again at

manuscript review: use the evolving data for an updated risk-benefit analysis, update the DURC/PEPP designation, and adapt mitigations if necessary. At responsible communication of data: identify information hazards and redact if necessary, and apply lessons learned to future research decisions. A footnote says "other risks" and "other benefits" are part of the scoring framework but are not drawn in the figure. What to notice is that the arrow does not stop at the funding decision: the same two scores are recalculated at four points, which is the authors' real argument about oversight.

The caveats, stated as a list

The authors devote a section to what their tool cannot yet do.

On purpose and scope, the tool is meant for research directly involving any human, animal or plant pathogen, toxin, agent or technology reasonably anticipated to produce one or more of the seven experimental effects in Section 8 of the order. Impacts across human, animal and plant sectors are included, but they say additional considerations for animals and plants could be added. Crucially, even finished, the tool would not be the sole method for deciding when research meets the order's definition of dangerous gain-of-function, because that will depend on specific experimental contexts that may need tailored discussion or further assessment.

On validation, the categories, indicators and scoring rules warrant further validation through more reviews of real high-consequence proposals, including the risk assessments institutions and agencies already perform before funding. Applying the tool to a larger set of studies, by a larger and more diverse group of experts, would improve how well it generalises.

On score weighting, the tool sets no thresholds for acceptable risk or benefit, and does not account for cases where some factors should count more than others. The authors expect use and validation to reveal which cases those are.

On pathogen-specific considerations, the tool as written contains no criteria specific to pathogen species or families. Defining them would make the scoring more objective and quantitative, but could also greatly change the scores, because of the innate differences between pathogen groups.

Their recommendation follows from those gaps: the Office of Science and Technology Policy and relevant agencies should build on this foundational work to develop a more robust tool and decisionmaking framework, through an expert working group led by the Executive Office of the President, coordinating with interagency partners and, where appropriate, non-federal stakeholders. They say twice that such a group needs members with scientific knowledge and practical hands-on expertise in implementing

biosafety and biosecurity measures, both to validate the tool and to judge whether the research community will actually adopt it. The pages after the assigned seven go on to define each factor, indicator and measurement in detail.

Words to know

- **Dual-use research of concern (DURC)**: research that is genuinely useful for science and could also be turned to harm. The Fink Report named it in 2004, and every American policy since has been an attempt to govern it.
- **Pathogens with enhanced pandemic potential (PEPP)**: the category covering work that could make an agent more dangerous at population scale. It travels alongside DURC in US policy, and earlier documents called it enhanced pandemic potential pathogens.
- **Gain-of-function research**: in Executive Order 14292's own words, research on an infectious agent or toxin with the potential to cause disease by enhancing its pathogenicity or increasing its transmissibility. The definition matters because it sets what the policy covers.
- **Institutional review entity (IRE)**: the committee each institution must form to write local policy and review potential DURC. It is the body that would use this tool most often.

- **Ordinal scoring matrix**: a grid that places a project in ranked bands rather than giving a precise number. It is honest about being a judgement aid, not a measurement.
- **Medical countermeasure (MCM)**: a vaccine, drug or diagnostic that can prevent, treat or detect a disease. Whether one exists appears on both sides of the ledger here, as a factor reducing risk and as a benefit the research might produce.
- **One Health**: the approach that treats human, animal and environmental health as one system. It appears as a benefit factor, which is how agricultural and ecological value enters the score.
- **Information hazard**: knowledge that causes harm simply by being known or published. It is the thing the final life cycle stage asks reviewers to look for, with redaction as the remedy.

Numbers in this reading

- 2004: the Fink Report, which introduced DURC and the seven categories of experiments of concern, at pages 114 to 115.
- 7: categories of experiments of concern in the Fink Report, and separately the 7 research activities listed in Section 8 of Executive Order 14292.
- May 5, 2025: the date of Executive Order 14292, which directed that the 2024 DURC/PEPP policy be revised or replaced.

- 5 risk factors and 7 benefit factors in the proposed tool, with 2 to 4 indicators defined for each factor.
- 4 research life cycle stages in Figure 1: grant proposal and experimental design, data generation, manuscript review, and responsible communication of data.
- Policy sequence cited: 2012, 2014, 2017 and 2024 federal guidance documents, plus the 2014 NIH companion guide.

What the author is unsure about

The authors are unusually frank, and the uncertainty is the point of the piece rather than an afterthought. They say the scoring scales warrant further validation, preferably by funders or government experts with better data. They say the tool sets no thresholds and no weighting, and that real use will reveal which factors should carry more weight. They say it has no pathogen-specific criteria, and that adding them would improve objectivity but could greatly alter the scores. They say some factors, societal and environmental implications in particular, are inherently qualitative and would benefit from future work to find more direct measurements. They say early-stage assessments will necessarily be more qualitative and anticipatory than later ones. And they state plainly that even in finished form this would not be the sole method for deciding whether a project is dangerous gain-of-function research. Everything

about the proposed workflow, the funder notification triggers, the bioresponsibility agency, is written as what could happen, not what does.

For the discussion

- The useful point to make is that twenty-one years after the Fink Report named the categories of concern, there is still no publicly available standard template for the assessment the 2024 policy required researchers to submit, so the gap in the Deny layer is not a missing rule but a missing method.
- The strongest feature of their design is that the same scores are recalculated at four stages, including before publication, which treats oversight as continuous rather than as a gate at the funding decision.
- A tension worth raising: the tool deliberately sets no risk threshold, so it can tell a committee that one project scores higher than another but never that a project is unacceptable, and the group could ask whether that is appropriate humility or a way of leaving the hardest call unmade.

2.2-4 · CORE · COURSE TIME 20 MIN

How to stop bioterrorists from buying dangerous DNA

Steph Batalis, Vikram Venkatram · 2025

Read the original paper → · the course assigns: whole piece, from a detailed structured account of the article rather than its original text

> **Screening DNA orders can reliably stop the obviously unqualified buyer, but it cannot read intent, so asking a synthesis company to spot the malicious researcher with perfect paperwork is asking for something screening cannot deliver.**

Average summary

Summarized from a detailed structured account of the article, not its original text, so nothing here is a direct quotation.

Steph Batalis and Vikram Venkatram, both at Georgetown's Center for Security and Emerging Technology, put the reader in a chair: you are the chief screening officer at a gene synthesis company, and six customers are waiting outside your office. All six have ordered DNA from a federally regulated pathogen or toxin. Your job is to decide who gets their order.

The setup matters because of where US policy sits. Dozens of gene synthesis companies worldwide make custom DNA strands for academic and industry customers, and that custom DNA has powered genetic engineering, new diagnostics and therapies, and gene-function research. Some orderable sequences

code for pathogen or toxin genes, so experts from industry, government and academia recommend screening both the sequence and the customer. A Biden administration executive order required federally funded researchers to buy only from companies that screen, which would have been the first time screening was standardised and compulsory. The Trump administration rescinded that requirement less than three months after it took effect. Many leading companies still screen voluntarily through an industry-led consortium.

The authors' diagnosis is that the real shortcoming is the absence of clear guidance or decision criteria for screening **customers**. Screening sequences alone is not enough, because genuinely risky sequences are needed for disease research and vaccine design. But sorting customers into legitimate and illegitimate is subjective and often ambiguous.

The six customers come in three pairs, each pair identical on paper, one member harmless and one malicious. The first pair, a high school teacher and a man with no laboratory background, are easy to deny, and intent never enters the decision: neither has experience with dangerous agents or access to the containment needed to work safely. The second pair, two doctoral researchers with appropriate containment and plausible research questions, are the genuine gray area, and today the provider decides

alone with no universal guidance. The third pair co-lead a national laboratory vaccine project with full documentation, biosafety sign-off and oversight by the Federal Select Agent Program; both orders should ship, because a commercial provider has no way of telling them apart.

Hence three recommendations: build standards around the clearly illegitimate, give providers an answer key for the gray areas, and remember that screening is one layer among several.

Why it's assigned

This is the Delay-layer reading that shows the chokepoint everyone reaches for, and its limits. Course page 2.7 frames the Deny sub-layer around the idea that a genome sequence is only a threat if someone can print it, which makes synthesis screening the obvious place to block. Batalis and Venkatram supply the counterweight: screening is good at excluding the unqualified, poor at reading intent, and only worth building if the institutional and law enforcement layers around it are built too. It is the best possible warm-up for the deep dive on safeguarding research infrastructure, and for the course page's own question about why know-your-customer checks are common in finance but not biotechnology.

In depth

What the piece is doing

This is an explainer and an argument rather than a study. The authors are Georgetown CSET researchers, Batalis a research fellow leading biotechnology work and Venkatram a research analyst on emerging technology. They use a role-play, with the reader as chief screening officer at a gene synthesis company, to show why screening customers is harder than it sounds, and to make three policy recommendations.

The policy moment

Custom DNA is ordinary infrastructure for modern biology. Dozens of gene synthesis companies worldwide manufacture custom strands for academic and industry customers, and that supply has driven genetic engineering, new diagnostics and therapies, and research on what genes do. Some sequences that can be ordered code for pathogen or toxin genes that could cause harm if misused, which is why experts across industry, government and academia recommend screening both sequences and customers before an order is filled.

American policy is at a crossroads. A Biden administration executive order required federally funded researchers to purchase only from companies that screen DNA orders. The Trump administration rescinded the requirement less than three months after it took effect. Many leading synthesis companies

screen anyway, as members of an industry-led consortium, but the revoked order would have been the first time screening was standardised and mandatory. The authors note that the new administration may keep some parts of the Biden order, so the status of requirements is unclear, and they add that the issue will not go away, because new science keeps demanding more synthetic genes.

Their central diagnosis: the main shortcoming in policy is that there is no clear guidance and no decision criteria for customer screening. Sequence screening by itself cannot carry the load, because sequences that look risky, such as pathogen genes contributing to virulence, are genuinely needed for disease research and vaccine design. And classifying a customer as legitimate or illegitimate is subjective and frequently ambiguous.

Malice aforethought: six customers

The device is sharp because of how it is built. Each pair of customers is matched, with identical qualifications and an identical order, and in each pair one is harmless and one is malicious. Nobody declares malice on an order form. So intent is invisible to the provider, while qualifications are visible, and that asymmetry is the article's whole point.

Pair one: Emma and Bob. Both ordered the sequence for botulinum toxin. Neither has a research background, biosafety training, or access to high-containment facilities. Emma is a high school teacher mentoring students interested in biology, who picked botulinum toxin as an interesting teaching example because it is a toxin that is also used cosmetically; she has no malicious intent and does not know the agent is controlled. Bob supplied no documentation of research need or capacity to work safely and has never worked in a laboratory; he does have malicious intent, having read online that he could order dangerous pathogen sequences from a provider. The verdict is easy: deny both. The authors' observation is the important part, that intent was irrelevant to the decision. Neither has experience with dangerous agents or access to the containment required to work with them safely, and that alone settles it.

Pair two: John and Nicole. Both are doctoral researchers at university laboratories, both ordering DNA coding for the hemagglutinin protein of an avian influenza virus. Both have extensive laboratory experience but have never worked with high-risk pathogens. Both have access to appropriate containment at their facilities, and both state that the sequence will not be used to generate infectious particles. John has studied how viruses infect cells using relatively low-risk circulating human influenza strains, has discovered a new inhibitor, and wants to

test whether it binds and disrupts the protein in H5N1, which could be a route to neutralising avian influenza. Nicole has also worked extensively with those genes from low-risk human influenza, and describes a study similar to John's. John means what he says and will most likely work safely. Nicole intends to use the DNA to cause harm. The verdict is that this is tricky: their qualifications put the decision in a gray area, and today it falls to the provider to weigh how much the customer needs the sequence, their experience, and what they disclosed about infrastructure and biosafety, with no standard universal guidance for how to do that weighing.

Pair three: Ralph and Alex. Both are highly respected researchers co-leading a national laboratory project to develop and optimise a Lassa fever vaccine for eventual inclusion in the US Strategic National Stockpile, work that routinely involves ordering modified and unmodified Lassa virus genes. They have extensive project documentation, are qualified for their institution's well-maintained high-containment facilities, have sign-off from the laboratory's biosafety officer, and are overseen both individually and as a laboratory by the **Federal Select Agent Program**, which regulates the possession, use and transfer of pathogens and toxins that threaten people, plants and animals. They have a proven compliance history, and in both cases the order ships to the national laboratory facility

rather than to the individual. Ralph is the model customer whose work furthers US security. Alex has malicious intent and plans to remove research material from the high-containment facility. The verdict is to ship to both, because a commercial provider has no way of knowing about Alex. Preventing theft or misuse at facilities that genuinely need sequences of concern belongs to the institution and the Federal Select Agent Program, not to the vendor. The authors add that given the background checks such researchers undergo, there are unlikely to be many Alexes.

Where do we go from here

Three recommendations, aimed at setting realistic expectations for what synthesis screening can achieve.

Build standards around the clearly illegitimate. However hard legitimacy is to define in general, one subset of customers, the Emmas and Bobs, is easy to identify as not legitimate, because they lack experience, infrastructure access and an appropriate research question. Those three factors are objective rather than subjective, so they could found a set of standards defining when it is unacceptable to supply a synthesised sequence of concern. And once the standards are clear, compliance evaluations could actually test whether providers deny that subset of orders as expected.

Give providers guidance for the gray areas. Without it, providers will reach different conclusions about customers like John and Nicole, and the judgement about whether a legitimate researcher is good enough will only get harder as more sequences are classified as concerning. The authors call for an answer key against which providers' decisions can be compared, especially for assessing compliance, built from more explicit criteria: how critical the sequence is to the requester's research, how experienced they are with similar sequences, and how much they disclosed about their facility's biosafety. They name a second gray area that gets less attention, referral. Should Emma and Bob be reported to the FBI? Should John and Nicole? Without knowing intent, which orders are suspicious enough to warrant investigation? Providers need guidance so those choices are systematic and consistent from company to company.

Remember screening is not the only defence. Some customers genuinely need sequences of concern; if there were no legitimate uses, legitimacy screening would not be necessary in the first place. It is unreasonable to expect a provider to fill Ralph's order and refuse Alex's, or to grade providers on catching Alex. The better move is to strengthen the safeguards that sit before and after the provider's involvement: the anti-theft and anti-misuse processes inside high-containment laboratories under the

Federal Select Agent Program, verified or strengthened institutional biosafety practice, and FBI and intelligence monitoring of actors sophisticated enough to beat provider screening. Building that capacity alongside synthesis screening is what reduces biological risk furthest.

The closing argument

Synthesis screening works best at preventing a baseline of clearly illegitimate customers from receiving potentially dangerous DNA, and is one piece of a larger risk-reduction puzzle. For the next iteration of required screening to work, providers need standards, guidance, support from other biosecurity tools, and clear compliance metrics. The authors argue that the administration will have to produce a plan, given how important the synthetic gene industry remains to biomedical research and biotechnology, and that a good starting point would be to keep the Biden-era screening requirements and build on them.

Words to know

- **Gene synthesis**: making a custom strand of DNA to order from a written sequence. It is the step that turns a sequence on a screen into physical material, which is why it is treated as a chokepoint.
- **Sequence screening**: checking the ordered DNA itself against lists of concerning genes. It is the

easier half of the job, and the authors argue it is not sufficient on its own.

- **Customer screening**: checking who is ordering, and whether their need, experience and facilities make sense. This is the half with no agreed criteria, and the subject of the article.
- **Federal Select Agent Program**: the US programme regulating possession, use and transfer of pathogens and toxins that threaten people, plants and animals. In the authors' argument it is the layer that should catch an insider like Alex, since a vendor cannot.
- **High-containment facility**: a laboratory built and certified to work safely with dangerous agents. Lack of access to one is one of the objective grounds for refusing an order.
- **Strategic National Stockpile**: the US reserve of medicines and countermeasures held for emergencies. It appears here as the destination for the Lassa vaccine work, which is what makes that order legitimate.

Numbers in this reading

- 6 fictional customers, arranged in 3 matched pairs, each pair identical on paper with one harmless and one malicious member.
- Less than 3 months: how long the Biden administration's screening requirement was in

effect before the Trump administration rescinded it.

- Dozens: the number of gene synthesis companies worldwide manufacturing custom DNA.
- 3 recommendations, and 3 objective factors behind the first of them: experience, infrastructure access, and an appropriate research question.

What the author is unsure about

The authors are clear about the limits of their own proposal. They say legitimacy is genuinely subjective and often ambiguous, and they do not claim their answer key would resolve the hard cases, only that providers would at least decide them consistently. They state that the status of US requirements is unclear, since parts of the rescinded order may be retained. They accept that nothing a provider does can catch an insider like Alex, and that it would be unreasonable to test providers on that. They acknowledge that the judgement about gray-area researchers will get harder as more sequences are classified as concerning. On referral to the FBI they pose the question rather than answering it. And their estimate that there are unlikely to be many Alexes is an expectation based on existing background checks, not a measurement.

For the discussion

- The sharpest point from this reading is that in the easy cases intent was irrelevant: the decision turned on experience, infrastructure and research question, all of which are visible, which is why the authors want standards built on the clearly illegitimate rather than on guessing who means harm.
- A tension worth raising: the article argues screening should not be judged on catching the insider, which conveniently moves responsibility to institutions and the FBI, and the group could ask whether those layers are actually funded and staffed to carry it.

2.7-1 · CORE · COURSE TIME 30 MIN

How to Secure the DNA Supply Chain

Janika Schmitt, Joshua Monrad · 2026

Read the original paper → · the course assigns: whole piece, including the appendix and footnotes

> **Synthetic DNA is the one place in the biological threat chain where almost every pathway to a pandemic-capable pathogen has to pass through a small number of companies, and the authors argue that getting roughly 80% of them to screen orders and customers, with nobody able to tell which ones screen, would buy the world years it badly needs.**

Average summary

This is a field strategy, published as part of IFP's Launch Sequence, written by two people at Sentinel Bio. It is a plan with targets, named projects and a recruiting pitch, not a neutral survey.

The motivating fact is a price collapse. When microbiologists synthesised the poliovirus genome from scratch in **2002** it cost **$300,000**; at today's commercial rates the equivalent genetic material costs less than **$1,000**. The authors also describe having a short fragment of the Ebola virus genome shipped to their coworking space last year with zero questions asked. The same capability cures cancers and builds synthetic viruses for the CDC to test treatments against, which is why their aim is to keep access easy for legitimate users.

Their central claim is that synthetic DNA is the highest-leverage chokepoint available. With research partners they assessed more than **200** alternative intervention points across the threat chain, from pipettes to biosafety cabinets, and nothing else came close, for three reasons: nearly every engineering pathway needs synthetic DNA, so it is hard to route around; a DNA order reveals both who is ordering and what they are ordering, which is far more signal about intent than a pipette gives; and gene-length synthesis is technically hard to commodify, so the market is concentrated in roughly **200** providers, with the

United States, China and the EU holding about **70%** of gene-length double-stranded DNA providers between them.

The counterintuitive part is that partial coverage works. They argue the weakest-link intuition, that one non-screening provider undoes everything, misreads how access controls function, because denial is only one of three mechanisms alongside deterrence and monitoring. Their analogy is LoJack, the hidden car tracker: auto theft fell sharply in cities that adopted it even though fewer than **2%** of cars carried one, because a thief could not tell which car was protected. Modelling with Rocco Casagrande found that patchy screening buys almost nothing, that effectiveness climbs rapidly once coverage passes **50 to 60%**, and that **70 to 80%** meaningfully reduces risk.

Hence their victory conditions for 2030: more than 80% coverage, preserved **strategic ambiguity** about who screens, low-cost and low-friction access for legitimate customers, and 90% of providers correctly identifying sequences of concern in realistic red-teaming. Voluntary measures over two decades have not got there, so they want mandates in a handful of jurisdictions, plus the infrastructure underneath: affordable customer checks, a real sequence standard, independent auditors, and security for benchtop synthesisers and short oligonucleotides.

Why it's assigned

This is the spine of her chosen deep dive. Course page 2.7 states the Deny premise, that a genome sequence is only a threat if someone can print it, so blocking the chokepoints prevents that, and this piece is the full argument for why DNA synthesis is the chokepoint and what securing it would actually require. It answers the page's guiding questions on nucleic acid synthesis and know-your-customer rules directly, including why those checks are rare in biotechnology: screening is a positive externality, costs the provider money, and brings them almost no private benefit, so voluntary adoption stalls. It also hands her the two unsolved problems at the edge of the field, benchtop devices and short oligonucleotides.

In depth

Buying time, and what the time is for

Until recently the only DNA supply chain was nature: to get a pathogen's genome you needed the pathogen. Cheap, reliable synthesis ended that, and the benefits the authors name are specific, since commercial synthesis has let patients drive their own cancers into remission and let the CDC build a synthetic Ebola virus to test field treatments faster.

The cost curve is the hinge. The poliovirus genome was synthesised from scratch in 2002 at a cost of

$300,000; at today's commercial rates the equivalent genetic material costs less than $1,000.

Meanwhile AI is approaching the point where a much wider set of people can execute sophisticated biological workflows, creating what they call a **risk window**: attack capability spreads before the defences that would eliminate pandemic risk exist. Their evidence on current uplift sits in a footnote, from a study they co-funded: the largest randomised controlled trial to date on AI-assisted biological performance, run by Active Site, put 153 biology novices through eight weeks of wet-lab experiments. Mid-2025 models gave no substantial uplift across the full set of tasks needed to synthesise a virus from scratch, but participants with randomised model access beat the control group on 16 out of 17 individual tasks, and the authors would bet this is the last null finding of its kind.

Why not simply build defences? Countermeasures take years: the COVID-19 vaccine rollout, the fastest in global history, still left low-income countries waiting more than two years for adequate coverage, and deploying anything to billions runs into bottlenecks like jurisdictional approvals. Their blunt conclusion is that an attack soon would leave the world with the COVID playbook, retreat and lockdown and wait.

So the strategy is to buy time by selectively raising barriers. DNA orders are attractive because each

carries two pieces of information, what is ordered and who ordered it, which lets providers stop dangerous orders before they ship and report suspicious actors. And yet, by the authors' count, fewer than 50 people worldwide work on DNA synthesis security full-time. They call it an orphan issue.

Why synthetic DNA is the chokepoint

A virus, as they put it, is essentially a stretch of genetic code wrapped in a protein shell, and viral genomes are short: SARS-CoV-2, which killed millions and cost the global economy trillions, is only 30,000 bases long, roughly 100,000 times shorter than the human genome's 3 billion base pairs. Building one needs the digital sequence, usually freely available online, plus synthetic DNA to reconstruct it, so no natural sample is required. Novel design is possible too: researchers used the genome language model **Evo 2** to design viable bacteriophages, some so genetically different from anything known that they would qualify as a new species.

The chokepoint claim rests on comparison. With research partners, building on a RAND report, the authors spent months assessing more than 200 alternative intervention points across the threat chain, from pipettes to biosafety cabinets. Three reasons nothing beat synthetic DNA. It is hard to route around, since other inputs cover only narrow classes of threat: restrict humanised ACE2 mice, used for

coronaviruses, and an actor switches to a pathogen that does not need them. It carries signal about intent, which common equipment does not, since sequence screening reveals whether someone is ordering fragments of smallpox or the gene for a fluorescent protein. And it resists commodification, because frontier uses such as gene therapy need error rates low enough to demand sophisticated methods and quality control, which concentrates the market: roughly 200 gene synthesis providers globally, with the United States holding 25% of gene-length providers, China 33% and the EU 11%. A footnote contrasts this with ordinary growth media that biohackers improvise at home, where restricting suppliers would achieve nothing.

The victory conditions

Four, set for 2030: coverage, with more than 80% of companies providing synthetic DNA or benchtop devices screening both orders and customers; strategic ambiguity, so nobody can tell which providers screen; access that stays low-cost and low-friction for legitimate customers; and effectiveness, with 90% of providers correctly identifying sequences of concern in realistic red-teaming. Two things get them there: mandates in the handful of jurisdictions covering the supermajority of providers, and the infrastructure underneath, including affordable know-

your-customer systems, standards, third-party red-teaming and benchtop security.

Why partial coverage is enough

The weakest-link intuition says one non-screening provider undoes the rest. The authors say that misreads access control, because most analyses of technology governance judge only denial, whether the system blocked the actor at the point of transaction, and on that measure customer screening often looks ineffective. But denial is one of three mechanisms, alongside **deterrence** and **monitoring**, which do at least as much work.

The illustration is motor vehicle theft. In the 1980s and 1990s American cities adopted LoJack, a hidden radio transmitter that let police track stolen cars, invisible from outside so thieves could not tell which cars had one. Auto theft fell sharply in LoJack cities even though fewer than 2% of cars were equipped, and a thief who took an equipped car led police straight to them, often dismantling whole operations. An attacker who cannot tell which providers screen must either order and risk an investigation, or not try.

That has a design consequence the authors flag as a trap: publishing a list of providers that screen, as a commercial incentive, would tell a bad actor exactly whom to route around. Mandates push the other way, because an actor must then assume any provider in

that jurisdiction might screen, and since some elsewhere screen voluntarily, none can be safely assumed not to.

To put numbers on it they worked with Rocco Casagrande, a biotechnologist who has spent over 25 years working to prevent biological misuse. His Markov-chain model follows an intelligent bad actor who may be deterred or interdicted before reaching a provider, succeeds by finding or randomly picking a non-screening provider or one whose screening fails, may be reported if caught, and tries again if neither disrupted nor discouraged. The finding is that all three components must hold, because failure on any one hands over the material.

The figure

The risk reduction curve plots the probability that a bad actor obtains the material against screening coverage, with effectiveness, deterrence and law enforcement's catch rate varied alongside it. The shape is the argument: nearly flat at low coverage, because moving from 10% to 20% barely changes the odds of routing around the screeners, then bending sharply upward once coverage crosses 50 to 60%, with 80% marked as the target. It covers only the mail-order gene-length pathway, though the same model applies to oligonucleotides or live pathogen repositories. What to notice is the knee rather than a straight slope, which is why incremental voluntary

gains are close to worthless while a mandate that pushes a jurisdiction over the knee is worth a great deal.

Why providers do not screen

For more than a decade a coalition organised around the **International Gene Synthesis Consortium**, from startups to multinationals across the US, Europe, the UK and China, has voluntarily screened orders for sequences associated with dangerous pathogens and toxins. That proves screening is feasible at industrial scale, and supplies strategic ambiguity outside mandated jurisdictions.

But it has not reached critical mass. A 2025 study suggested many providers shipped potentially dangerous sequences without adequate screening; its methodology was hotly debated, but multiple non-public studies developed with industry and government engagement validated the broader conclusion that industry screening is inadequate to deter, deny or monitor determined actors. Those studies stay unpublished to avoid disclosing vulnerabilities and to preserve strategic ambiguity.

Their diagnosis of the incentives is the sharpest passage in the piece. Providers perceive the risk as limited, an inference the authors call historically valid, since there are no known cases of synthetic DNA being used to build a bioweapon. More fundamentally,

the benefit of screening is a positive externality, accruing to the world rather than the provider, especially since misused DNA is hard to trace to its manufacturer without recordkeeping, while the costs are not zero because customer checks are still largely manual. A provider facing non-zero cost and near-zero private benefit will not act unless required to, and voluntary regimes penalise responsible providers who carry costs their competitors avoid, which is why they keep asking governments to level the playing field.

Governments are moving

In the United States, the bipartisan Biosecurity Modernization and Innovation Act and the administration's AI Action Plan have moved synthesis screening into the legislative and executive mainstream. The EU's Biotech Act is largely deregulatory but carved out synthesis screening as a core security provision. The UK's 2025 Biological Security Strategy named mandatory screening a priority, and New Zealand has introduced legislation. In what the authors call a rare moment of consensus, leading technology executives and life sciences researchers, from the 2024 Nobel Laureate in chemistry David Baker to OpenAI chief executive Sam Altman, called on Congress in a public letter to make screening mandatory. A footnote records the American whiplash: the Biden administration's 2024

Framework for Nucleic Acid Synthesis Screening made federally funded researchers buy only from companies that screen, and a Trump executive order of May 2025 told the Director of the Office of Science and Technology Policy to revise or replace it, with no new framework released yet.

The economics favour action everywhere. Cost-benefit analyses by the British Centre for Long-Term Resilience and RAND Europe find societal benefits vastly outweighing the costs, with the UK gaining £3.50 in avoided economic damage and avoided loss of life and health for every £1 spent, a net benefit of £150 million per year over 20 years. The same argument holds in Shenzhen and Seoul as in Brussels or Washington, and Asia, particularly China, where the plurality of manufacturers operate, is high priority, needing sustained engagement across governments, industry and scientific institutions with bodies such as ISO ensuring interoperability. National policy also reaches past borders through import controls and procurement, as with the German provider that adopted a premium screening tool so its customers could qualify for federal US research grants.

The projects they want built

Affordable screening. Sequence screening is the easy half: free tools already screen thousands of base pairs per second, so the authors state flatly that no provider in 2026 can credibly call it too expensive or

infeasible. The cost sits on the customer side, where verifying legitimacy means reviewing publication histories, institutional affiliations and sanctions lists. A recent study found AI-assisted customer screening cut costs more than tenfold, to $1.18 per customer against $14.04 with a human reviewer, while matching or exceeding human accuracy, and roughly $0.20 for information-gathering tasks AI handles alone. They size that against aviation screening, which US passengers fund at $5.60 per one-way trip, around 2% of the average domestic fare, making customer checks a small share of a typical $70 to $210 gene-length order. Savings multiply if the check happens once, since centralised credentialing would let a researcher verify once across synthesis providers and AI platforms, while surfacing patterns across vendors that individual checks miss.

Standards. Providers still face the question of what counts as dangerous. With no standard at the level of actual sequences, they screen against different databases and must choose between flagging too much and losing legitimate customers, or too little and risking penalties under future mandates, while governments cannot enforce a standard that does not exist. The scope is tractable: only 1,400 known viruses, bacteria and fungi cause human disease, and only a small fraction threaten public health, agriculture or national security. Industry has asked governments for nearly two decades to name the

regulated sequences, with no sufficiently detailed standard yet issued. In 2024 the **Sequence Biosecurity Risk Consortium**, catalysed philanthropically and now partly sustained by industry dues, stepped in, categorising over 1.1 million sequences into a risk-based standard ready for regulators, which needs regular updates as science and AI advance. On the customer side, the Engineering Biology Research Consortium and IBBIS are developing a minimum viable customer standard for government adoption.

Extending the perimeter. Screening covers gene-length sequences from mail-order providers, and the two hardest problems sit outside it. Benchtop synthesisers print DNA on demand on a laboratory bench, which democratises biotechnology and enlarges the risk surface; the installed base is small but the devices are on track to make longer sequences and spread further, and cryptographic methods can screen even air-gapped machines. Working with manufacturers now on access controls, usage logging and integrated screening is orders of magnitude easier than retrofitting thousands of machines later, and the authors say the world needs a general manager for this, because not a single person works on it full-time. The second problem is short **oligonucleotides**: the market is more decentralised, with up to five times as many manufacturers as for gene-length DNA, and short fragments carry less

signal, so it is harder to judge one in isolation or to spot orders split across providers. SecureDNA screens down to 30-mers with few false positives, but catching split orders is a coordination problem as much as a technical one, and an industry-wide clearinghouse has only begun. Their research also surfaced live pathogens as a comparably important chokepoint, since an actor holding one can modify what they have rather than build a genome, with contract research organizations and cloud laboratories close behind.

Auditing, and AI evasion. The way to know whether screening works is to test it. The 2024 Screening Framework shows why: it drove adoption in the US and abroad but only required providers to self-attest, and independent red-teaming afterwards showed some who claim to screen do not do it well enough. They want dedicated organisations stress-testing at scale, government-backed eventually, with civil society pioneering the methods. On evasion, a study in Science found AI protein design tools circa 2023 could already reformulate sequences of concern to evade screening; the tools were patched, and a few similarly reformulated benign proteins did not reliably retain function, so they call this not the most pressing problem today while refusing to bet against AI eroding that margin. The same advances cut both ways, and Fourth Eon Bio uses AI for adaptive function-based screening, moving the question from

whether a sequence matches a known pathogen to whether it could have a dangerous function.

They insist on sequencing the work: people gravitate to the evasion problem because it is interesting, but where few providers screen a bad actor needs no AI, only a provider that does not screen. None of these problems is close to solved, and the number of people working on them needs to increase fivefold.

How it closes

Their final position is pro-abundance: personalised gene therapy cures rare diseases once deemed untreatable, new diagnostics and vaccines can be designed in days at the start of an outbreak, and they want vastly more of that. The principle they rest on is that consequential technologies carry safeguards proportionate to their power, with the line that nobody seriously argues private citizens should be able to order plutonium to their doorstep. Done well, screening is fast and cheap enough that legitimate researchers barely notice it while the handful of catastrophic orders are caught. The piece ends as a recruitment call to Sentinel Bio and IFP's Launch Sequence request for proposals, with disclosures stated: the authors work at Sentinel Bio, which funded Fourth Eon Bio and supports credentialing work, and took no compensation from IFP.

Words to know

- **Chokepoint**: a step in a process that is hard to avoid, so controlling it controls everything downstream. The whole strategy depends on synthetic DNA being one.
- **Gene-length synthesis**: making DNA long enough to constitute a gene, as opposed to short fragments. It is technically demanding, which is why the market is concentrated and screenable.
- **Oligonucleotide (oligo)**: a short DNA fragment, which can be assembled into longer sequences. The decentralised oligo market is one of the two unsolved problems here.
- **Benchtop synthesiser**: a machine small enough to sit on a laboratory bench that prints DNA on demand. It moves synthesis out from under mail-order screening entirely.
- **Know-your-customer (KYC)**: verifying who a customer is and whether their stated use is legitimate. The authors argue this, not sequence screening, is where the real cost sits.
- **Strategic ambiguity**: deliberately keeping it unclear which providers screen, so an attacker cannot route around them. It is why publishing a list of good actors would backfire.
- **Red-teaming**: testing a defence by trying to beat it, here by sending realistic orders to see whether providers catch them. It is how the gap between claimed and actual screening was found.

- **Positive externality**: a benefit that goes to everyone except the party paying for it. It is the economic reason voluntary screening stalls.
- **Defence in depth**: stacking independent safeguards so no single failure is decisive. The authors invoke it because even 90% coverage leaves residual risk.

Numbers in this reading

- $300,000 in 2002 to synthesise the poliovirus genome from scratch, against less than $1,000 for the equivalent genetic material at today's commercial rates.
- 30,000 bases: the length of the SARS-CoV-2 sequence, roughly 100,000 times shorter than the human genome's 3 billion base pairs.
- More than 200 alternative intervention points assessed across the biological threat chain.
- Roughly 200 gene synthesis providers globally; 25% of gene-length providers in the United States, 33% in China, 11% in the EU, with those three accounting for about 70% of gene-length double-stranded DNA providers.
- Fewer than 50 people worldwide working full-time on DNA synthesis security, and a call to increase the workforce fivefold.
- Victory conditions by 2030: more than 80% coverage, and 90% of providers correctly identifying sequences of concern in red-teaming.

- Fewer than 2% of cars carried LoJack in cities where auto theft fell sharply.
- Coverage thresholds from the model: 10% to 20% changes little, effectiveness improves rapidly above 50 to 60%, and 70 to 80% meaningfully reduces risk; 75% already much better than 30%, and 85% better still.
- £3.50 of pandemic-prevention benefit per £1 spent in the UK, a net benefit of £150 million per year over 20 years.
- $1.18 per customer for AI-assisted screening against $14.04 with a human reviewer, and roughly $0.20 for tasks AI handles alone, set against a typical $70 to $210 gene-length order and the $5.60 per one-way trip US passengers pay for aviation screening, about 2% of the average domestic fare.
- 1,400 known viruses, bacteria and fungi that cause human disease, of which only a small fraction threaten public health, agriculture or national security.
- Over 1.1 million sequences categorised by the Sequence Biosecurity Risk Consortium since 2024.
- Up to 5 times as many manufacturers of short oligos as of gene-length DNA; SecureDNA screens down to 30-mers with few false positives.
- 153 biology novices over eight weeks in the Active Site randomised controlled trial, with model-assisted participants beating the control group on

16 out of 17 individual tasks despite no substantial uplift on the full set.

- More than two years: how long low-income countries waited for adequate COVID-19 vaccine coverage in the fastest rollout in global history.

What the author is unsure about

They separate their findings from their bets fairly carefully. The 80% target is explicitly a target rather than a threshold, with no magic at the exact number. They concede that even at 80% or 90% coverage a small risk remains, which is why they argue for defence in depth. They acknowledge that providers' perception of low risk is historically valid, since no known case exists of synthetic DNA being used to build a bioweapon. They report that the methodology of the 2025 red-teaming study was hotly debated, and that the studies they rely on instead are unpublished, which means a reader cannot check them. They say hardening against AI-enabled evasion is not the most pressing problem now, while refusing to bet against AI eroding that margin, and they predict the Active Site null result will probably be the last of its kind, labelled as a bet. Their model is a simulation with assigned transition probabilities, offered on request rather than published in the piece. And they are open about interest: the authors work at Sentinel Bio, which has funded one of the organisations they praise and is supporting the credentialing work they recommend.

For the discussion

- The LoJack argument is the thing worth bringing to the group, because it inverts the usual intuition: you do not need every provider to screen, you need an attacker to be unable to tell which ones do, which means publishing a list of responsible companies would actively make things worse.
- The reason this has not happened already is economic rather than technical, since sequence screening is cheap and available while the benefit of screening goes to the public and the cost falls on the provider, so the group could ask what else in biosecurity has that same shape.

2.7-2 · CORE · COURSE TIME 35 MIN

Know your scientist: KYC as biosecurity infrastructure

Jonathan Feldman, Tal Feldman, Annie I. Anton · 2026

Read the original paper → · the course assigns: whole piece

> **Since nobody can reliably tell whether a designed protein is dangerous, the authors argue we should stop trying to judge the output and start verifying the user, borrowing the banking sector's know-your-customer machinery and letting research institutions vouch for their own people.**

Average summary

The authors start from a technical dead end. Biological AI tools for protein design and structure prediction are advancing fast, and the standard safeguards, keyword filtering, output screening and content-based access denials, are badly suited to biology. Keyword filters can be circumvented. Sequence homology catches known pathogens but misses novel designs by construction. Functional annotation is incomplete, and reliably predicting whether something will be pathogenic or toxic is beyond current capability. That leaves a bad choice: restrict access so heavily that legitimate research is blocked, or allow access so broadly that misuse is hard to detect or attribute.

Their alternative is to change the question. Instead of asking only whether a particular output is dangerous, which current systems cannot answer, also ask whether the user generating it is one we trust. The model they borrow is **anti-money laundering** regulation in finance, where layered identity checks at entry are combined with monitoring of behaviour over time, and where **know-your-customer** rules require institutions to verify identity and refuse service to sanctioned entities before granting access.

The proposal is three tiers. **Tier I** is institutional gatekeeping: a researcher applies through their home institution, not the model provider; the institution

confirms identity and role and vouches for the proposed use; the host keeps a list of trusted institutions whose endorsements it accepts. Exclusions are objective, drawn from government lists such as OFAC's Specially Designated Nationals, the State Department's Foreign Terrorist Organizations list and the Commerce Department's Entity List, plus people convicted under biological weapons statutes. **Tier II** screens outputs in real time with tools such as BLAST homology searches and functional annotation, freezing and logging flagged results for human review rather than automatic refusal. **Tier III** watches patterns over time for activity inconsistent with a declared research purpose.

The architecture rests on shared responsibility, because institutions can judge researchers but cannot see what they do inside the model, while hosts can see every interaction but cannot judge a researcher's standing. Enforcement runs on access revocation plus shared exclusion lists across platforms, so a revoked user cannot simply move to a competitor. Everything can start voluntarily, through terms of service, and the authors point to the Federal Select Agent Program as a precedent where the federal government sets exclusion rules and institutions control access day to day.

Their stated limits are real: open-source models bypass the whole scheme, well-resourced state actors

can build front institutions, thresholds need empirical tuning, and no access control solves the insider problem.

Why it's assigned

This is the conceptual heart of her chosen deep dive. Page 2.7 asks why know-your-customer protocols are common in financial services but not in biotechnology, and this paper is the most complete answer in the pack: it takes the finance machinery apart and rebuilds it for biology, tier by tier, with the enforcement and the funding levers attached. It also connects the Deny sub-layer's two halves, because it argues that screening DNA orders without governing the design tools upstream leaves a gap, since a designed sequence can be carried elsewhere or synthesised abroad.

In depth

The problem with judging the output

Systems for structure prediction, sequence modelling and generative protein design now support work that was recently confined to specialised laboratories, and the authors accept the benefits for medicine and biotechnology as real. Their argument is about the safeguards, not the tools.

Most current approaches are model-level restrictions: keyword filtering, output screening, or denying access

based on content. The authors say each fails in a specific way. Keyword filtering can be circumvented. Sequence homology searches detect known pathogens but miss novel designs by construction, because a novel design is precisely one that does not resemble anything on the list. Functional annotation remains incomplete. And reliable prediction of pathogenicity or toxicity is beyond current capabilities. Taken together, these limits produce the tradeoff the paper exists to escape: restrict so heavily that legitimate research is impeded, or permit so broadly that misuse becomes difficult to detect or attribute.

The pivot is one sentence: rather than asking solely whether a particular output is dangerous, a question current systems cannot reliably answer, one can ask whether the user generating that output is one we trust. The authors are clear this is a supplement to content inspection, not a replacement for it.

Why finance is the model

Anti-money laundering frameworks exist to stop illicit flows of money for terrorism, organised crime and sanctions evasion, and they work through layered identity verification of users plus behavioural monitoring of their actions. At the root sit know-your-customer requirements: institutions must verify customer identity and refuse service to sanctioned entities before granting access, and transaction

monitoring then detects suspicious patterns over time. The authors draw the parallel plainly, that the risk anti-money laundering prevents, enabling financial crime, is instructive for the risk here, enabling pathogen design. They also state what is new in their contribution: know-your-customer approaches have long been discussed in AI governance and are already used in parts of the biotechnology supply chain, particularly nucleic acid synthesis, but have never been formalised into an operational framework centred on biological AI. On timing, they note protein design capability has advanced rapidly over the past three years, naming AlphaFold, ESM3 and RFdiffusion, and that many tools are now reached through APIs, which creates natural chokepoints for access control.

The figure

Figure 1 is the framework diagram, showing the three tiers as a stack that a user passes through and then stays inside: institutional vouching at the point of entry, real-time screening of each output, and longitudinal monitoring of behaviour across sessions, with human review rather than automatic refusal at the decision points. The saved text of the article carries the reference to the figure but not its caption or image, so this description comes from how the body text lays the tiers out rather than from the drawing itself.

Tier I: the institution as trust anchor

A researcher seeking to use a biological AI tool applies through their home institution rather than directly to the provider. The institution evaluates the request, confirms the researcher's identity and role, and vouches for the legitimacy of the proposed use, while the hosting entity keeps a list of trusted institutions whose endorsements it accepts.

By research institutions the authors mean organisations conducting biological research under recognised oversight: universities and colleges, government research laboratories, non-profit research institutes, hospital-affiliated research centres, and corporate research and development divisions. The common thread is institutional accountability, since many maintain biosafety committees or ethics review boards and all employ researchers under formal relationships. The institution, in their phrase, functions as the compliance officer, responsible for confirming the researcher is who they claim, that the use is legitimate, and that they are not on government exclusion lists.

Following finance, they prefer objective exclusions to subjective trustworthiness judgements. Hosts should deny access to anyone on OFAC's Specially Designated Nationals list, the State Department's Foreign Terrorist Organizations list or the Commerce Department's Entity List, plus individuals convicted

under biological weapons statutes. Beyond those bright lines, institutions keep discretion for context-specific vetting, which the authors compare to the enhanced due diligence banks apply to higher-risk customers.

Two practical gaps get handled. Startups and small firms without equivalent administrative infrastructure can partner with an established institution that assumes sponsorship, or a third-party vetting institution could emerge to vouch for them for a fee, taking on the same accountability. Visiting scholars, collaborators and independent researchers can be sponsored the same way, with the sponsor vetting both person and purpose and accepting ongoing oversight; the authors note that an institution's willingness or refusal to sponsor someone is itself a risk signal.

Trust criteria are left to the host and scale with the model. A host might weigh the presence of an **Institutional Biosafety Committee**, federally funded research subject to established oversight, prior compliance history, or an existing relationship, read against the model's risk profile rather than in isolation. For lower-risk systems, such as small protein language models comparable to the ESM family, the threshold may reasonably be lower; for models capable of designing a genome from scratch, requirements may be substantially stricter, effectively

limiting access to institutions with demonstrated oversight capacity and clearly legitimate uses.

Tier II: screening outputs, without automating the verdict

Generated sequences or designs are analysed in real time with whatever tools exist, currently sequence homology searches such as BLAST and ontology-based functional annotation, to find similarity to known pathogens or hazardous functions. The authors concede these are imperfect while arguing they give useful signals in context, and they make the tier modular so better function predictors and toxicity assessments can be slotted in without changing the access-control structure. Emerging watermarking and output-provenance techniques can also be integrated, which adds accountability particularly when a generated sequence is later synthesised.

When an output is flagged and the user's declared use case does not justify it, an inquiry can begin: the output may be temporarily frozen and held for review, and the user's activity is logged. The point is not automated enforcement but generating signals that prompt human oversight, and users knowing their activity is monitored and attributable is itself part of the mechanism.

Their worked example is the clearest statement of how the tiers interact. A researcher developing

influenza vaccine candidates may legitimately generate sequences resembling viral proteins; the system flags them, review confirms the declared purpose, and access continues. But if a user whose stated purpose is analysing mitochondrial genomes produces similar sequences, the mismatch triggers an escalated review. The flag is not the judgement; the mismatch is. This mirrors transaction monitoring in finance, which escalates for review rather than enforcing automatically, and the authors lean on an admission built into that regime: perfect automated detection is explicitly recognised as unattainable, so the objective is signals that enable human oversight, accountability and attribution.

Tier III: patterns over time

The third tier covers risks that emerge gradually rather than from any single output. User activity is analysed longitudinally for patterns that deviate from declared use cases or expected research behaviour: repeated interactions with flagged outputs, aggregate risk signals accumulating, unexpected access patterns, or inconsistencies between stated goals and observed usage. Rather than reacting to single events, the tier uses threshold-based accumulation of signals to trigger review.

Two honest notes sit in this section. Decisions to suspend or revoke access are not automated, because human review is required before enforcement, which

allows contextual judgement and reduces erroneous penalties. And judging whether outputs are inconsistent with a declared use case may require considerable domain expertise and familiarity with the researcher's work, which the authors say is not a trivial undertaking. Where possible the home institution and the host should assess legitimacy together, particularly after repeated flagging, since the institution can see whether the activity fits the researcher's broader programme.

Shared responsibility, stated as the foundation

Neither party can do this alone. Institutions vet applicants, evaluate proposed uses and keep researchers inside oversight structures, but their visibility into downstream model use is inherently limited, so institutional gatekeeping cannot be a complete safeguard. Hosts must decide which institutions to trust, calibrate requirements to the model's risk, and run the technical oversight, and because they see every interaction they are uniquely placed to spot anomalous usage an institution never would. The framework works, in their words, precisely because responsibility is distributed rather than centralised. One structural condition is stated twice: all of it depends on the tool running in a managed access environment, since open-source models can be run or modified locally, bypassing access control entirely.

Why anyone would adopt it

For researchers, the authors accept that preventing biosecurity risk requires some access friction, since misuse cannot be deterred without restrictions, and give three reasons the friction is tolerable. It is minimal, because researchers already navigate institutional approvals for laboratory work, grants and protocols, so a one-time access request is a marginal addition and most continue uninterrupted afterwards. It is targeted, because unlike general-purpose chatbots there is little reason for someone not doing cutting-edge biological research, with no institutional affiliation, to independently access tools capable of designing agents with pandemic potential, just as they should not work unsupervised with live biological agents. And the finance comparison holds, since banks verify every customer and monitor transactions worldwide while modern payment systems still operate at global scale, because risk-based controls are designed to be proportionate.

For hosting entities, the argument is that this is far more tenable than the alternative of building systems that reliably predict whether a designed protein is dangerous. Current biological AI cannot determine pathogenicity reliably, and solving that would need breakthroughs in structural biology, immunology and machine learning, which the authors call expensive, uncertain and potentially impossible. Know-your-customer sidesteps it by deferring vetting and

attribution to institutions: the institution vouches, the user operates under institutional oversight, and misuse traces back to both. They add that high-profile misuse could bring regulatory backlash and institutional distrust, so proactive safeguards act as reputational insurance.

Enforcement: revocation, and a shared list

The framework is designed for immediate voluntary adoption: companies offering API access to biological design tools, and they name AlphaFold Server, ESM APIs and RFdiffusion implementations, can implement all three tiers unilaterally through their terms of service while regulatory pathways develop.

The primary mechanism is access revocation, resting on the position that access to these tools is a privilege rather than a right. Users who misrepresent themselves, circumvent safety measures, or generate outputs inconsistent with their declared purposes lose access, and institutions that repeatedly vouch for problematic users, or fail to address misconduct by their own researchers, risk institution-wide revocation.

What makes revocation bite is information sharing between platforms. As financial institutions share information about suspicious actors through regulatory reporting and authorised mechanisms, biological AI hosts could maintain shared exclusion lists, so a revocation on one platform puts the person

or institution on a registry others can see. The authors are careful about what gets shared: predefined, objective categories such as documented terms-of-service violations or confirmed misrepresentation, because standardised criteria reduce reliance on subjective trustworthiness judgements and mitigate legal and governance concerns. The problem this solves is competitive, since a single platform may hesitate to revoke under commercial pressure, and if a problematic user can move elsewhere immediately, enforcement has no deterrent value. With shared lists, an institution that loses trusted status with the major platforms effectively loses access to the whole ecosystem.

Precedent, and the government levers

The authors argue delegated oversight is already how American biosecurity works. The **Federal Select Agent Program** is their exemplar: federal law sets which agents are regulated and who is barred, while entities must register, designate responsible officials and maintain internal controls, with individual access decisions made inside the institution and informed by federal security risk assessments conducted through the Department of Justice. Export control and research security regimes split the same way, with institutions screening partners and visitors against government lists and facing penalties for violations.

Three federal pathways could accelerate adoption. Funding agencies already impose biosafety and dual-use research requirements on recipients, so the NIH, NSF and DOE could condition awards on institutions implementing this kind of access control, which the authors say would need no new legislation and could be done through administrative rulemaking. NIST, through CAISI, has authority to develop voluntary consensus standards and could convene hosts, institutions and biosecurity experts to write technical ones, which would probably become de facto requirements through market pressure and federal procurement. And the Biden administration's 2024 Screening Framework Guidance for Providers of Synthetic Nucleic Acids already makes DNA synthesis companies screen orders against databases of concerning sequences and verify customer identity, which the authors read as federal recognition that these risks require verifying who has access rather than only what is being made. Their framework is the logical upstream extension: if synthesised DNA orders are screened, access to the tools that design those sequences should be governed too. They note the Trump administration's AI policy discussions have likewise treated synthesis screening as a national security priority.

Their argument against stopping at synthesis is specific. A design must eventually be synthesised to cause physical harm, but governing only that stage

leaves gaps, because researchers can accumulate dangerous knowledge, share designs through other channels, or synthesise abroad. Federal action could also attack the open-source hole by conditioning funding or using export controls to keep the most capable models under hosted access, which they concede would not prevent all diffusion but would slow proliferation of the highest-risk tools. The most ambitious pathway is direct federal oversight for high-risk models, which they call the most challenging option, needing dedicated funding, infrastructure and sustained administrative capacity: the government would vet individual researchers, through an existing agency or a dedicated subsidiary, with the Department of Energy and the Centers for Disease Control named as possible homes, while hosts kept monitoring usage and the government maintained a registry of high-risk models. Given resource limits, federal vetting could cover only the riskiest models, with institutions still the primary evaluators.

Limitations, in their own words

Four, and the authors soften none. Open-source models bypass access control entirely, which renders the framework ineffective without centralised hosting, so for it to apply, biological AI tools must not be open-sourced. Know-your-customer is not a complete defence against determined, well-resourced state actors that can establish front institutions; its strength

is deterring lone actors and raising the cost of misuse through accountability and traceability, though they add that the financial experience shows well-resourced actors hiding behind complex ownership structures can still be detected. Calibration is unresolved, since thresholds for flagging outputs and accumulating behavioural signals need empirical tuning to balance false positives against security sensitivity, and they deliberately leave review timelines, appeals and specific trust criteria to implementers, on the grounds that hosts face different threat models, resources and user populations. Finally, a malicious actor with genuine institutional affiliation could use the tools legitimately while probing boundaries or accumulating knowledge for harmful ends; Tier III gives some protection, but they state plainly that no access control system fully solves the insider threat problem.

The conclusion restates the choice: either restrict access so heavily that legitimate research suffers, or keep open access with minimal safeguards as capabilities grow more dangerous. Their framework is the middle path, implementable today and scalable later, and their closing line is that the question is not whether such infrastructure will eventually be necessary but whether it is implemented proactively or reactively. The paper declares no financial support, no commercial or financial conflicts of interest, and no use of generative AI in writing it.

Words to know

- **Know-your-customer (KYC)**: verifying who a user is, and refusing service to sanctioned parties, before granting access. Imported here from banking as the entry layer of the framework.
- **Anti-money laundering (AML)**: the financial regime that combines identity verification with ongoing transaction monitoring. It is the template, including its admission that perfect automated detection is impossible.
- **Biological AI tool, or biological design tool**: a model for predicting protein structure, modelling sequences or generating new protein designs. AlphaFold, ESM3 and RFdiffusion are the named examples.
- **Trust anchor**: the party whose vouching the system relies on. Here it is the research institution, chosen because it already vets people through hiring, training and biosafety review.
- **Sequence homology search**: comparing a sequence against known ones to find similarity, with BLAST as the standard tool. It is Tier II's workhorse and also its weakness, since novelty evades it by definition.
- **Institutional Biosafety Committee (IBC)**: the in-house body that reviews a research institution's biological work. Its existence is one signal a host can use to decide whether to trust an institution.

- **Managed access environment**: a tool reachable only through a host that controls who uses it, usually via an API. Without it, the entire framework collapses, which is why open-source release is treated as disqualifying.
- **Enhanced due diligence**: the extra scrutiny banks apply to higher-risk customers. It is the model for the discretion institutions keep after the objective exclusions are applied.
- **Output provenance and watermarking**: techniques for marking generated content so its origin can be traced later. They matter most at the moment a designed sequence is handed to a synthesis provider.

Numbers in this reading

- 3 tiers: institutional gatekeeping, output screening, and behavioural pattern monitoring.
- 3 years: the period over which the authors say protein design capabilities have advanced rapidly, naming AlphaFold, ESM3 and RFdiffusion.
- 3 government exclusion lists named as bright lines: OFAC's Specially Designated Nationals, the State Department's Foreign Terrorist Organizations list, and the Commerce Department's Entity List, plus conviction under biological weapons statutes.
- 3 federal funders named as potential levers: NIH, NSF and DOE, with NIST through CAISI as the standards body.

- 2024: the Screening Framework Guidance for Providers of Synthetic Nucleic Acids, which the authors treat as precedent for verifying who has access rather than only what is synthesised.
- 4 stated limitations: open-source models, state actors with front institutions, unresolved calibration, and the insider threat.

What the author is unsure about

The paper is unusually explicit about its own boundaries, and most of the uncertainty is structural rather than empirical. The framework does not work at all for open-source models, which the authors state twice. It is not a complete defence against state actors able to create front institutions, and they say its strength is deterring lone actors and raising costs, not eliminating threats. The thresholds for flagging outputs and for accumulating behavioural signals are unset and need empirical tuning, which they would rather resolve by iteration than by writing fixed rules. They deliberately leave review timelines, appeals and institutional trust criteria undefined, arguing that over-prescription would be counterproductive. They concede that judging whether a researcher's outputs match their declared purpose takes real expertise and is not trivial. And they say plainly that no access control system fully solves the insider problem. The regulatory pathways are proposals about what agencies could do, not descriptions of anything in

force, and the claim that voluntary NIST standards would become de facto requirements is a prediction.

For the discussion

- The reframing is the thing worth saying out loud: because function prediction is unreliable, the authors move the question from what was generated to who generated it, which is the same move synthesis screening already makes when sequence checks run out of signal.
- The enforcement insight is that revocation only deters if a revoked user cannot move to a competitor, so shared exclusion lists are what turn individual terms of service into an actual regime, and the group could ask who would be willing to run such a registry and on what evidence.
- She has argued for the monitoring half of this in her own name: she submitted a public comment to the U.S. Food and Drug Administration on generative AI, in her individual capacity, asking for monitoring after approval, version tracing and a way to roll back, which is close to what Tiers II and III ask of a model host after the gate has already opened.

2.7-3 · CORE · COURSE TIME 10 MIN

No, there are not hundreds of cloud labs in biology

Lennart Justen · 2026

Read the original paper → · the course assigns: whole piece

> **Cloud labs are a real technology and a small industry, three or four providers rather than dozens, and treating them as today's big biosecurity hole gets the facts wrong and spends credibility that biosecurity cannot spare.**

Average summary

This is a correction with a named target. On Sam Harris's Making Sense podcast, episode 463, Privatizing the Apocalypse, Rob Reid said there are dozens, maybe a hundred cloud labs and suggested they are a major emerging source of biorisk. Lennart Justen agrees with much of that podcast but says Reid is wrong on two counts: the number of cloud labs, and how much risk they currently pose.

A **cloud lab** is a highly automated, flexible biological laboratory that users operate remotely, without owning hardware. The appeal is obvious: a wet lab is expensive in both equipment and staff, so why not send the experiments to a centralised foundry full of robots? In theory you control everything with code, which is easy, and skip handling molecules, which is hard, and because everything is automated and logged the results should be more precise and more reproducible.

In practice, Justen counts a handful of commercial cloud labs that a scientist or startup can actually use: **Emerald Cloud Lab**, **Strateos** (formerly Transcriptic) and **Ginkgo Bioworks**, whose cloud lab product launched the same month he was writing. Strateos runs two automated labs in California and has partnered with Eli Lilly, but has been moving toward installing its software at client sites rather than expanding its own hosted labs. Emerald and Ginkgo each serve only a small number of customers, and Emerald contracts reportedly start above $250k a year.

Getting an experiment running is not a button click. Each provider has its own software layer, a proprietary Symbolic Lab Language at Emerald and Catalyst at Ginkgo, and the work involves checking compatibility with their instruments and iterating on a real automation setup with their staff. Justen quotes Tessa Alexanian's framing, that starting with a cloud lab looks more like establishing a client relationship than opening an account.

On risk, he takes the threat model seriously but says its proponents treat cloud labs as a black box where any protocol goes in and material comes out. They are not: providers ask about your protocol and your goals. He thinks **contract research organizations**, which are much the same but with humans doing the work, deserve more worry, and that the real cloud lab risk is

generating data to optimise a pathogen. The safeguards he wants are the DNA synthesis ones: screen requests, screen customers, and build the standards and law that make switching suppliers useless.

Why it's assigned

This is the reality check inside her chosen deep dive. Page 2.7 lists cloud labs and contract research organizations alongside DNA synthesis and specialised equipment as chokepoints worth blocking, and asks directly how dangerous cloud labs are. Justen answers that question with a count and a judgement rather than a vibe, which is what the Deny sub-layer needs if it is going to spend scarce effort in the right place. It also hands her the better target: if the workflows that matter are bespoke rather than high-throughput, the oversight question moves from cloud labs to the contract research organizations beside them.

In depth

The claim he is answering

Justen opens by naming the source: Rob Reid, on Making Sense episode 463, Privatizing the Apocalypse, saying there are dozens, maybe a hundred cloud labs, and presenting them as a major emerging source of biological risk. Justen says he agrees with many points made on the podcast, and

disagrees with this one on both the count and the current magnitude of the risk. The disagreement is friendly and specific, which is the tone of the whole piece.

What a cloud lab is, and why the idea is attractive

Cloud labs are highly automated and flexible biological laboratories that a user can operate remotely, without needing their own hardware. The commercial and scientific case, as Justen lays it out, is compelling on paper. Building a wet lab is expensive, in equipment and in the staff to run it, so outsourcing to a centralised foundry that can do everything with robots looks like an obvious trade. The approach should lower the barrier to doing biology, because you control things with code, which he calls easy, instead of manipulating molecules, which he calls hard. And because everything is automated and logged and not subject to the whims of human performance, experiments should come out more precise and more reproducible, which is good for science.

How many there actually are

His answer is that the vision remains largely a vision. As of writing there are only a handful of commercial cloud labs accessible to scientists or startups. The primary ones are Emerald Cloud Lab, Strateos,

formerly Transcriptic, and Ginkgo Bioworks. Ginkgo's cloud lab product launched the month he wrote, as part of a broader pivot into lab automation as a service. Strateos runs two automated labs in California and has partnered with Eli Lilly, but has been pivoting toward deploying its software on-site at client facilities rather than expanding its own hosted cloud labs. That, he says, is about it: a handful of providers, not dozens.

Both Ginkgo and Emerald are, as far as he is aware, still serving only a small number of customers, and Emerald contracts reportedly start above $250k per year.

The friction is the heart of his argument. Getting a protocol running on any lab automation system takes more than clicking a button or asking an AI agent. Each cloud lab has its own niche software layer for specifying experiments: Emerald uses a proprietary Symbolic Lab Language, Ginkgo uses its own Catalyst software. Getting a protocol running means verifying compatibility with their instrumentation, developing the programmatic protocol inside that software environment, and iterating on an actual automation setup. The provider's team will help, but it is an involved process that requires consultation, and it is probably only worth the investment if you need to screen a large number of molecules or run the same protocol many times. For bespoke, low-throughput

work, he says a cloud lab is probably more trouble than it is worth. He borrows Tessa Alexanian's line for the current experience: starting with a cloud lab looks more like establishing a client relationship than opening an account and making an order.

Why the hype runs ahead of the reality

Part of the excitement, he argues, comes from flashy demonstrations. The example he gives is the OpenAI and Ginkgo collaboration in which GPT-5 helped direct thousands of cell-free protein synthesis experiments, optimising reaction conditions and reducing costs by 40%. His reading of it cuts two ways: that is precisely the kind of high-throughput work where lab automation excels, and the process was far from fully autonomous, needing manual reagent adjustments by Ginkgo engineers and some nudging of the model along the way.

He also covers the decentralised route, since it is the obvious reply: liquid handlers at $16k and open-source Python libraries to control them. You can indeed point a coding assistant at a liquid handler or build a small self-driving lab, but you are not removing the need for biological know-how. Liquid handlers remain quite constrained in which instruments they support, and anyone who thought lab automation was purely a coding problem will meet hardware problems instead.

The threat model, and what he thinks is wrong with it

The threat model he hears is that cloud labs could lower the barrier to making dangerous pathogens, by removing the need for wet lab skills and expensive equipment, and by reducing the risk of self-infection or of being noticed.

He calls that a concern worth taking seriously, and then says most proponents of it overestimate both the number of cloud labs and their maturity, treating them as a black box into which arbitrary protocols go and out of which molecules or data come. That is not the current state of affairs. Anyone trying to use a cloud lab to make a pathogen would have some awkward conversations about their goals with the provider, because providers are not a neutral black box and do ask about protocols and end goals.

He adds a second mismatch. For many core dual-use biology workflows, he names reverse genetics as the example, cloud labs are not the best fit, because the work is fairly bespoke rather than enormous and high-throughput. For those workflows he says he would be more concerned about **contract research organizations**, which he describes as rather like cloud labs except that humans do the work.

Where he does see real risk is in using cloud labs to generate valuable data for pathogen optimisation, and he says operators should take meaningful steps to

reduce it. But surveying what he calls the current, highly neglected state of biorisk, cloud lab risk is not among his top concerns. His forward-looking caveat is explicit: lab automation is going to get better, so it is reasonable to start developing and integrating safeguards now even though the overall risk is low today.

The safeguards he recommends

They are deliberately borrowed from commercial DNA synthesis, where providers try to stop malicious actors obtaining dangerous pathogen DNA. Three of them: screen incoming requests for hazardous features; perform **know-your-customer** screening so the end users are legitimate; and develop the standards, legislation and industry groups that get screening and customer checks enforced nationally and ideally globally, so that evading them is harder than simply switching suppliers.

On the first, he argues cloud labs should screen both the nucleic acid substrates and the actual protocols, to be sure they are not facilitating the creation or modification of dangerous pathogens. He points out that this is a biosafety measure as much as a biosecurity one: if automation engineers are walking around and handling equipment, the operator wants to know nobody is producing something dangerous on the setup without their knowledge.

His conclusion, and the definitional loose ends

He closes on his own motivation, saying he is enthusiastically committed to making the world safer against biological threats and unlocking the beneficial applications of biology, the mission around which he has built his PhD and career. The argument for accuracy is strategic as well as factual: given how neglected biosecurity is, resources and attention should go where they matter most, and the field should stay evidence-based, or it risks alienating much-needed allies by crying wolf where there is only a pup.

Three loose ends he leaves open on purpose. First, counting depends on definitions: a 2025 RAND report identified 15 cloud lab organizations globally, but from what he can tell most are internal platforms belonging to private companies, not something a startup or an individual scientist could access. Second, some service companies, he names Arctoris and Adaptyv Bio, brand themselves as cloud labs while being more specialised and less autonomous. Third, there is an official definition to work from: the Biden-era Framework for Nucleic Acid Synthesis Screening defines a cloud lab as a highly automated research laboratory with a diversity of analytical and synthesis capabilities across the life sciences that can be remotely operated by specifying experimental protocols via software. His own recommendation for policy is to stop worrying about the label and apply

much the same security principles either way, treating cloud labs as a subset of contract research organizations that happen to have less direct human intervention and oversight. He also points readers to a RAND report and to Tessa Alexanian's 2023 primer on cloud labs for policymakers, which he says is still relevant.

Words to know

- **Cloud lab**: an automated laboratory you operate remotely by sending it code, without owning equipment. It is the thing being counted, and the count is the argument.
- **Wet lab**: a physical laboratory where someone handles real materials. The contrast with a cloud lab is what supposedly removes the need for hands-on skill.
- **Contract research organization (CRO)**: a company that runs experiments for a client, with people rather than robots doing the work. Justen thinks these deserve more biosecurity attention than cloud labs do.
- **Liquid handler**: a benchtop robot that moves small volumes of liquid. At $16k it is the cheap, decentralised version of automation, and it still needs biological judgement to use.
- **Know-your-customer (KYC) screening**: checking that a customer is who they claim and has a

legitimate reason to order. Borrowed here from DNA synthesis as the second of three safeguards.

- **Reverse genetics**: his example of a core dual-use workflow that is bespoke rather than high-throughput, which is why he argues cloud labs are a poor fit for it.
- **High-throughput**: running very many variations of the same experiment. It is the only regime where cloud labs currently pay for themselves, which bounds who uses them.

Numbers in this reading

- Dozens, maybe a hundred: Rob Reid's claimed number of cloud labs on Making Sense episode 463, and the claim Justen is correcting.
- A handful: Justen's count of commercial cloud labs a scientist or startup can actually use, naming 3 providers.
- 2: automated labs Strateos runs in California.
- Above $250k per year: the reported starting point for Emerald Cloud Lab contracts.
- Thousands of cell-free protein synthesis experiments in the OpenAI and Ginkgo collaboration, with costs reduced by 40%.
- $16k: the price of the liquid handlers behind the decentralised automation route.
- 15: cloud lab organizations identified globally by a 2025 RAND report, most of which Justen believes

are internal corporate platforms rather than services.

- 3 safeguards he recommends: request screening, customer screening, and enforceable standards.

What the author is unsure about

He marks his own uncertainty throughout, and most of it is about the future rather than the present. He says the number of customers at Ginkgo and Emerald is as far as he is aware, and that the Emerald contract figure is reported rather than confirmed. He assumes some nudging of the model happened in the OpenAI and Ginkgo work rather than knowing it. He accepts the threat model is worth taking seriously, and concedes a real risk in using cloud labs to generate pathogen optimisation data, while placing it outside his top concerns. He says plainly that lab automation will get better and that safeguards are worth building now even though current risk is low, which is a judgement about trajectory, not a finding. And he flags the definitional problem honestly: his count differs from RAND's 15 because they are counting different things, and he hesitates to call some automated labs cloud labs at all.

For the discussion

- The most useful correction to bring is the direction of concern: if dual-use workflows are bespoke rather than high-throughput, then contract research

organizations, where humans do the work, are the bigger gap, and cloud labs are best treated as a subset of them with less human oversight.

- The friction is doing security work by accident right now, because onboarding requires consultation with the provider's staff about protocols and goals, and the group should ask what happens to that accidental safeguard as the software gets good enough to remove the conversation.
- A tension worth raising: Justen's argument for accuracy is partly about credibility, that overstating a small risk costs allies the field needs, which is a real cost but also an argument that could be used to delay any safeguard before the harm shows up.

2.7-4 · FURTHER · COURSE TIME 10 MIN

Customer Screening Resources

International Biosecurity and Biosafety Initiative for Science (IBBIS) · 2026

Read the original paper → · the course assigns: the whole page, including its description of the screening exercise and its forms and guidance

> **IBBIS is trying to turn the judgement call at the heart of DNA order screening into paperwork anyone can follow: two forms, two step-by-step decision guides, and a set of customer profiles drawn from real scientists**

> **and real attempted bioterror that providers practise on.**

Average summary

This is an organisation's working page, not an article. **IBBIS**, the International Biosecurity and Biosafety Initiative for Science, builds resources to help life science providers make know-your-customer decisions, which it also calls know-your-collaborator. Its own timeline reads: first customer screening resources released in 2024, currently developing practical tools for synthetic DNA providers, and in future, legitimacy verification standards across the life sciences.

The problem statement is blunt. Many providers already screen customers before selling or sharing materials such as synthetic DNA encoding pathogens and toxins. What is hard is designing a process consistent and rigorous enough to catch potentially illegitimate actors while still enabling legitimate science, and IBBIS says that difficulty becomes especially clear when serving customers and collaborators outside the Global North and outside established academic settings. At present screening often rests on subjective judgements about legitimacy, which is why IBBIS consulted global stakeholders to build structured templates and flowcharts instead.

The logic linking the two halves of screening is stated plainly: the decision whether to fill an order hinges on whether the customer should be trusted with those sequences, and orders for potentially risky sequences require a higher level of confidence in the customer, so sequence screening and customer screening both feed one decision. IBBIS also offers a free tool to identify sequences of concern, through its Common Mechanism project.

The documents themselves come in two tiers. The **New Customer Form** is filled in once, at onboarding, to establish that a customer is a member of the life sciences community with a legitimate reason to order synthetic nucleic acids; it is available as a fillable PDF, a Word document or plain text. Its companion decision guide walks a provider, which might be a company, a core facility or a biofoundry, through how to read the answers, how to verify a customer's identity, and what the red flags are. The second tier applies to orders containing sequences that contribute to pathogenicity, virulence or toxicity. For those, the **SOC Order Form**, for sequences of concern, asks about the order, its intended end use, and institutional approvals or other biorisk management practices, and has its own step-by-step decision guide covering interpretation, verification of legitimacy and red flags.

The practical exercise the course wants her to try is built on a series of customer profiles based on real examples of legitimate scientists and attempted bioterror, designed to show how hard verifying legitimacy actually is. Providers have used them in interactive screening exercises, and they have run at workshops hosted by the Global Biofoundries Alliance, the iGEM Competition, the Engineering Biology Research Consortium and SynBioBR, among others. The slides and profiles are published under a Creative Commons licence, with permission to translate them so long as the content stays faithful. IBBIS says refinement is ongoing and asks for suggestions by email.

Why it's assigned

This is the hands-on half of her chosen deep dive, Deny by safeguarding research infrastructure. Where the page's other readings argue about how the DNA supply chain should be governed, this one is the actual instrument: the forms a provider sends, the red flags they look for, and an exercise that puts her in the screener's chair to feel how thin the evidence is. It answers the course page's question about why know-your-customer protocols are standard in finance but not in biotechnology, because here you can see what a biotechnology version has to invent from scratch, including how to treat a legitimate scientist who is not at a well-known Western university.

Words to know

- **KYC (know-your-customer, or know-your-collaborator)**: checking who you are selling to and whether they have a legitimate reason to buy. Borrowed from finance, and IBBIS is writing the biology version.
- **Sequence of concern (SOC)**: a sequence that contributes to pathogenicity, virulence or toxicity. Ordering one triggers the second, stricter form.
- **Biofoundry**: a facility that builds biological material to order using automated equipment. It is named as one of the provider types the decision guides are written for, alongside companies and core facilities.
- **Common Mechanism**: IBBIS's freely available tool for identifying sequences of concern. It handles the sequence half of the decision, which the customer forms complement.
- **Legitimacy verification**: establishing that a customer is genuinely part of the life sciences community with a real use. IBBIS's stated future goal is a standard for this across the life sciences, not only for DNA.
- **Red flags**: the specific warning signs a decision guide tells a provider to look for. They are what turns a subjective impression into something another provider could reproduce.

Numbers in this reading

- 2024: the year IBBIS released its first customer screening resources.
- 2 forms, the New Customer Form at onboarding and the SOC Order Form for sequences of concern, each with its own step-by-step decision guide.
- 3 file formats offered for the New Customer Form: fillable PDF, Word document and plain text.
- 3 white papers and reports listed: Implementing Emerging Customer Screening Standards for Nucleic Acid Synthesis, August 2025, by Sarah R. Carter, Lucas Boldrini and Tessa Alexanian; Verifying Legitimacy: Findings from the Customer Screening Working Group, 2020-2023, February 2024, by Tessa Alexanian and Sarah R. Carter; and Biosecurity Innovation and Risk Reduction: A Global Framework for Accessible, Safe and Secure DNA Synthesis, January 2020, by Elizabeth Cameron, Sarah R. Carter, Jacob Jordan and Ryan Morhard for the NTI-WEF working group on preventing illicit gene synthesis.

What the author is unsure about

IBBIS is open about the limits of its own work. It says designing screening that is consistent and rigorous enough to catch illegitimate actors while enabling legitimate science is not easy, and that the problem is especially apparent for customers outside the Global North and outside established academic settings. It says screening today often relies on subjective

judgements about legitimacy, which is the gap its templates are meant to close rather than something it claims to have closed. It describes refinement of the documents as ongoing and asks readers to write in with ways to make them work for a wider range of international providers and customers. The future item on its own timeline, legitimacy verification standards across the life sciences, is stated as an ambition.

For the discussion

- The structural point worth making is that IBBIS put the strictness in the right place: one light form at onboarding for everyone, and a second, harder form only for sequences that contribute to pathogenicity, virulence or toxicity, so the burden scales with the risk rather than falling on every order.
- The exercise is built from real cases of both kinds, legitimate scientists and attempted bioterror, which is the honest way to train a screener, and it is worth saying out loud that the hard profiles are the legitimate ones who look wrong rather than the attackers who look suspicious.
- A question worth asking: IBBIS names serving customers outside the Global North and outside established academia as where this gets hardest, so whose science gets slowed down when

legitimacy is proved by institutional letterhead, and who decides what counts as proof?

2.4-1 · CORE · COURSE TIME 35 MIN

Expanding on Future Biological Weapons Convention Verification: An Interview Dr. Piers Millett

Dr. Dan Regan · 2023

Read the original paper → · the course assigns: whole interview

> **Piers Millett argues that the Biological Weapons Convention cannot make political progress on verification until states redo the technical work they last did in 1993 and 1994, and that in biology the realistic goal is a collaborative process of showing there is nothing to worry about, not traditional arms-control proof.**

Average summary

This is an interview. **Dan Regan**, writing for the Council on Strategic Risks, asks the questions. **Dr. Piers Millett** answers. Millett is Executive Director of the International Biosecurity & Biosafety Initiative for Science (IBBIS) and a co-author of a recent article called "Feasibility of On-Site Verification," written with

experts from academia, private industry and non-government organisations.

Regan sets the scene. The biodefense community went into the Ninth Review Conference of the Biological Weapons Convention (BWC), held in Geneva at the end of 2022, uneasy: earlier review conferences had stalled, Russia was spreading disinformation, and there were concerns about non-compliance. The outcome was better than expected. A working group was set up to prepare for the Tenth Review Conference, covering seven topics, one of them "measures on compliance and verification."

Millett's core point is that the technical groundwork is three decades old. The last serious technical work was the 1991-1994 Ad Hoc Group of Governmental Experts, which produced the VEREX Report. Drafting of a binding verification instrument ran from 1995 to 2001 and collapsed after rejections in 2001 by the President George W. Bush administration. Nobody had touched the technical side since 1993/1994. His favourite line from the paper: the report of that last technical review was written on a typewriter.

He makes a distinction he says sounds academic but matters: verification versus compliance. He thinks traditional arms-control verification in biology would need a level of intrusiveness that genuinely puts national security and commercial proprietary information at risk. His chapter reframes the question.

Instead of proving yes or no whether someone breached the treaty, states would run a more collaborative process of showing there is nothing to worry about. Either way, he says, the final call on whether a country has complied is a political decision, not a technical one.

On resources he is blunt. When he worked in the BWC Implementation Support Unit, the only standing body the convention has, staff used to say its annual budget was roughly the same as a local McDonald's. Regan adds that the Unit has just got consensus to add its fourth full-time staff member.

Two further themes: use the data that already exists in the public domain rather than piling new reporting burdens on companies and governments, and face the fact that we lack good enough baseline disease data to know what normal looks like.

Why it's assigned

Day 2 is the Delay layer, and this page is its first sub-layer, Deter: making a state not want a biological weapons programme in the first place. The course page says the BWC currently has no verification mechanism and fewer than 5 permanent staff, and asks why it has no verification mechanism. This interview is the inside answer from someone who watched the negotiations collapse and then worked in the convention's own support unit, and it names the

live process, the working group heading to the Tenth Review Conference in 2027.

In depth

The setup: a review conference people expected to fail

Dan Regan of the Council on Strategic Risks asks the questions; Dr. Piers Millett, who runs IBBIS and co-authored "Feasibility of On-Site Verification," answers. Regan opens on the mood going into the **Ninth Review Conference** of the BWC, Geneva, end of 2022: uneasiness, from the stalled progress of earlier review conferences, Russian disinformation and concerns about non-compliance. The result was generally positive, and attention has returned to verification. A working group was set up to prepare for the Tenth Review Conference, with seven topics, one of them "measures on compliance and verification."

The history: the last serious technical work was the 1991-1994 Ad Hoc Group of Governmental Experts to Identify and Examine Potential Verification Measures. Its VEREX Report set out possible on-site and off-site measures and stressed that no single measure would do. From 1995 to 2001 the Ad Hoc Group drafted a binding verification resolution, shuttered after rejections in 2001 by the President George W. Bush administration.

Why Millett went back to the technical work

Millett entered biosecurity as an academic in an NGO just as the protocol talks were ending. His first trip to the convention was in about 2000, to the Ad Hoc Group, and he was interning inside the UN for the Fifth Review Conference, where the collapse played out. Then the topic disappeared for two decades at least.

His motivation: you cannot make political progress without going back to the technical work. States Parties keep calling to return to the old negotiations, but those rested on an earlier process about the science of telling when someone is breaking the treaty, untouched since 1993/1994. His favourite anecdote: the last technical review report was written on a typewriter.

He assembled co-authors with different skills on purpose: academic research and development, industry, people running multiplexed platform facilities that produce a wide variety of biological materials. One assumption was borne out: it is not individual developments that made the difference, it is the way science is conducted and transitioned into engineering and real-world applications.

Why you cannot future-proof the convention

Future-proofing is virtually impossible, Millett says, because he is fundamentally a believer in development. He hopes we are living in the century of

biotechnology and does not want to get in the way; he wants more of it done responsibly, safely and securely. He is not comfortable claiming he knows what the industry will look like in another 20 to 30 years, but he is 100% sure it will not look like the industry of 20 or 30 years ago. So the task is not future-proofing. It is catching up with where we already are.

Verification versus compliance

The distinction seems academic, he says, but is vitally important in biology. He personally thinks we will struggle to reach the certainty traditional arms-control verification needs without intrusiveness that genuinely puts national security and commercial proprietary information at risk.

The chapter's compliance framing changes the question. Rather than proving yes or no whether somebody breached the treaty, he envisages a more collaborative process of proving there is nothing to worry about: different questions, similar tools. And whether a country has complied is, in the end, a political decision rather than a technical one.

Will the Tenth Review Conference take this on

Getting consensus at review conferences is very, very hard and has got harder, Millett says; he does not think we live in an epoch with enough global consensus on how to reach a better future. That is not unique to the BWC, and compared with other arms

control treaties' recent review conferences the BWC did particularly well in 2022, given the geopolitics and the accusations around development and use of biological weapons. His own judgement: compared with other global threats, biological weapons and malign misuse of biotechnology are massively underrated.

The budget of a local McDonald's

Millett worked in the BWC **Implementation Support Unit**, which he calls the only body the convention has at the moment. Staff used to say its annual budget was roughly the same as a local McDonald's. If that is what the world collectively spends on preventing biological weapons, he says, that is kind of ridiculous: there is low-hanging fruit available, and a real need for more ambition and bigger systems. Regan adds that the Unit has just got consensus to add its fourth full-time staff member, which invites comparison with the chemical weapons convention's implementing body, the OPCW.

Does the BWC need help from outside

Asked whether an external or parallel entity is needed, Millett says the challenge is big enough that all of the above is most appropriate. Not every state is ready to work at the same speed, and those that see a greater threat can move faster together than the whole membership; he names the Global Partnership and its

Africa Signature Initiative as a model. (The Global Partnership is a multi-country funding partnership against weapons of mass destruction; not from this reading.)

Then he plugs IBBIS, which is building the **Common Mechanism**, an open and accessible gene synthesis screening tool meant to let more people in more places screen gene synthesis. Such an initiative is probably too risky for an intergovernmental organisation to launch, he argues: hard to tell members you will build a tool when nobody knows how it will work. Easier when an international organisation builds it, proves it works, and shows everyone what they would be signing up to; IBBIS can then hand it to any intergovernmental organisation ready to take it.

Transparency: use the data that already exists

Regan notes that virtually every part of biotech could count as dual-use, and asks whether a few countries withholding things like CBMs can stonewall the BWC. (CBMs are confidence-building measures, the annual reports States Parties have politically committed to submit; not from this reading.)

Millett calls existing data one of the paper's most interesting aspects. A huge amount of information is already in the public domain, and ignoring it because a state did not officially provide it looks to him like a

wasted opportunity. If the worry about a compliance regime is the extra burden on companies and governments, existing data is how you minimise it.

The baseline-data problem

His second point cuts the other way: existing data sources are nowhere near good enough. A key finding from the report was that we lack good enough baseline data on disease to know what normal is. If we do not know what normal is, we do not know what unusual is; if we do not know what unusual is, we cannot tell what is unusual but unnatural. He is not sure finding something unnatural could ever be automated, certainly not in the short to medium term, but a lens that flags incidents unusual enough to deserve extra attention would be a step forward. He points to metagenomic sequencing and the sequencing capability proven during the recent pandemic, which show what is normally circulating and how it changes.

The obstacle: moving from diplomacy back to technical talk

The biggest challenge, Millett says, is moving from the diplomatic and political process back to technical discussions. Diplomacy, in his experience, runs on **constructive ambiguity**: words everybody can go home believing mean what they wanted. That delivers an outcome at a high-level meeting, but if everyone

went home thinking something different, then when it comes time to act the starting points are dissimilar and you reopen all the national gaps you had cunningly painted over.

He then caveats his own paper. It is one snapshot by a subset of experts with particular backgrounds; States Parties collectively have not done that exercise since 1993; so nobody should rush to conclude that a verification protocol should not be negotiated. He would prefer states first discuss how things have changed. He is not sure there will be consensus that this was the purpose of the process, and suggests a **VEREX 2.0** could be negotiated in the next few years, perhaps not until after the 10th Review Conference in 2027. That is a forecast, not a finding.

Where the problem now sits: almost every country

Two closing reflections. First, this is no longer a challenge for a small subset of countries. In the 1990s only a small number had a biotech industry advanced enough for a breakout biological weapons programme with militarily relevant production capability. Now, he says, the vast majority of countries could run one if they chose to.

He does not want that capability reduced. He wants more countries to have biotech industries: more vaccine manufacturing, more multiplex facilities

producing nanoparticles to encase mRNA technologies, because the next natural pandemic will need them. (The transcript prints "paramedic" here, a likely transcription slip.) The spread is not a bad thing, but verification can no longer be pushed onto a few countries to figure out.

Second, with dual-use technologies, and especially biological warfare programmes rather than bioterrorism, the cases warranting higher focus are where advanced biotechnology overlaps with military or national security capability. Dealing only with programmes under direct government control would not solve every problem, but it may be palatable, because a government-run facility is far easier for a government to grant access to and share information on.

The further reading he suggests

Richard Guthrie's Daily Reports from the BWC. iGEM's March 2022 Wilton Park event, the Responsibility Workshop: Designing a BWC Compliance Regime. And Jez Littlewood's 2005 book on the failed verification push at the 5th Review Conference, *The biological weapons convention: a failed revolution*.

Words to know

Verification - in arms control, the technical business of establishing whether a country is breaking a treaty, to a standard of certainty that supports a formal

accusation. Millett doubts biology can reach that standard without unacceptable intrusiveness.

Compliance - whether a state is in fact meeting its obligations. Millett separates it from verification on purpose: his framing asks states to show there is nothing to worry about, and he says the final judgement is political, not technical.

Review Conference - the meeting where BWC States Parties review the treaty and decide what happens next. The Ninth was Geneva, end of 2022; the Tenth is in 2027.

VEREX - the report of the 1991-1994 Ad Hoc Group of Governmental Experts, the last technical survey of possible verification measures for the BWC. Millett floats a "VEREX 2.0" to redo it for today's biotech.

Implementation Support Unit - the BWC's only standing body, which Millett says had an annual budget roughly the size of a local McDonald's and has just reached consensus to hire a fourth full-time staff member.

CBMs, confidence-building measures - the abbreviation Regan uses without unpacking it. They are the annual reports States Parties have politically committed to submit about their biological facilities and biodefence work (not from this reading).

Common Mechanism - the open, accessible gene synthesis screening tool IBBIS is building, meant to let more organisations in more places screen DNA synthesis orders.

Metagenomic sequencing - sequencing everything in a sample at once rather than hunting for one known organism, which is how you learn what is normally circulating and how it is changing.

Numbers in this reading

- Ninth Review Conference of the BWC: held in Geneva at the end of 2022.
- Seven topics for the working group preparing the Tenth Review Conference, including "measures on compliance and verification."
- 1991-1994: the Ad Hoc Group of Governmental Experts that produced the VEREX Report.
- 1995-2001: the Ad Hoc Group's work drafting a binding verification resolution.
- 2001: the year the drafting effort was shuttered, after rejections by the President George W. Bush administration.
- About 2000: Millett's first trip to the bioweapons convention, to the Ad Hoc Group.
- 1993/1994: the last time anyone touched the technical question of telling when a state is breaking the treaty.

- Two decades at least: how long Millett says the verification topic disappeared for.
- 20 to 30 years: the span over which he says he cannot predict the biotech industry, and the span of the past industry it will not resemble.
- 100%: his stated confidence that the future biotech industry will not look like the industry of 20 or 30 years ago.
- One local McDonald's: what the BWC Implementation Support Unit's annual budget was said to be roughly equal to.
- Fourth full-time staff member: what the Implementation Support Unit has just got consensus to add, per Regan.
- 2027: the year of the 10th Review Conference.
- 1993: the last year States Parties collectively did a technical landscape exercise.
- March 2022: iGEM's Wilton Park event on designing a BWC compliance regime.
- 2005: Jez Littlewood's book on the convention and the failed verification push.

What the author is unsure about

Millett flags his own limits repeatedly. He says future-proofing the convention is virtually impossible, and that he is not comfortable claiming to know what biotech will look like in 20 to 30 years. He is not sure the search for an unnatural disease event could ever

be automated, certainly not in the short to medium term. He is not sure there will be consensus that landscape analysis was even the purpose of the current working group, and his VEREX 2.0 suggestion is a possibility he floats, not a plan anybody has adopted. Most pointedly, he puts a caveat on his own paper: it is one snapshot by a subset of experts with particular backgrounds, States Parties have not done that exercise since 1993, and so nobody should rush to conclude that a verification protocol should not be negotiated. Several of his strongest statements are explicitly labelled as personal, including that biological weapons and malign misuse of biotechnology are massively underrated, and that the overlap of advanced biotech with military capability is where focus belongs.

For the discussion

- The course page asks why the BWC has no verification mechanism; Millett's answer is that the political argument has been running for two decades on top of technical work last done in 1993 and 1994, when the report was typed on a typewriter.
- Worth raising as a tension: Millett doubts biology can ever reach arms-control standards of certainty without intrusion that risks national security and trade secrets, so his fallback is a collaborative

process of showing there is nothing to worry about, which is a weaker thing than proof.

- Worth asking: if we have no baseline data good enough to know what normal disease looks like, how would any verification regime, or any attribution claim, tell unusual from unusual but unnatural?

2.4-2 · CORE · COURSE TIME 20 MIN

Tracing engineered biothreats with AI forensics: Five steps to improve attribution

Oliver M. Crook, Anemone Franz and Aaron Maiwald · 2025

Read the original paper → · the course assigns: whole piece, from the saved structured capture in raw/2.4-2.txt (a detailed section-by-section capture of the full article, not verbatim text; the saved HTML is only a Cloudflare challenge page)

> **Three researchers argue that the world cannot currently trace an engineered pathogen back to whoever made it, and set out five steps, mostly about data, that would turn AI attribution from an academic result into a forensic capability.**

Average summary

Summarized from a detailed structured account of the article, not its original text, so nothing here is a direct quotation

This is an **opinion** piece in the Bulletin of the Atomic Scientists, published 1 December 2025, by Oliver M. Crook, Anemone Franz and Aaron Maiwald. It is an argument and a set of recommendations, not a study with findings of its own.

The gap they name is concrete. The 2001 US anthrax letter attacks, the 2007 UK foot-and-mouth outbreak and the COVID-19 pandemic all share one feature: investigators struggled to establish their origins despite extensive effort. Modern forensics cannot reliably trace biological threats to their sources.

They separate two jobs. **Detection** establishes whether a pathogen was modified at all. **Attribution** traces the modifications back to a source. Attribution is possible in principle because engineering leaves traces, from obvious insertions of foreign genes to subtle single-base replacements, and the many choices a designer makes add up to a genetic fingerprint.

AI has shown it can read those fingerprints. The number they lead with: in a 2021 genetic engineering attribution challenge, the best system identified the correct laboratory among more than 1,300 possibilities in 81.9 percent of cases. But that work was mostly done on plasmids in a curated academic repository. Real investigations face degraded samples, incomplete sequences and adversaries actively obscuring their tracks.

Their five recommendations are mostly about data, not algorithms. One, build comprehensive datasets of both natural and engineered pathogen sequences, including short fragments, because wastewater surveillance yields fragments rather than whole genomes, and because existing natural-sequence databases are often contaminated with mislabelled engineered material. Two, train attribution systems on pathogens rather than plasmids, starting with viral vectors; Addgene currently hosts over 1,000, and they estimate a capable system would need that scaled by one or two orders of magnitude, with at least 10 data points from each of several thousand laboratories. Three, cover Biosafety Level 3 and higher labs, which are absent from public repositories, while recognising that such a dataset is itself a biosecurity risk. Four, extend attribution to countries and regions so covert laboratories are not invisible; a 2020 MIT study predicted nation of origin with over 75 percent accuracy across 33 countries. Five, prioritise transparency, because current systems are mostly black boxes and a confident wrong answer could be catastrophic.

They are careful about limits. Some techniques leave no detectable signature, genomic evidence alone will probably not identify a perpetrator, and adversaries may plant false signatures.

Why it's assigned

Day 2 is the Delay layer and this page is its Deter sub-layer, where the course asks how you reduce a state's or group's intent to build a biological weapon. Attribution is the piece that makes deterrence credible: if nobody can say who made it, nobody can be held responsible. The course page asks how attributing a virus to a country or lab would actually work, and this is the current technical answer plus an honest account of how far it is from being usable.

In depth

A note on the source text

The saved raw file is a detailed, section-by-section capture of the full article rather than its verbatim words. It keeps the article's own headings, every number it reports, and its caveats. Two things are missing from the capture: one photograph (a wastewater sample being analysed for SARS-CoV-2 contamination, credited to Adobe) and the comment thread, which had 0 comments when the page was read.

The gap the piece names

Crook, Franz and Maiwald start with three cases: the 2001 US anthrax letter attacks, the 2007 UK foot-and-mouth outbreak and the COVID-19 pandemic. In each, investigators struggled to establish origins despite extensive effort. They call this a critical gap in biosecurity.

When a novel pathogen emerges, investigators face a cascade of questions. What is it. Is it familiar or new. Did it come from natural spillover, a lab accident, or a deliberate attack. And if the origin is not natural, who made it.

That last question grows in importance as biotechnology advances, along two dimensions. More laboratories doing high-risk pathogen research raises the chance of accidents involving engineered organisms. And advancing biotechnology may let bad actors design pathogens that spread more efficiently or kill more effectively than anything natural. Either way, attack or accident, tracing the thing to its source is crucial.

Then the distinction the whole piece rests on. **Detection** capability establishes whether a pathogen was modified. **Attribution** traces the modifications back to a source.

Why genetic attribution matters

Three payoffs, in their framing. Reliable detection and attribution would deter actors whose designs could be traced to them. It would enable accountability through diplomatic or legal action. And it could inform rapid response, by revealing a pathogen's design characteristics and the creator's objectives.

Attribution is possible because engineering techniques leave traces: from obvious insertions of

foreign genes down to subtle single-base replacements. The many techniques and choices involved in engineering a sequence leave genetic fingerprints that could identify the designer.

Recent AI work has proven effective at finding those fingerprints. Most of that research trains on **plasmids**, small bacterial DNA molecules commonly used in genetic engineering. Systems are trained on thousands of plasmid sequences plus host species and other metadata, and learn lab-specific patterns automatically. The benchmark they report: in a 2021 genetic engineering attribution challenge, the best system identified the correct laboratory among more than 1,300 possibilities in 81.9 percent of cases.

They immediately stress this is not a solved problem. Real investigations are harder than sorting plasmids in a curated academic repository: degraded samples, incomplete sequences, and adversaries actively obscuring their tracks.

Recommendation 1: Build comprehensive detection datasets

Academic studies usually start from sequences already known to be engineered. Forensic work has to do the earlier job first: establish whether engineering happened at all. Current pathogen detection systems usually specialise in one form of engineering, such as

large insertions of foreign gene content, and struggle across a wider range of practices.

The first step toward a generalised detector is comprehensive datasets of both natural and engineered pathogen sequences, DNA or RNA, covering complete genomes and short fragments. Fragments matter especially for environmental monitoring such as wastewater surveillance, where sequencing yields many small fragments rather than whole genomes. A further problem they flag: existing natural-sequence databases are often contaminated with mislabelled engineered material and need careful curation. Robust detection is the foundation for attribution, so dataset building is the essential first step for both.

Recommendation 2: Train reliable attribution systems for pathogens

Pathogen attribution hits the same bottleneck. Publicly available engineered pathogen genomes are scarce and insufficient for training robust models, and systems trained mainly on plasmids will at best have limited accuracy on pathogen genomes.

Their proposed starting point is **viral vectors**: engineered viral particles used widely in research and medicine to deliver genetic material, modified so they cannot replicate, which makes them safer to work with. The figure they report: the global plasmid

repository Addgene currently hosts over 1,000 viral vectors. Their own estimate, labelled as an estimate: a highly capable attribution system for viral vectors would require Addgene or a similar initiative to scale that dataset by one or two orders of magnitude, covering at least 10 data points from each of several thousand laboratories.

Recommendation 3: Expand laboratory coverage to Biosafety Level 3 and higher

A system trained on public data can only identify laboratories that contributed training sequences, which in practice means academic repositories. The labs posing the greatest biosecurity and biosafety risk, those working with dangerous pathogens at **Biosafety Level 3** and above, are notably absent from repositories like Addgene.

So they recommend collecting engineered sequences from BSL-3-and-higher labs, and in the same breath note that the dataset itself would be a significant biosecurity risk requiring strict access control and monitoring to prevent misuse. Because publishing research on such datasets carries security risk, they argue academics may no longer be the right people to develop these attribution systems, and government agencies may have to carry the work forward outside the public domain. That is an argument about who should hold the capability, not a finding.

Recommendation 4: Develop attribution for covert laboratories

Data collection from secret state or non-state laboratories may be impossible. To cover covert operations, they recommend expanding attribution beyond individual labs to countries or regions, leveraging training data from registered labs in the same region, on the reasoning that nearby covert facilities likely share methodological practices.

The proof of concept they cite: a 2020 study by MIT researchers predicted the nation of origin for engineered plasmids with over 75 percent accuracy across 33 countries. They recommend developing systems that attribute sequences to nations, to regions, and potentially to individual scientists.

Recommendation 5: Prioritize transparency

Attribution that reaches covert laboratories matters for deterrence. But using attribution evidence after an incident, to prosecute a terrorist group or a state actor, raises a further problem the authors put as a question: would governments trust a result enough to impose diplomatic sanctions or bring criminal charges?

Credibility requires understanding how these systems reach their predictions, and current attribution AI operates mostly as **black boxes**: accurate, but arriving at conclusions humans cannot verify.

Transparent systems would let investigators examine why a conclusion was reached and confirm it rested on genuinely distinctive features. The failure mode they name is specific: an opaque system producing a confident but flawed attribution, for example fingering a lab on the basis of a sequence that looks unique but is actually common contamination across many labs.

Their example that transparency is achievable: researchers at Rice University developed **PlasmidHawk**, an attribution system reaching near state-of-the-art performance while also identifying the specific DNA fragments that supported its laboratory predictions. PlasmidHawk is inherently interpretable. The other route is building better interpretability into black-box models. They say both are worth pursuing, and that interpretability should be a core goal so systems supply defensible reasoning alongside predictions.

Limitations of genetic engineering attribution

The authors give limitations their own section. Some techniques leave no detectable signature: modern methods can join DNA fragments without leaving scars, and traditional approaches such as serial passage can evolve pathogens toward desired traits through selective pressure alone, without directly modifying sequences.

Even where engineering is detected, genomic evidence alone is unlikely to identify perpetrators definitively. Reliable attribution will probably need multiple lines of evidence: geographic emergence patterns, epidemiological data, intelligence sources. Sophisticated adversaries may also use evasion techniques, or plant false signatures to mislead investigators.

Their conclusion on this point: attribution is one component of a broader forensic approach rather than a complete solution, and these limitations are challenges to overcome rather than reasons to abandon attribution.

A critical capability

The closing argument. Persistent uncertainty about past outbreaks, with COVID-19 the most salient example, has eroded public trust, strained international relations and left security questions unanswered. To avoid repeating that as genetic engineering becomes more accessible, they argue the United States needs forensic capability to trace engineered threats to their sources, supplying evidence for accountability, intelligence for response, and deterrence against future attacks. The technologies are advancing quickly, and their five recommendations chart a path to operational attribution systems. Attribution will not solve every biosecurity problem, but they call it a critical

capability that should not be neglected, and the difference between uncertainty and accountability in the next crisis.

Who wrote it, and who they thank

Oliver M. Crook is a computational and biophysical chemist at the University of Oxford. Anemone Franz is a visiting research fellow at the American Enterprise Institute, focused on biosafety and biosecurity. Aaron Maiwald is a second-year PhD student in Statistics at Oxford University. (The site truncated each biography with a Read More link.) They thank Ethan Alley, Jeff Kaufmann, Gregory Lewis, Mike McLaren, Shrestha Rath, Ben Snyder and Jake Swett for conversations and feedback on earlier drafts, and take responsibility for remaining mistakes.

Words to know

Detection, in this piece - establishing that a pathogen was engineered at all. It comes before attribution, and the authors say current detectors usually handle only one kind of engineering.

Attribution - tracing the engineering back to a source: a lab, a country, a region, possibly a person.

Plasmid - a small bacterial DNA molecule commonly used in genetic engineering. Most attribution AI was trained on plasmids, which is why the authors say it will not transfer well to pathogen genomes.

Addgene - the global plasmid repository the authors treat as the main public data source. It currently hosts over 1,000 viral vectors, and BSL-3-and-higher labs are notably absent from it.

Viral vector - an engineered viral particle used in research and medicine to deliver genetic material, modified so it cannot replicate, which makes it safer to work with. The authors propose it as the practical starting point for pathogen attribution.

Biosafety Level 3 and above - the containment tiers used for dangerous pathogens. These are the labs the authors say matter most for biosecurity and are missing from the training data.

Black box - a model that is accurate but cannot show its reasoning. The authors argue an attribution result nobody can inspect will not carry the weight of sanctions or criminal charges.

PlasmidHawk - an attribution system from researchers at Rice University that reaches near state-of-the-art performance while naming the specific DNA fragments behind its prediction.

Numbers in this reading

- 2001: the US anthrax letter attacks, one of the three cases where origins proved hard to establish.
- 2007: the UK foot-and-mouth outbreak, the second case.

- 81.9 percent: the share of cases in which the best system in a 2021 genetic engineering attribution challenge identified the correct laboratory.
- More than 1,300: the number of possible laboratories in that challenge.
- Over 1,000: the viral vectors Addgene currently hosts.
- One or two orders of magnitude: how much the authors estimate that viral-vector dataset would need to grow.
- At least 10 data points from each of several thousand laboratories: the coverage their estimate implies.
- Over 75 percent accuracy across 33 countries: the result of a 2020 MIT study predicting nation of origin for engineered plasmids.
- Biosafety Level 3 and above: the containment tier the authors want brought into the training data.
- 1 December 2025: the publication date of the piece.

What the author is unsure about

The authors are explicit that attribution is not a solved problem and that real investigations are harder than curated academic benchmarks, because samples degrade, sequences are incomplete, and adversaries obscure their tracks. They say plasmid-trained systems will at best have limited accuracy on

pathogen genomes. Their dataset-scaling figure is labelled as their own estimate. They concede that data collection from covert labs may be impossible, and that their regional workaround rests on an assumption, that nearby covert facilities likely share methodological practices. They name techniques that leave no detectable signature at all, so a negative result proves nothing. They doubt genomic evidence alone can definitively identify a perpetrator and say attribution will probably need epidemiological, geographic and intelligence evidence alongside it. And they raise, without answering, whether a government would trust an AI attribution enough to impose sanctions or bring charges. The whole piece is published as opinion.

For the discussion

- The course page asks how you would attribute a virus to a country or lab; this piece says the method works in a curated repository and the real obstacle is data, specifically that the labs that matter most, Biosafety Level 3 and above, are the ones missing from it.
- Worth raising as a tension: their own recommendation creates a new hazard, since a dataset of engineered sequences from the highest-containment labs would itself be a significant biosecurity risk, which is why they suggest moving

the work out of academia and into government hands.

- Worth asking: if 81.9 percent accuracy among 1,300 labs is the current best, what accuracy would a government need before it would impose sanctions or bring criminal charges on that basis, and who decides that threshold?

2.4-3 · FURTHER · COURSE TIME 25 MIN

The biosecurity benefits of genetic engineering attribution

Gregory Lewis, Jacob Jordan and Gregory D. Koblentz et al. · 2020

Read the original paper → · the course assigns: whole paper

> **The authors argue that reading a genetically engineered organism's sequence for the design habits of whoever built it could become a real forensic tool, one that would make attribution of an engineered outbreak more accurate and so deter reckless and malicious work, and they are blunt that at 70% accuracy it is nowhere near a smoking gun.**

Average summary

This is a comment piece in Nature Communications, 2020, by Gregory Lewis, Jacob Jordan, Gregory D.

Koblentz and colleagues; the journal's own citation lists Lewis, Jordan, Relman and others.

Their starting point is **attribution**: after a human-caused biological event, working out who was responsible. They give it three security benefits. Knowing who did it informs the response, because it reveals motives and capabilities. It identifies parties for civil, criminal or diplomatic penalty. And successful attribution followed by real consequences deters reckless or malicious work in the first place.

They split the evidence into three kinds. Non-technical indicators give contextual clues to intent: the victims, the location, epidemiological features. Intelligence covers human sources, intercepted communications and surveillance data. Technical forensics reads the agent itself. Microbial forensics is still young: it helped the FBI identify a suspected lab of origin for the anthrax used in the 2001 attack, and a suspected perpetrator.

Technical forensics needs two separate capabilities: establishing that the organism was engineered at all, and identifying who engineered it. Detection work is underway, including IARPA's Finding Engineering-Linked Indicators project, FELIX, and an international trusted laboratory network under the UN Secretary-General's mechanism for investigating alleged biological attacks. The authors note that performance

across laboratories in detecting genetic modifications is currently variable.

The paper's idea is that the second capability is possible because engineers have habits. Which genes or features to use, where they come from, how they are incorporated: those choices reflect training, prior experience, habits and available resources, and in aggregate they form a **methodological signature**. Machine learning was shown to read those signatures in late 2018, with a limited accuracy of 48%. Then Alley and colleagues used deep learning on plasmids submitted to Addgene, the largest repository of its kind with 70,000 submissions from labs in 37 countries, reaching 70% accuracy distinguishing between over 1000 labs. They also added uncertainty estimation, tracking of engineering group "genealogies", and inference of the nation where the lab sits.

The authors then argue this is an unusually attractive biodefense investment. It does not impede scientific enquiry, and offers responsible engineers credit and protection for their intellectual property. It has limited offensive use, so it is less likely to trigger the security dilemma, where one state's defensive work provokes an arms race. And its effectiveness grows with the same biotechnology trends that make misuse more worrying: more engineered sequence data, and more diverse methods to leave signatures in.

The last third is limits, and they are heavy. Attribution is not detection and should not be improvised into it. Failure to attribute does not rule out engineering. Misuse by someone other than the designer breaks the chain. And the tools have only been tested on engineers working in the clear.

Why it's assigned

Day 2 is the Delay layer, and this page is its Deter sub-layer: reducing the intent to build a biological weapon. The course page names genetic engineering attribution as one of three topics to dig into, and asks how attributing a virus to a lab or country would actually work. This paper is the case for why that capability is worth building, written by people close to the field, with a sober account of what it cannot do.

Words to know

Attribution - determining who was responsible after a human-caused biological event. The authors keep it strictly separate from detection, which only establishes that engineering happened.

Microbial forensics - reading the properties of the agent itself for clues about who made or released it. The field helped the FBI name a suspected lab of origin in the 2001 anthrax case.

Methodological signature - the pattern left by a genetic engineer's own choices, which genes or

features, from where, incorporated how, shaped by their training, experience, habits and resources. It is what the machine learning models are picking up.

Plasmid - the kind of engineered DNA sequence used in the attribution studies, submitted by labs to the Addgene repository.

Addgene - the largest repository of its kind for these sequences, with 70,000 submissions from labs in 37 countries.

Gain of function - the 2011 influenza experiments that raised concern because adapting a highly virulent avian strain to be transmissible between mammals meant a laboratory escape could lead to a pandemic.

Security dilemma - the problem, cited to Jervis, that a state's purely defensive activity can still frighten other states into an arms race. The authors argue attribution forensics is less likely to do this than other biodefense work.

Wild-type - an unengineered, naturally occurring genome. The authors use it for their false-positive worry: misused attribution could name the "engineer" of a sequence nobody engineered.

Numbers in this reading

- 183 states are party to the Biological Weapons Convention, which categorically bans development

and production of biological weapons; the authors note multiple states have been alleged to have violated it.

- 751 cases of food poisoning caused by the Rajneeshee cult's use of Salmonella in Oregon in 1984.
- 2007: the UK Foot and Mouth disease outbreak, attributed to a leaking pipe at the Institute of Animal Health at Pirbright.
- 1977: the influenza pandemic caused by a strain closely related to ones isolated in the 1950s, which suggests an anthropogenic origin.
- 2011: the "gain of function" influenza experiments that raised the laboratory-escape concern.
- 2001: the anthrax attack where microbial forensics helped the FBI identify a suspected lab of origin and a suspected perpetrator.
- 48%: the accuracy of the first demonstration, in late 2018, that machine learning could read methodological signatures.
- 70,000 submissions from labs in 37 countries: the size of the Addgene repository used by Alley and colleagues.
- 70%: the accuracy Alley and colleagues reached when distinguishing between over 1000 labs.

What the author is unsure about

The authors are explicit that their case is a belief about potential, not a demonstrated capability: genetic engineering forensics is at an early stage and there is a long way to go from published proof-of-principle studies to a robust forensic capability. They state flatly that 70% accuracy is far from a smoking gun, that it may improve but the performance ceiling is not known, and that this should be seen as one tool in the attribution toolkit rather than a standalone silver bullet. They name a key uncertainty: the technique has only been developed and tested on genetic engineers operating in the clear, people who publish their sequences and make no attempt to conceal authorship. A sophisticated adversary might obfuscate or misdirect, including by adopting someone else's methodological signature to deflect attribution or confuse the analysis. They say forecasting how a contest between forensics and counter-forensics would play out is difficult, and that one side or the other may have an intrinsic advantage; their fallback claim is only that evasion imposes a cost, a design constraint and a residual risk of discovery, each a disincentive. They also list hard boundaries: attribution follows detection and is not a substitute for it, inability to attribute does not rule out engineering, improper use could "attribute" engineering that never happened, the method does not apply to releases of non-engineered agents, and the designer identified may not be the actor who misused the organism. Finally, they say the deterrence

value is sensitive to political context: it offers little against actors who intend to claim responsibility rather than conceal it, or who plan to escape consequences through disinformation or other political means.

For the discussion

- The page asks how attribution would work: the answer here is that engineers have habits, and the diversity of design options that makes misuse easier is the same diversity that leaves a traceable methodological signature.
- Worth raising as a tension: the authors call attribution forensics unusually safe to invest in because it has little offensive use and so is less likely to set off a security dilemma, which is a political argument for a technical programme, and worth testing against the 70% number they themselves call far from a smoking gun.
- Worth asking: if a mistaken attribution could lead one state to believe wrongly that it is the victim of a biological attack, what standard of proof should sit between a model's output and any diplomatic or criminal consequence?

2.4-4 · CORE · COURSE TIME 10 MIN

Project BEACON: Using OSINT methods in BWC verification

Neha Suresh · 2024

Read the original paper → · the course assigns: whole excerpt as published on the BlueDot blog

> **Neha Suresh argues that because the Biological Weapons Convention has never had verification provisions, the practical way forward is open-source intelligence, the methods UNSCOM improvised in Iraq now available to anyone with an internet connection, and she proposes a student project to build that skill and test what happens after a detection.**

Average summary

This is a student project write-up on the BlueDot blog, not a paper. Neha Suresh submitted it; it won the **"Build a Thing"** prize in BlueDot's Pandemics course in May 2024, after four weeks of work, and the published text is an excerpt from the final project.

The problem she sets up: genome editing and artificial intelligence are converging, and together with the ability to genetically modify pathogens that creates a threat landscape verification methods struggle to keep pace with. Unlike the major chemical and nuclear weapons treaties, the BWC has lacked verification provisions since its inception, and global discussions on it have been halted since 2001. Verification is exceptionally difficult because

biotechnology is dual-use: telling legitimate research from weapons development is hard.

She notes one opening. At the Ninth BWC Review Conference in late 2022, 185 countries agreed to establish a working group to revisit compliance and verification. At a follow-up meeting in December 2023, participants acknowledged the potential of open source verification tools and microbial forensics.

Her case study is Iraq. The Iraqi biological weapons programme ran from the late 1980s to the early 2000s and violated the BWC after it had come into force. After the 1991 Gulf War, the United Nations Special Commission, **UNSCOM**, was created. Biological work was the hardest thing to verify: agents can be destroyed quickly, and facilities can hide in plain sight disguised as pharmaceutical companies. Al-Hakam was declared as a civilian plant making animal feed proteins, but its remote location and high security raised suspicions. UNSCOM analysed equipment and material purchase records to build a watchlist, used publicly available import and export data and commercial records, looked for red flags in Iraqi scientists' publication patterns, and used historical satellite images to spot infrastructural changes. Al-Hakam turned out to be one of Iraq's primary biological weapons production facilities.

That, she argues, is where OSINT came from. Open source intelligence means leveraging publicly

available information: commercial satellite imagery, social media, transport flows, publications about commercial enterprises. Unlike traditional intelligence gathering, which needs on-site equipment and clearance, it can be done by anyone with an internet connection and the right skillset.

Applied to the BWC, OSINT offers continuous monitoring without physical access, which matters because on-site inspections are politically sensitive and resource-intensive. She lists seven project ideas, from research publication and patent analysis to supply chain tracking, health data integration, social media clusters, economic monitoring and satellite imagery, plus two complementary policies.

Her proposal, Project BEACON, would bring student scientists together to study historical BWC violations, build OSINT analytics skills, and write policy recommendations. She is clear OSINT does not replace a negotiated verification regime.

Why it's assigned

Day 2 is the Delay layer, and this page is its Deter sub-layer, where the course asks how to reduce the intent to build a biological weapon. The course page names open source intelligence as one of its three deep-dive topics and asks how OSINT groups build credibility. This piece answers that from the ground up: it shows where the methods came from, what they

can and cannot see, and it is itself a worked example of a course participant turning the question into a project.

In depth

What this document is

A project submitted by Neha Suresh, which won the "Build a Thing" prize in BlueDot's Pandemics course in May 2024. Participants worked on these projects for four weeks, and the published text is an excerpt from the final project. Read it as a proposal and an argument, not as research findings.

The gap: a treaty with no verification provisions

Suresh opens on two pressures at once. Recent years have seen a convergence of genome editing and artificial intelligence, and that, together with the ability to genetically modify pathogens, creates a constantly evolving landscape of potential threats that verification methods struggle to keep pace with.

Then the structural fact. Unlike major chemical and nuclear weapons treaties, the BWC has lacked verification provisions since its inception, and global discussions on this critical aspect have been halted since 2001. Her explanation for why it is so hard: biotechnology is **dual-use**, which makes distinguishing legitimate research from weapons development challenging.

She does report movement. At the Ninth BWC Review Conference in late 2022, 185 countries took what she calls a significant step forward, agreeing to establish a working group to revisit compliance and verification, a decision made against a backdrop of rising international tensions and a potential turning point. The subsequent meeting in December 2023 saw constructive dialogue that moved beyond past impasses, with participants acknowledging the potential of open source verification tools and microbial forensics.

The Iraqi case study, and why it matters

The Iraqi biological weapons programme operated from the late 1980s to the early 2000s. Suresh treats it as the crucial case because it violated the BWC after the treaty had come into force, which gives researchers evidence about how verification can actually be done.

After the 1991 Gulf War came the question of how to verify and dismantle Iraq's weapons of mass destruction programmes. The **United Nations Special Commission**, UNSCOM, was created, and she says it set new standards for international arms control and verification.

The biological part was particularly challenging. Unlike chemical weapons, biological agents can be quickly destroyed, and they can be hidden in plain

sight disguised as legitimate pharmaceutical companies.

Her example is **Al-Hakam**. Iraq initially declared it a civilian facility producing animal feed proteins. Its remote location and high security raised suspicions. Three things UNSCOM did:

- Analysed records of equipment and material acquisitions to create a watchlist for monitoring suspicious purchases that could be linked to bioweapons development, examining publicly available import and export data and commercial records to track dual-use equipment and materials.
- Examined changes in Iraqi scientists' publication patterns for potential red flags, such as sudden shifts in research focus or unexplained gaps in productivity.
- Used historical satellite images to identify infrastructural changes at suspected sites, giving visual evidence of possible covert activities.

Uncovering Al-Hakam was a significant breakthrough in UNSCOM's mission: it was one of Iraq's primary biological weapons production facilities.

What OSINT is, and why the skill has spread

UNSCOM, she argues, pioneered many techniques now commonplace in **Open Source Intelligence**. At its core OSINT is about leveraging publicly available

information to gain insights: data from commercial satellites, social media, transport flows, and publications about commercial enterprises. It is most commonly used in security and business intelligence.

The point she draws out is access. Traditional intelligence gathering needs on-site equipment and clearance. OSINT can be conducted by anyone with an internet connection who cultivates the right skillset, and she names two beginners' guides, Exposing the Invisible and Bellingcat. This democratization of intelligence, she says, has led to some remarkable discoveries by journalists and researchers.

OSINT applied to BWC verification

Traditional on-site inspections to verify the treaty are politically sensitive and resource-intensive. OSINT allows continuous monitoring of potential biological weapons activities without requiring physical access to facilities.

Drawing on diverse sources, scientific literature, patent databases, satellite imagery, social media and public health records, OSINT can paint a comprehensive picture of a nation's biological capabilities and activities. The breadth is the point: it lets analysts identify patterns or anomalies that would not be apparent in any single source.

She then lists seven projects that could be pursued:

- **Research publication analysis**: tools to scan academic journals, preprint servers and conference proceedings for keywords, unusual research patterns, or collaborations that might indicate dual-use biotechnologies.
- **Supply chain tracking**: a database of companies producing or selling relevant materials and equipment, with monitoring of their transactions, shipping records and customer bases for unusual patterns or undeclared capabilities.
- **Health data integration**: aggregating public health reports, hospital admissions data and pharmaceutical sales information to identify unusual disease patterns or drug demands that do not align with known natural outbreaks.
- **Patent analysis**: a system to scan patent databases globally, flagging applications that involve certain techniques or processes.
- **Social media analysis**: looking across platforms for clusters of health-related complaints that could indicate an accidental release of biological agents.
- **Economic analysis**: monitoring financial transactions, company formations and government budget allocations related to biotechnology, flagging unusual patterns that could indicate covert programmes.
- **Satellite imagery**: analysing imagery for structures consistent with potential dual-use

facilities, and tracking changes over time to detect expansions or modifications.

Her argument for this approach is adaptability: as new dual-use technologies arise, OSINT techniques can quickly incorporate them into monitoring frameworks, so verification efforts stay relevant as the science moves.

Two policies to go with it

She proposes complementary policies so OSINT is not left on its own:

- A **Global Biosurveillance Network**, implemented as a global environmental DNA (eDNA) sampling programme with international cooperation and data sharing.
- An **International Biosecurity Rapid Response Network**, a collaborative international network of health agencies, biosecurity units and experts to respond rapidly to biological threats.

Her framing: OSINT complements traditional verification methods by offering continuous and broad spectrum monitoring that can guide more targeted investigations, and with both policies the global biosecurity architecture is strengthened through better regional cooperation and preparedness.

Project BEACON

She quotes Henrietta Williams, Policy Researcher at KCL, who says both UNSCOM and OSINT come down to people: "talented, inventive, meticulous people" finding solutions to very difficult monitoring problems.

Her proposal is to create Project BEACON, bringing together student scientists who would examine historical cases of BWC violations for insights into past challenges and shortcomings in verification; build on that history by using OSINT to develop advanced data analytics skills and new methodologies for detecting potential biological weapons activities; and then craft policy recommendations that complement the technical approaches.

For this project she took on one question: what happens after detection using OSINT. Her method was a simulated tabletop exercise, with the findings in a linked report.

Her closing argument

The prevalence of OSINT verification technologies, she says, demonstrates the widespread capability for nongovernmental monitoring. These tools already track various illicit activities and global transport networks. That presents an opportunity to integrate disparate efforts into a comprehensive global bioweapons tracking service, and so to build a powerful complement to traditional high-level diplomatic initiatives.

Her caveat is clear and she states it herself: OSINT work is not a replacement for internationally negotiated verification regimes, though it offers unique advantages in biosecurity monitoring. The ambition she claims for the integration is bridging the gap between grassroots technological innovations and formal interstate cooperation.

Words to know

Verification provisions - the parts of a treaty that let other states check compliance. Suresh's central fact is that the BWC has never had them, unlike the major chemical and nuclear treaties.

Dual-use - research or equipment with both legitimate and weapons applications. It is her explanation for why BWC verification is exceptionally difficult.

UNSCOM - the United Nations Special Commission, created after the 1991 Gulf War to verify and dismantle Iraq's weapons of mass destruction programmes, and the source of many techniques now standard in OSINT.

Al-Hakam - the Iraqi facility declared as a civilian animal feed protein plant that turned out to be one of Iraq's primary biological weapons production sites.

OSINT, open source intelligence - gaining insight from publicly available information: commercial

satellite imagery, social media, transport flows, company publications. Its defining feature in this piece is that it needs no clearance or on-site access.

Watchlist - the purchasing list UNSCOM built from acquisition records so suspicious buys of dual-use equipment could be monitored.

eDNA, environmental DNA - genetic material collected from the environment rather than from an organism directly. It is the basis of her proposed Global Biosurveillance Network.

Tabletop exercise - a simulated run-through on paper, with no real-world action, which is how she tested what happens after an OSINT detection.

Numbers in this reading

- Four weeks: how long participants worked on these projects in the Pandemics course.
- May 2024: when the project won the "Build a Thing" prize.
- 2001: the year global discussions on BWC verification were halted.
- Ninth BWC Review Conference, late 2022: where 185 countries agreed to establish a working group on compliance and verification.
- 185 countries: the number she gives as agreeing to that step.

- December 2023: the follow-up meeting where participants acknowledged the potential of open source verification tools and microbial forensics.
- Late 1980s to early 2000s: the span of the Iraqi biological weapons programme.
- 1991: the Gulf War, after which UNSCOM was created.
- Seven OSINT project ideas and two complementary policies.

What the author is unsure about

Suresh does not present findings, so there is little to be uncertain about in the usual sense; most of the piece is a proposal and a set of ideas she thinks are worth pursuing. The limit she states plainly is the important one: OSINT work is not a replacement for internationally negotiated verification regimes, only a complement to high-level diplomatic initiatives. Her seven project ideas are described as things that could be pursued, not things that exist, and her two policies are given only as an idea plus a one-line implementation note. Her own contribution to the question of what happens after a detection rests on a simulated tabletop exercise, with the findings in a separate linked report that is not part of this excerpt. She does not claim that any OSINT signal would be strong enough to establish a treaty violation, and she does not discuss how a false positive would be handled.

For the discussion

- The course page asks how OSINT groups build credibility; her answer runs through UNSCOM, which built its case on purchase records, publication patterns and historical satellite imagery rather than on access, and that pedigree is the credibility.
- Worth raising as a tension: she says OSINT gives continuous monitoring without the political cost of on-site inspection, which is exactly why it is attractive, and also why it will never be the thing a state accepts as proof.
- Worth asking: of her seven project ideas, the health data and social media ones would mean watching ordinary people's hospital visits and posts, so who sets the limits on a global monitoring capability that needs no clearance to run?

2.4-5 · FURTHER · COURSE TIME 20 MIN

Adherence to and Compliance with Arms Control, Nonproliferation, and Disarmament Agreements and Commitments (2025)

U.S. Department of State · 2025

Read the original paper → · the course assigns: the biological weapons sections in full (the BWC entry under U.S. compliance, and all of Part V with its China, DPRK, Iran and Russia assessments), plus the report's introduction (purpose,

scope, the compliance vocabulary, and who prepares it) and its table of contents and section headings

> **The U.S. government's annual verdict on who is keeping the arms control treaties: on the biological side it finds Russia and North Korea in violation of the Biological Weapons Convention, has open concerns about China and Iran, and says all U.S. activity in 2024 was consistent with the treaty.**

Average summary

From the report's biological weapons sections, its introduction and its section headings.

This is the U.S. Department of State's annual compliance report, April 2025, 50 pages, transmitted under Section 403 of the Arms Control and Disarmament Act, as amended (22 U.S.C. § 2593a), which requires the President to report every year. It assesses U.S. compliance and other states' compliance and adherence, primarily for activities from January 1, 2024 through December 31, 2024, and says some findings may be addressed only in its SECRET or TOP SECRET annexes. It covers many agreements, not only biological ones: New START, the Nuclear Non-Proliferation Treaty, the Threshold Test Ban Treaty and nuclear testing moratoria, the Missile Technology Control Regime, the Chemical Weapons Convention, and conventional agreements including

the Vienna Document and the Treaty on Conventional Armed Forces in Europe. The biological part follows.

The vocabulary is deliberate. **Violation**, also called non-compliance, is used only for legally binding agreements; **adherence** is for political commitments. Where the U.S. believes a state may not have complied but factual or legal uncertainty remains and no determination has been reached, the report says **compliance concern** instead. Findings rest on intelligence collection, cooperative verification measures, open source information, diplomatic means, and whatever the country itself provides. The State Department is the lead agency, its Bureau of Arms Control, Deterrence, and Stability leads, and the report is prepared with the concurrence of the Director of National Intelligence.

On the Biological Weapons Convention: it opened for signature in 1972 and entered into force in 1975. As of the end of 2024 there were 186 States Parties and four signatory States for which the treaty is not yet in force. The annual data exchange known as the Confidence-Building Measures was established in 1987, modified and expanded in 1991 and streamlined in 2011. Submitting them is a politically binding commitment, not all States Parties routinely do it, and submission rates have risen for several years, climbing over 100 annual submissions in 2023 and over 110 submissions 2024.

The U.S. finds all its own 2024 activities consistent with the BWC. Its four country assessments:

- **Russia**: assessed to maintain an offensive biological warfare program, in violation of Articles I and II. The U.S. assesses the Soviet program was absorbed, not dismantled. Russia's annual submissions report nothing new to declare, and renovations continued in 2024 at the Ministry of Defense-related 48th Central Scientific Research Institute in Sergiyev Posad.
- **DPRK**: assessed to have a dedicated, national-level offensive program, in violation of Articles I and II, with capabilities since at least the 1960s. It filed a null report in 1990 and none since.
- **China**: concerns under Article I over toxin research for military purposes, and insufficient information to determine whether China eliminated the offensive program the U.S. assesses it ran from the 1950s to at least the late 1980s.
- **Iran**: concerns under Article I; assessed not to have abandoned its intention to research biological agents for offensive purposes. Only two submissions since 2011, both incomplete.

Why it's assigned

Day 2 is the Delay layer and this page is its Deter sub-layer, where the course asks what makes a state decide not to build a biological weapon. This is what

deterrence looks like in its current paperwork form: the treaty exists, it has no verification mechanism, so one government publishes its own assessment of who is breaking it, based partly on open source information. Reading it beside the course page's point about the BWC's missing verification machinery shows exactly what fills the gap, and how much of the judgement is national rather than international.

Words to know

Adherence - the report's word for whether a state is keeping a politically binding commitment, as opposed to a legal obligation. It never calls a political lapse a violation.

Compliance concern - the middle category. The U.S. has reason to believe a state may not have complied, but factual or legal uncertainty remains, so no determination has been reached. Most of the biological section lives in this category rather than in findings of violation.

Confidence-Building Measures (CBMs) - the annual data exchange BWC States Parties established in 1987, modified and expanded in 1991 and streamlined in 2011. Submitting is politically binding, not legally, which is why non-submission is an adherence problem and not a violation.

Article I and Article II - the two BWC obligations the findings turn on. Article I bars developing,

producing, stockpiling, acquiring or retaining biological agents or toxins of types and in quantities with no justification for prophylactic, protective or other peaceful purposes, and the weapons or delivery means designed to use them for hostile purposes. Article II requires a state to destroy, or divert to peaceful purposes, the items specified in Article I.

Article X - the cooperation provision. The U.S. cites it for its own transparency about scientific and technical engagements and laboratory support that help other countries detect, prevent, prepare for and respond to disease outbreaks.

State biological warfare program (U.S. definition) - a leadership-approved effort intended to acquire, develop, modify, produce or retain biological warfare agents for use or potential use as a weapon. The report says such a program would probably include research, acquisition, development, modification, production, retention or testing of agents or dispersal devices; facilities producing or intended to produce them; training, doctrine or plans for use; and use or attempted use.

Bioregulators - the subject of BW-relevant projects the report says open source reports attribute to Iranian military-associated universities and affiliated research centers.

Numbers in this reading

- April 2025: the report's date; the document runs to 50 pages.
- 22 U.S.C. § 2593a, Section 403 of the Arms Control and Disarmament Act: the law requiring this annual report.
- January 1, 2024 through December 31, 2024: the activities the report primarily reflects.
- 1972 and 1975: when the BWC opened for signature and entered into force.
- 186 States Parties to the BWC as of the end of 2024, plus four signatory States for which the treaty is not yet in force.
- 1987, 1991, 2011: the CBM data exchange established, then modified and expanded, then streamlined.
- Over 100 annual CBM submissions in 2023, and over 110 submissions 2024.
- February 2025: Comoros acceded to the BWC, noted as outside the reporting period.
- January 9, 2018 and March 5, 2018: the Palestinians' purported instrument of accession, and the U.S. letter detailing its objections to the purported accession of the "State of Palestine."
- China: a State Party in 1984, CBMs submitted every year since 1989, an assessed offensive program from the 1950s to at least the late 1980s, production at two facilities in Beijing and Lingbao before China signed in 1972, and compliance

questions reported since the 1993 Report. The report says China reportedly weaponized ricin, botulinum toxins, and the causative agents of anthrax, cholera, plague and tularemia.

- November 2024: the U.S. raised concerns about China's interest in pharmaceutical based agents and toxins in its National Statement at the OPCW's 29th Conference of the States Parties.
- DPRK: a State Party in 1987, BW capabilities assessed since at least the 1960s, a null CBM report in 1990 and no report since.
- Iran: a State Party in 1973, compliance addressed every year since the 1993 Report, and only two CBM submissions since 2011, in 2016 and 2021, both incomplete.
- Russia: the Soviet Union became a State Party in 1975; Russia's compliance was first addressed in the 1993 Report and Soviet noncompliance in the January 1984 Report to Congress; Yeltsin confirmed the Soviet program on April 11, 1992; key military biological facilities remain inaccessible and not transparent nearly 30 years after his decree; multi-million dollar renovations at the 48th Central Scientific Research Institute in Sergiyev Posad, formerly the Soviet Center of Virology in Zagorsk, continued in 2024.
- May 1, 2024: the most recent sanctions. During the reporting period the U.S. designated three Russian government entities associated with its chemical

and biological weapons programs and four Russian companies supporting them, and under a separate WMD non-proliferation authority three entities and two individuals involved in procuring items for military institutes.

What the author is unsure about

The report is unusually open about the edges of what it knows, and its hedging words are doing real work. It says there can be legal or factual uncertainty about whether a violation occurred, and it keeps a separate category, compliance concerns, for cases where no determination has been reached. It states that some compliance and adherence findings may be addressed only in the SECRET or TOP SECRET annexes, so the unclassified document is not the whole assessment. On China it says plainly that the United States does not have sufficient information to determine whether China has fulfilled its Article II obligation, and that there is no available information demonstrating China took those steps; it also uses "probably" for China's capabilities, including that China probably is capable of using publicly available artificial intelligence and machine learning tools to advance efforts related to BW applications. On Iran it says it is unknown whether Iran's leadership has set a directive to maintain a biological weapons option, and that it remains difficult for the United States to differentiate some of Iran's public health research,

pharmaceutical production and biodefense activities, which are allowed, from activities that would be prohibited. On the DPRK it uses "probably" for weaponization capability. It also notes that lack of reporting in Iran's CBMs contributes to uncertainty, and that consistent and complete Iranian reports may help provide clarity. One structural limit it states itself: it does not include findings that relate solely to potential future compliance. And it says the issues covered occurred during the previous administration, with the current one continuing to assess developments going forward.

For the discussion

- With no verification mechanism in the treaty, the compliance judgement is national: this report says the State Department prepares it with the concurrence of the Director of National Intelligence, which is a very different thing from an international inspectorate's finding.
- Worth raising as a tension: the report's own evidence base includes open source information alongside intelligence collection and diplomacy, which is the same material the open-source-intelligence reading on this page wants to build a monitoring capability from.
- Worth asking: CBM submission is only politically binding, so Russia can file "nothing new to declare" every year and the DPRK can file nothing at all

since 1990 without that itself being a violation, and over 110 submissions 2024 is called a steady improvement; what would make the reporting worth anything?

2.5-1 · CORE · COURSE TIME 30 MIN

Confronting risks of mirror life

Katarzyna P. Adamala and 37 co-authors · 2024

Read the original paper → · the course assigns: the abstract, author list, acknowledgements, references and metadata only; the body text is paywalled

> **Thirty-eight scientists, including several who helped build the field of synthetic biology, used a Science policy forum to say that building life from mirror-image molecules carries unprecedented and largely overlooked risks, and that because the capability is probably at least a decade away there is still time to decide against it.**

Average summary

Read the provenance line first, because it limits everything below. This summary comes from what was readable on Science's page: the abstract, the subtitle, the full author list, the acknowledgements, all 15 references and the article metadata. The body of the policy forum is behind the paywall and was not available, so the argument's structure, its examples,

its policy recommendations and any caveats in the main text are **missing here**. The course also links a supplementary-material PDF, which opens but could not be read as text. The free route to the actual analysis is reading 2.5-2, the same working group's roughly 300-page technical report on mirror bacteria, which is reference 1 of this paper and is downloaded in full in this pack.

What the abstract says is this. All known life is **homochiral**: DNA and RNA are built from right-handed nucleotides, proteins from left-handed amino acids. Driven by curiosity and by plausible applications, some researchers had begun working toward lifeforms composed entirely of mirror-image biological molecules. Such mirror organisms would be a radical departure from known life, and the authors say their creation warrants careful consideration.

Their timing judgement is the practical core: the capability to create mirror life is likely at least **a decade** away and would require large investments and major technical advances. That gap is the opportunity, because the risks can be considered and pre-empted before anyone can realise them.

The paper rests on an in-depth analysis, carried out for the technical report, of three things: the current technical barriers, how those barriers might be eroded by technological progress, and what the authors call unprecedented and largely overlooked

risks. The published subtitle states the ask plainly, that broad discussion is needed to chart a path forward, and the abstract names who should be in it: the global research community, policy-makers, research funders, industry, civil society and the public.

The author list is itself part of the argument, running to 38 names and including George M. Church, Drew Endy, Kevin M. Esvelt, John I. Glass, Jack W. Szostak, David A. Relman, Thomas V. Inglesby, J. Craig Venter and Gregory Winter. Glass and Szostak are the corresponding authors and chaired the working group the article emerged from. The acknowledgements record that the **Mirror Biology Dialogues Fund**, set up to support discussion and research on this topic with contributions from the Alfred P. Sloan Foundation, the David and Lucile Packard Foundation, the Gordon and Betty Moore Foundation, Open Philanthropy and Patrick Collison, supported the working group, and that Open Philanthropy funded several authors' contributions. The views are stated as those of the individuals, not their institutions.

Why it's assigned

Page 2.5 is the Deny option on regulating discovery, and it says the case for banning mirror life research is unusually clear while the argument over other areas continues. This is the paper that made that case publicly, which is why it opens the page: a large group of eminent scientists, several of whom spent careers

arguing for open and permissive biology, concluding that one line of research should not be pursued at all. It is the clearest available example of the course's first category of intervention, regulating what gets researched, decided by the research community itself rather than by a regulator.

Words to know

- **Homochirality**: the fact that all known life uses only one handedness of its building blocks, right-handed nucleotides and left-handed amino acids. It is the shared property that mirror life would break.
- **Mirror organism, or mirror bacteria**: a hypothetical lifeform built entirely from mirror-image versions of normal biological molecules. The concern is that it would be unrecognisable to defences that evolved against ordinary life.
- **Policy Forum**: Science's section for argued pieces on research policy rather than new experimental results. Knowing the genre matters, because this is a considered recommendation from named scientists, not a finding.
- **Working group**: the self-organised group of researchers, chaired here by Glass and Szostak, that produced both this paper and the technical report. It is the body that actually made the judgement.

Numbers in this reading

- 38 authors in total, with 2 corresponding authors, John I. Glass and Jack W. Szostak.
- At least a decade: the authors' estimate of how far away the capability to create mirror life is, requiring large investments and major technical advances.
- 15 references, the first being the working group's own technical report on mirror bacteria in the Stanford Digital Repository.
- Science Vol. 386, Issue 6728, pages 1351 to 1353, published online 12 December 2024 and in print 20 December 2024.
- 92 citations in Crossref and 70 in Web of Science at the time the page was read, which is one measure of how quickly the argument travelled.

What the author is unsure about

The readable part is modest about what is known. The decade estimate is hedged twice over, as likely and as at least, and it is a forecast about technology rather than a measurement. The authors describe the risks as largely overlooked, which is a claim about the state of the discussion rather than a quantified result, and they call for broad discussion to chart a path forward rather than presenting a finished policy. The acknowledgements note that a list of competing interests for all authors sits in the supplementary materials, which could not be read, and that individual

affiliations are there too. Because the body text is paywalled, any doubts, dissents or limits the authors stated in the article itself are not represented in this summary, and that absence should not be read as confidence.

For the discussion

- The strongest point to make from this paper is who signed it: a group that includes several of the people who built modern synthetic biology, arguing against pursuing a line of research, which is a very different kind of evidence from a regulator saying no.
- Their reasoning is about timing rather than imminence, since they judge the capability at least a decade away and treat that gap as the reason to decide now, which is the opposite of how most technology governance actually happens.
- A tension worth raising: the working group that reached this conclusion was convened and funded by philanthropies set up to support exactly this discussion, which is either a model for how fields should police themselves or a small group deciding what the rest of biology may not do, and the group could ask which it looks like.

2.5-2 · FURTHER

Technical Report on Mirror Bacteria: Feasibility and Risks

Katarzyna P. Adamala, Deepa Agashe, Damon J. Binder et al. · 2024

Read the original paper → · the course assigns: the Summary and the Introduction

> **A working group of scientists from synthetic biology, immunology and ecology concludes that bacteria built entirely from mirror-image molecules could become feasible within decades, could evade most immune systems and most predators, and if released could do irreversible harm that no countermeasure would undo.**

Average summary

This is the December 2024 technical report that accompanies the Science article "Confronting risks of mirror life". The course assigns the Summary and the Introduction, which is Chapter 1 and also an overview of the whole eight-chapter report.

The starting fact is **chirality**. Many of life's key molecules come in two mirror-image forms that cannot be superimposed, and all known life uses only one of each: almost all amino acids are the left-handed form, the ribose in RNA is always the right-handed form. A cell in which every molecule was swapped for its mirror image would work normally on

its own. What changes is every interaction with natural life, because recognition between molecules depends on shape.

Three consequences drive the report. First, immunity. Immune systems detect bacteria by binding conserved chiral molecules, so pattern recognition would likely fail; so would many of the mechanisms that kill bacteria once detected; and because mirror proteins resist the processing step, antibody production would likely be impaired too. The authors say a lethal outcome from a systemic mirror bacterial infection "appears highly plausible" absent an effective immune response, while noting the clinical picture is unclear. Treatment would be thin: most antibiotics act stereospecifically, leaving only a few achiral or racemic ones, and there would be no existing vaccine.

Second, other species and the environment. Vertebrates would likely share human defects; invertebrate immunity is more variable and less well characterised. Almost all known plant immune receptors recognise chiral targets. Mirror bacteria would be completely resistant to all bacteriophages, largely indigestible to predators, and largely resistant to the antimicrobials other microbes use, so a growing population would not be met by growing predation. Key crops could probably be engineered to resist; wild plants and ecosystems could not.

Third, feasibility and control. Making a living cell from non-living parts is not yet possible, but the authors think it could plausibly be done within as little as a decade given sufficient resources, and a mirror cell could follow by the same methods. The first one would likely be fragile, and then easy to make robust. A biocontained mirror bacterium is probably achievable; one that a moderately skilled actor with access to mirror DNA could not un-contain does not appear feasible.

The report is explicit about its own limits: conclusions are "necessarily tentative and uncertain", written at one point in time, and the authors hope others will find what they missed.

Why it's assigned

Day 2 is the Prevent layer, and page 2.5 is Deny: deciding what should not be researched, written down or made available. The course page calls the case for banning mirror life research "unusually clear", and this is the document that case rests on. Its Rationale for Public Release is also the unit's cleanest worked example of the publication dilemma: the authors argue in print why they chose to describe a danger rather than keep it to themselves.

Words to know

Chiral Having a mirror image that cannot be laid on top of the original, like a left and right hand. The

whole report follows from life using only one version of each chiral molecule.

Enantiomers The two mirror forms of such a molecule. The report notes they have nearly identical physical properties but rotate polarised light in opposite directions.

Achiral The same as its own mirror image, like a sphere. It matters because achiral nutrients, such as glycerol, would feed a mirror bacterium without any engineering.

Mirror bacterium A bacterium in which the proteins, nucleic acids and metabolites are all replaced by their mirror images. It cannot evolve from existing life, only be built.

Pattern recognition receptors, and the patterns they detect The sensors of the innate immune system and the conserved bacterial molecules they bind. Because those molecules are almost all chiral, this is where detection would fail first.

Biocontainment Engineering an organism so it cannot survive outside the lab, for example by making it depend on a nutrient absent from nature. The report's point is that containment can be reversed by someone who wants to reverse it.

Autotroph An organism that makes its own food from light or chemistry rather than eating others. A mirror

autotroph is the report's worst ecological case, because it needs no chiral nutrients at all.

Numbers in this reading

Eight chapters: a general introduction, then the creation of mirror bacteria, further engineering with its biosafety and biosecurity implications, then human health, medical countermeasures, animals, plants, and ecological consequences.

As little as a decade: how soon the authors think a living cell could plausibly be made from non-living precursors, given sufficient resources. A mirror cell would take longer, and they say estimating timelines is difficult.

One to a few decades: the window in which, they write, scientific progress could lead to mirror bacteria were it not for safety and security concerns.

Half a million base pairs: the smallest known bacterial genome size, in engineered reduced-genome strains, which they cite as a measure of how feasible building a cell from scratch is becoming.

$4 million: the US National Science Foundation award made in 2019 to investigators aiming to design, construct and deploy synthetic mirror cells. They also name grants from the National Natural Science Foundation of China and the European Commission's

ERA-NET MirrorBio consortium, supported by the German Federal Ministry of Education and Research.

Three ways immunity could be compromised: initial detection, the mechanisms that kill bacteria, and the processing step that activates adaptive immunity and antibodies.

About 10 to the minus 11 joules per mole: the theoretical energy difference between enantiomers from the weak nuclear force, which the report calls about 10 to the 14 times smaller than thermal noise at room temperature. The most precise measurement to date found no such difference at a sensitivity of about 10 to the minus 8 joules per mole. This is why they treat a mirror cell as a perfect mirror.

What the author is unsure about

The authors say plainly that many of their conclusions are necessarily tentative and uncertain, that any risk assessment must extrapolate from limited information and a handful of well-studied examples, and that a single report written at a single point in time cannot be considered definite.

They hedge the big claims in their own words: it "appears plausible, even likely" that robust mirror bacteria could spread unchecked; a lethal outcome from systemic infection "appears highly plausible"; the exact clinical presentation is unclear.

Specific unknowns they name: whether the first mirror cell would be fragile is plausible but not certain; whether mirror bacteria could spread through plant vasculature is unclear and may differ between plant types; whether a nematode could survive infection may or may not follow from its reliance on avoidance rather than recognition; how effectively predators could kill mirror bacteria even if they ingested them is not clear.

On timelines they say estimating is difficult, and that normal-chirality synthetic cells are likely still years away.

One asymmetry is stated as settled rather than open: they argue it is almost certain that a correctly built mirror organism would mirror natural life exactly, because no known physical process violates chiral symmetry at molecular scales to any degree that could matter in biology.

They also address the obvious reassurance and reject it: mirror life is absent from nature because chirality inversion is all or nothing and evolution proceeds stepwise, so its absence is no evidence that it would be harmless.

For the discussion

The report's Rationale for Public Release is the most interesting governance argument in the unit: the authors say disclosure is necessary precisely because

the default path of well-meaning science would arrive here anyway, which flips the usual presumption that naming a danger spreads it.

Their guardrail proposal is narrow and testable, so it is worth raising: if well-intentioned scientists simply decline to build particular components, and critical materials such as mirror nucleic acids are monitored, the technical barrier stays high for everyone else.

A tension worth naming: the same report says robust biocontainment that a moderately skilled actor could not undo does not appear feasible, which means the usual lab-safety answer does not apply, and the decision has to be made before anything is built.

2.5-3 · FURTHER · COURSE TIME 90 MIN

James Smith on why he quit everything to work on a biothreat nobody had heard of

80,000 Hours Podcast (host: Luisa Rodriguez) · 2026

Read the original paper → · the course assigns: the episode page, not a full transcript

> **James Smith, who dropped his other work within two weeks of hearing about mirror bacteria, argues the threat is catastrophic, close enough to matter, almost benefit-free, and therefore one of the most preventable catastrophes going.**

Average summary

What was saved here is the **episode page**, not a transcript: the show's own blurb, a note from Rob Wiblin about infohazards, a section the show calls the episode in a nutshell, a run of quoted highlights from the conversation, and the reading list. So the quotations are real but the connecting summary is the show's, not the guest's. Episode #233, recorded on November 5-6, 2025, with host Luisa Rodriguez.

Smith is director of the Mirror Biology Dialogues Fund and a coauthor of the December 2024 Science paper, written with 38 leading scientists including two Nobel Prize winners and several people who had previously wanted to create mirror organisms. He was sceptical at first, as he says most people are. Two things convinced him: the best immunologists, ecologists and synthetic biologists had been asked to find the hole in the analysis and could not; and imagining the world it describes, where any exposure to the outdoors could be fatal.

The mechanism he explains is recognition. Immune defence depends on molecules binding by shape, and he uses a hand-in-glove image: a left hand will not seat properly in a right-handed glove. Pattern recognition receptors would not bind a mirror bacterium's reversed molecules, and because the antibody response depends on that first step, the whole chain can break. His grim analogy is MyD88

deficiency, a human disorder in which that receptor repertoire fails, where something like 30% of patients die before the age of two.

The asymmetry he keeps returning to: to cause harm a mirror bacterium only has to grow, and there are enough achiral nutrients in human blood to let it, while the immune system has to recognise it first, and that is the step that fails. Viruses that infect bacteria outnumber them 10 to 1 and would be useless here, because a mirror cell cannot read a normal virus's genetic code.

On timing, he gives a spread rather than a date: synthetic biologists mostly say 10 to 30 years, roughly $500 million to $1 billion might be enough to do it much faster, George Church thinks it could happen in the next couple of years, and he says it is hard to be confident in any estimate.

His case for tractability is that there is almost nothing on the other side of the ledger, unlike AI or nuclear technology. Three things would need to be true within two to five years: a strong scientific norm against the work, regulation of the precursor technologies, and governments treating it the way they treat terrorists seeking nuclear weapons. He notes fewer than 10 people work on this full time.

Why it's assigned

Day 2 is the Prevent layer and this page is Deny, the layer that asks what should not be built or published. This episode is the clearest statement of the strongest version of that argument: a technology with catastrophic downside and, on his account, almost no upside, where the realistic lever is governing the precursor technologies before anyone can take the final step. It also answers the course's guiding question about whether anyone still pursues mirror life, including the awkward detail that some of the paper's authors used to.

Words to know

Mirror bacteria Bacteria built from the mirror images of life's normal molecules. Smith's point is that this small-sounding switch breaks the recognition that all immunity rests on.

Chirality The handedness of a molecule. All life on Earth uses one handedness, which is why the page says the DNA of all life twists the same way.

Achiral Having no handedness, so it looks the same mirrored. Acetate, acetoacetate, glycerol, succinate and pyruvate are the examples he gives from human blood, and they are what a mirror bacterium could eat without any engineering.

Pattern recognition receptors The innate immune system's first-line sensors, which bind molecules on

invading bacteria such as bacterial DNA or flagellin, the protein in a bacterial tail.

Conjugate vaccine A vaccine that attaches a target the body ignores to something it reacts to, tricking it into making antibodies. He notes this has been done for cocaine and might work for mirror molecules.

Precursor or enabling technologies The capabilities developed on the way to a goal. Governing these, rather than banning the end goal, is the policy ask he thinks is hardest and most important.

Mirror ribosome The mirror version of the cell's protein-making machine, made of 54 proteins and 3 RNA molecules. It is the main technical gap between now and mirror life.

Numbers in this reading

38 leading scientists coauthored the December 2024 Science paper, including two Nobel Prize winners and several who had previously wanted to create mirror organisms.

About 30% of patients with MyD88 deficiency die before the age of two, his analogy for what losing one part of innate immunity does.

10 to 1: how much the viruses that infect bacteria outnumber bacteria, a control that would not work on

mirror bacteria at all. He says we can be "very confident" about that one.

10 to 30 years: what most synthetic biologists who work on this say, before pricing in AI acceleration. George Church thinks it could be the next couple of years.

$500 million to $1 billion: what people estimate it would cost to make mirror life from where we are now, which could make it happen much more quickly.

54 proteins and 3 RNA molecules make up the ribosome that would have to be mirrored. Scientists can already make mirror DNA polymerases, mirror RNA polymerases, and 70% of the ribosome by weight.

Fewer than 10 people work full time on mirror life risks and governance. He says you could probably become your country's policy expert on it within a few weeks or months.

About 100 scientists across all continents signed a statement from a recent summit on responsible biotechnology saying this should not be pursued.

Two to five years: his window for the three conditions that would make him comfortable the problem was basically solved.

What the author is unsure about

He is open that he was sceptical at first and that most people are, and that his confidence comes from the failure of expert attempts to poke holes rather than from direct evidence.

He separates what he is confident about from what he is not. Immunity to bacterial viruses: "very confident". Plants: much less so, because plant defence leans on physical barriers rather than recognition, mirror bacteria would lack the enzymes to get into the plant's vascular system, and plant immune systems are less well understood. He says there is a reasonable chance of serious harm to plants, including crops, and that we will struggle to rule it out.

On nutrition he flags a real constraint and then removes it: a plain mirror E. coli would probably grow a bit more slowly, but a malicious actor could engineer it to eat common chiral nutrients such as glucose, and he thinks that constraint is hard to rely on.

On timelines he says it is really hard to be confident in any particular estimate, and that is part of why he thinks the discussions must happen now.

He lists open questions rather than answers: where exactly the regulatory line should be drawn, what mix of law, policy, voluntary commitments and monitoring would work, and what further research would clarify the risks without moving anyone closer to the goal.

The page also carries a disclosure note from Rob Wiblin arguing that the risk has been covered in major outlets since December 2024, so the interview adds little to how widely known it is. That is the show's judgement about publishing, not a finding.

For the discussion

The precedents he cites are the most usable thing in the episode: the Environmental Modification Convention of the late 1970s, which the United States and the USSR both signed, banned weather as a weapon before anyone could do it, and the treaty on blinding laser weapons did something similar, so prohibiting a technology in advance is not unheard of.

The countermeasure point is worth raising because it inverts the usual comfort: mRNA and DNA vaccines, the platforms we are best at, would not work at all, because a cell cannot read mirror genetic code. Detection has the same problem, since PCR relies on enzymes that would not act on mirror molecules.

A tension: he says mirror life sits in a tiny corner of the technology tree and giving it up costs almost nothing, while also saying regulating the precursor technologies is the hard ask because people want to work on them in the near term. Both cannot be fully true, and that gap is where the governance fight will happen.

2.5-4 · FURTHER · COURSE TIME 15 MIN

Mirror Biology Dialogues Fund (MBDF) website

Mirror Biology Dialogues Fund · 2026

Read the original paper → · the course assigns: the whole site page

> **This is the website of the small organisation that funds and convenes the global conversation about mirror life, and it doubles as a running scoreboard of which governments, scientific bodies and funders have accepted that the risk is real.**

Average summary

Not an article: a short organisational site. It is worth reading as evidence of how a brand-new risk gets institutionalised, and how fast.

The **Mirror Biology Dialogues Fund**, MBDF, exists to advance and support the global effort to understand and address the risks posed by mirror life. Its tagline is advancing the conversation. Mirror life, as the site defines it, means hypothetical synthetic organisms built from molecules that are the mirror images of their naturally occurring counterparts, which if created could pose catastrophic risks to humans and global ecosystems.

The origin story is two documents. In a December 2024 article in Science, a working group of leading scientists raised concerns about creating mirror life, supported by a comprehensive technical report that gave the first systematic assessment of its feasibility and risks. Since then, the site says, a growing international community has acknowledged that mirror life could pose serious risks: government bodies, multilateral organisations, scientists and research funders. The authors of that article, and many others, are calling for governance to prevent the creation of mirror life, with one conditional attached: absent significant new evidence for reassurance.

What MBDF actually does is convening, not laboratory work. It funds and co-organises international conferences, workshops, public seminars and other initiatives, aiming at constructive dialogue and decision-making between academic, policy, commercial and public stakeholders. Its team and advisors are based in the United States, the United Kingdom, Canada, China, the Netherlands and France, which the site presents as the point rather than an accident. Under its own work it lists three things: supporting the working group of 38 researchers to publish in Science, funding and hosting international events that bring together scientists, policymakers, ethicists, funders, industry leaders and members of the public, and supporting technical work to

understand the key precursor technologies for mirror life and the possibilities for governing them.

The recent news list is the substance for a reader studying the Deny layer. Each line is an institution moving. The UN Scientific Advisory Board and the World Health Organization have issued brief explainer articles on the risks and the need for governance. UNESCO's International Bioethics Committee recommended a precautionary global moratorium on creating mirror organisms. Germany's Central Commission for Biological Safety said mirror organisms could pose serious risks and that a broad scientific and societal debate was necessary. The Bulletin of the Atomic Scientists cited mirror life as a key factor pushing the 2026 Doomsday Clock closer than ever to midnight, and said preventing its creation could help turn the clock back. Two funders, the Alfred P. Sloan Foundation and Renaissance Philanthropy, have stated they will not fund research aimed at creating mirror organisms. There are also two events: an international symposium at the National University of Singapore on 28 September, co-hosted with the Asia Centre for Health Security, SINERGY and SynCTI, and a workshop with Professor Patrick Cai at the University of Manchester on precursor technologies and their potential benefits.

Why it's assigned

Day 2 is the Prevent layer and this page is Deny, which is the layer that depends on institutions actually saying no. This site is the shortest route to seeing who has said it so far, across science academies, UN bodies, a national biosafety commission and two philanthropic funders, and it shows the three levers of the Deny layer in one place: a norm among scientists, governance of precursor technologies, and funders declining to pay.

Words to know

Mirror life Hypothetical synthetic organisms built from mirror-image molecules. The site's own one-line definition, and the thing all of its work is about.

Moratorium An agreed pause on doing something. UNESCO's bioethics committee recommended a precautionary global one, meaning adopted before harm is demonstrated rather than after.

Precursor technologies The capabilities that would have to exist before mirror life could be built. MBDF funds technical work on these and on how they might be governed, which is where the practical regulation would bite.

Multilateral organisations Bodies made up of many governments, such as the UN and the WHO. The site counts their attention as evidence the concern has moved beyond academia.

The Doomsday Clock The Bulletin of the Atomic Scientists' annual symbolic reading of how close humanity is to catastrophe (not from this reading). The site notes the 2026 setting is the closest ever to midnight, with mirror life named as a key factor.

Numbers in this reading

December 2024: the Science article that started this, with its accompanying technical report described as the first systematic assessment of feasibility and risks.

38 researchers in the working group MBDF supported to publish that article.

Six countries where its team and advisors sit: the United States, the United Kingdom, Canada, China, the Netherlands and France.

28 September: the date of the international symposium at the National University of Singapore, co-hosted with the Asia Centre for Health Security, SINERGY and SynCTI. The site gives no year for it.

2026: the year of the Doomsday Clock setting the Bulletin of the Atomic Scientists pushed closer than ever to midnight, with mirror life cited as a key factor.

Two named funders, the Alfred P. Sloan Foundation and Renaissance Philanthropy, have said they will not

fund research with the goal of creating mirror organisms.

What the author is unsure about

This is an organisation describing itself, so it states positions rather than doubts. The honest way to read it is to notice what it does not claim.

It does not claim mirror organisms exist or are imminent: the definition is "hypothetical", and the risk is always "could pose".

The call for governance carries a condition, which is the closest thing here to an admitted uncertainty: prevention is urged absent significant new evidence for reassurance, which concedes that such evidence could in principle arrive.

The site says a growing international community has acknowledged the risks, but acknowledging a risk is not the same as regulating it, and the news list is mostly statements, recommendations and explainers rather than binding rules. The two funder commitments are the only items that change what someone can actually get money to do.

No dollar figures, staff numbers or funding sources are given, and the September symposium is dated without a year.

For the discussion

The two funders are the most concrete thing on the page: a grant-maker declining to fund a goal is a real constraint on what gets researched, and it arrived faster than any law would have.

Worth raising as a question: nearly every item on the news list is a recommendation or a statement, so what would it take for the UNESCO committee's precautionary moratorium to become something a laboratory is actually bound by?

The conditional in the ask is the subtlest point here, and the strongest: governance is urged unless significant new evidence for reassurance appears, which puts the burden of proof on anyone who wants to proceed rather than on the people raising the alarm.

2.5-5 · CORE · COURSE TIME 13 MIN

The Deep Lunacy that was (almost) 'DEEP VZN'

Rob Reid · 2026

Read the original paper → · the course assigns: whole piece

> **Rob Reid argues that DEEP VZN, a $125 million USAID programme, would have hunted down the world's most dangerous unknown viruses, ranked them, and published their genomes to everyone, and that killing the**

> **programme before it started was the right call.**

Average summary

In 2021, USAID authorised a five-year, **$125 million** budget for a programme called Deep Vision, written as the acronym DEEP VZN. Reid, who campaigned against it, calls it one of humanity's worst ideas and lays out its three planned steps.

Step one was virus hunting. Local teams in a dozen developing countries would search remote places like bat caves and rural bush-meat markets for thousands of undiscovered viruses of completely unknown deadliness, then bring them back to dense population centres and store them in laboratories. Reid's objection is that laboratories leak. Some of the world's top labs have fatally leaked some of history's deadliest agents, and he says all classes of laboratory leak routinely, at unknown rates, because reporting requirements are largely missing and the few that exist are opaque. He counted the word "leak" once in the programme's 114-page description, in a context unrelated to lab leaks, against 34 uses of "gender" and over 50 of "climate".

Step two was characterisation: experiments to work out which of the collected viruses were most likely to be pandemic-grade, with the scariest shipped to American labs.

Step three was publication. The programme would rank the viruses by most terrifying potential and publish their genomes to the whole world, a world Reid says contains roughly 30,000 people with the tools and know-how to build a live virus from a published genome.

The programme was stopped. Reid learned about it shortly after authorisation and tried to blow the whistle in early 2022 by interviewing MIT scientist Kevin Esvelt on his own podcast, After On, episode #62. Sam Harris reposted that episode to a much larger audience, people who knew their way around Washington got in touch, and a loose alliance of people from a wide range of political backgrounds worked to terminate the programme.

Most of the piece is a thought experiment about what would have happened if it had succeeded. Reid supposes it found seven pandemic-ready viruses and names three kinds of people publication would have armed: lone-wolf mass murderers, whose death tolls he says are mostly a function of the weapon available; chaos agents, meaning rogue states or gangs who could make an unverifiable pandemic threat and use it to spread panic and demand ransoms; and omnicidal groups that actually want human extinction, for which Aum Shinrikyo is his case study.

He closes by dismantling the arguments for the programme one by one: vaccines made in advance

could not be shown to work; "if we don't do it, China will" would hand China starting points; and the claim that the horse has already left the barn, because the 1918 flu genome is public, is to him a reason not to make it worse. This is an argued opinion piece, not a study.

Why it's assigned

Day 2 is the Prevent layer, and page 2.5 is the Deny option: stopping an adversary from getting the knowledge in the first place. This reading is the clearest worked example of all three Deny categories failing at once in a single programme: what gets researched, what gets written down, and who can use it. It is also the course's answer to its own guiding question about virus hunting, told by one of the people who helped shut the programme down.

In depth

What DEEP VZN was going to do

In 2021, USAID authorised a five-year, $125 million budget for Deep Vision. Reid describes three activities, connected, he says, in ways that verged on a suicide pact for humanity.

Virus hunting: local teams in a dozen developing countries scouring remote places such as bat caves and bush-meat markets for thousands of undiscovered viruses of completely unknown

deadliness, then hauling them back to dense population centres and storing them in laboratories.

Characterisation: experiments to determine which of those viruses were most likely to be true weapons of mass destruction, with the scariest shipped home to American labs.

Publication: rank the viruses by most terrifying potential and publish their genetic recipes to the entire world. What such a recipe yields, Reid stresses, is not inert data but a live virus, ready to start a pandemic in the right set of lungs, and roughly 30,000 people can build one, in dozens of countries, among them, he says, Iran, Pakistan and North Korea. In success, the programme would have handed those 30,000 unvetted strangers the killing power of a small nuclear arsenal.

Why laboratories are the weak point

His case against step one is containment, not fieldwork. Some of the world's top labs have fatally leaked history's deadliest agents, and he says all classes of laboratory leak routinely, at unknown rates, because reporting requirements are largely missing and the few that exist are opaque. So an undisturbed cave is safer than a lab staffed by flawed humans. His evidence that the programme was unworried is a word count of its own 114-page description, given in the numbers below.

How the programme was stopped

Reid learned of Deep Vision shortly after it was authorised and tried to blow the whistle in early 2022 by interviewing Kevin Esvelt, an MIT scientist who knew the programme well, on his own podcast After On, episode #62. Sam Harris reposted it to his much larger audience. Then people who knew their way around Washington reached out, and a loose alliance from a wide range of political backgrounds collaborated to terminate the programme. From there he pivots to now: biothreats where AI meets synthetic biology, which he calls synbio.

Group one: lone wolf mass murderers

How many pandemic-grade viruses would it have found? Impossible to know, he says, with a small chance of none, but a high single-digit number "seems conservatively likely according to various analyses", and USAID would not have funded the work if it expected failure. For the rest of the piece he assumes seven.

His first group kills as many strangers as possible and dies in the act, and their death toll, he argues, is mostly a function of the weapon to hand: mass stabbers kill fewer than mass shooters, and both kill fewer than a pilot who crashes a plane. He doubts the programme would have armed such a person immediately, because their prevalence seems lower

than 1 in 30,000. But that 30,000 would not have stayed still. The technique, reverse genetics or viral rescue, was first achieved in 1999, and perhaps a hundred people could do it by 2005, against Esvelt's 30,000 in 2022: over two orders of magnitude in just over 15 years.

His first rule: time multiplies the threat

He expects more orders of magnitude as AI, cloud labs and other factors lower the barriers of education, equipment, budget and permission. Two reference points: 300,000 is the rough global count of large commercial jet pilots, one of whom becomes a top-grossing mass murderer every five or so years, and 3,000,000 Americans produce about six mass shooters a year.

Hence his first rule: the passage of time is the ultimate threat multiplier. He unpacks it himself. Deadly information made public today stays available, because the internet never forgets, and the population that can act on it keeps compounding, so assume that lethal skills needing a post-doc today will soon be taught in first-year biology.

None of this, he says, argues against synthetic biology or against melding it with AI, which brings countless benefits and cannot be stopped. What he objects to is judging the risk by what bad actors could do today rather than tomorrow, the same failure of imagination

that left cockpit doors unhardened for decades before 9/11.

Group two: chaos agents

Actors who may have no intention of launching a pandemic, and perhaps no ability, but whom publication would still empower. His illustration is invented, not reported: a rogue state announcing that it has tested the published viruses on political prisoners, found one with a 20% fatality rate, and stored samples in its embassies. Today that would be dismissed, because nobody alive can design such a virus. In a Deep Vision world it would be credible and completely unverifiable, because we do not infect people with viruses to see how deadly they are, and the spies who would have to call the bluff are the ones who swore Saddam Hussein had weapons of mass destruction.

So it is a nuclear-grade deterrent almost any rogue state could claim with partial credibility, while gangs and terror cells could invent similar threats to spread panic, tank markets and demand ransoms. This havoc, he stresses, would have been enabled even if the programme found zero pandemic-grade viruses, because only the most vicious groups could have tested its findings.

Group three: omnicidal groups

Not standard terrorists, who want to kill some people and not others, but groups animated by fringe beliefs that seek human extinction: eco extremists, animal-rights zealots, antinatalists, religious zealots. Aum Shinrikyo is his case study: a Japanese death cult, active in the 1980s and 1990s, seeking a purifying Armageddon their sect would survive, which released sarin gas in the Tokyo subway, killing over a dozen people and injuring thousands. Its money, followers, weapons and in-house scientists are in the numbers below.

The scenario turns on the imperfection of step two. Because characterisation cannot be perfect, the pandemic-grade viruses would be scattered among the top thirty or more rather than sitting neatly at the top, so a group with Aum's resources could have generated every virus near the top, then released all of them at a dozen international airports at once. Seven pandemics at once would leave doctors unable to tell who is sick with what, many people carrying several infections, and a combined fatality rate that could be staggering. That is where he thinks civilisation could topple: nobody would risk a walk around the block, which ends frontline work, and so ends food, law enforcement, electricity and communications.

The arguments for it, and why he rejects each

In the original interview he and Esvelt built every steelman argument they could, meaning the strongest version of the other side's case. The best of them was vaccines made before pandemics hit. His answer: you would not know whether any of them worked, because vaccines are tested by inoculating people during an outbreak and seeing whether the vaccinated stay healthy, so before anyone is sick you get 40 to 50 candidates that might be worthless. Even a successful one would save almost no time: for COVID the saving would have been about two days, the time Moderna took to reach its final candidate after China published the genome, against the 340 days before the vaccine was available, lost to testing, regulators and building a manufacturing line. About 1% off the wait, for the risk of triggering multiple pandemics.

"If we don't do it, China will" he rejects twice over: publishing starting points would accelerate rather than slow a hostile state, and pandemics are not strategically useful weapons, because they are not targetable and will eventually hit whoever launched them. A culprit could not secretly vaccinate its own people first, because a vast national effort cannot be hidden from modern intelligence. On ethnically targeted pandemics, many doubt they are possible, and COVID showed how fast pandemics evolve around pockets of immunity.

"The horse is already out of the barn", because the smallpox and 1918 flu genomes are already available, gets his sharpest line: that is like saying we already have two hijackers on board, so why not invite a few more before the plane leaves the gate.

The 1918 flu genome

He takes the 1918 flu because it is the easiest of those genomes to work from. It killed almost 1 out of 30 humans the last time it ran free, and it has been resurrected, with live virus in multiple labs we know of and almost certainly in labs we do not. It probably would not be as deadly now, because most people have some immunity from its cousins and descendants, and antibiotics control the secondary infections that killed many of its victims. But we can only guess: one expert he surveyed puts the odds of it triggering another pandemic at 10%, which he calls monstrously unacceptable.

The body that made that genome universally available was the United States Centers for Disease Control, which published it in 2005, six years after viral rescue was first achieved, when roughly a hundred people could probably have used it. That brings him to the argument he finds most chilling, and probably the only one that mattered when Deep Vision was approved: that some people see nothing wrong with giving universal access to genomes that could kill at the

scale of a world. He takes that up, with his second rule, in his next post.

Words to know

DEEP VZN Short for Discovery & Exploration of Emerging Pathogens, Viral Zoonoses, a USAID programme Reid reads phonetically as Deep Vision. It is the thing the whole piece is about.

Reverse genetics, also called viral rescue The technique for making a live virus starting from its genetic sequence. It is the hinge of his argument: it turns a published genome into a working weapon, and the number of people who can do it keeps growing.

Characterisation Running experiments on a collected virus to work out what it can do, in this case how likely it is to cause a pandemic. Reid's point is that it is imprecise, so the ranked list it produces would be wrong in dangerous ways.

Steelman The strongest, fairest version of an argument you disagree with. Reid uses it to signal that he and Esvelt tried to make the case for the programme before knocking it down.

Synbio His shorthand for synthetic biology, the engineering of biological systems. He is explicit that he does not want it stopped.

Omnicidal Wanting to kill everyone, as distinct from ordinary terrorism, which wants to kill some people and spare others.

Numbers in this reading

$125 million, the five-year budget USAID authorised for the programme in 2021.

114 pages, the length of the programme description Reid searched; "leak" appears once in it, in an unrelated context, against 34 uses of "gender" and over 50 of "climate".

Roughly 30,000, his estimate of the people with the tools and know-how to build a live virus from a published genome, sourced to Esvelt's ballpark at their 2022 interview and checked with other biologists.

1999, when reverse genetics was first achieved; by 2005, perhaps a hundred people could do it. He calls that growth over two orders of magnitude in just over 15 years.

300,000, the rough global count of large commercial jet pilots, a population that produces a mass murderer every five or so years, with seven deliberately downed jets since 1994, not counting 9/11.

3,000,000, a population of Americans that produces about six mass shooters a year.

Seven, the number of pandemic-ready viruses he assumes the programme would have found, from a "high single-digit number" he calls conservatively likely.

20%, the fatality rate in his invented Iranian announcement. It is an illustration, not a finding.

Over a billion dollars in assets, over 50,000 followers and over 300 members of an internal science ministry: Aum Shinrikyo's capacity. Their 1995-era sarin attack in the Tokyo subway killed over a dozen and injured thousands.

40 to 50, the number of iffy vaccine candidates he thinks the programme would have generated.

Two days, the time saved on a COVID vaccine candidate, against 340 days from Moderna's candidate to availability, which he converts into about 1% off the wait.

1 out of 30 humans, the share killed by the 1918 flu the last time it ran free. One expert he surveyed gives the resurrected genome 10% odds of triggering another pandemic.

2005, when the US Centers for Disease Control published the 1918 genome, six years after viral

rescue was first achieved, when roughly a hundred people could probably have used it.

What the author is unsure about

He says plainly that it is impossible to know how many pandemic-grade viruses the programme would have found, and that there is a small chance it would have found none. Seven is an assumption he adopts "for the sake of our exploration", built on a high single-digit number he attributes to various analyses he does not name.

The 30,000 figure is a ballpark from Esvelt, checked with other biologists, not a count.

He doubts the programme would have armed a lone-wolf killer immediately, because the prevalence of such people seems lower than 1 in 30,000.

The 10% odds on the 1918 flu come from one expert he surveyed, and he says we can only guess what happens if that virus gets loose. He also lists reasons it might be less deadly now.

The Iranian announcement is his own invention, written as an illustration.

He concedes there is no group today openly seeking human extinction, and argues from the Aum precedent and the reticence of such groups rather than from current evidence.

He flags his own standing twice: he notes he is a science-fiction writer making a large claim, and he writes throughout as an advocate who helped kill the programme. The piece is an argument, not a study, and it ends mid-series with his second rule still unwritten.

For the discussion

Reid's three steps map exactly onto the unit's three Deny categories, so the useful question is which step was the fatal one: if a virus-hunting programme collected and characterised but published nothing, would he still have fought it?

His 1% point is the sharpest thing in the piece worth testing out loud: if pre-positioned vaccine candidates would only have saved about two days against a 340-day wait, then the real bottleneck is manufacturing and regulation, not discovery, which is an argument about where prevention money should go.

The uncomfortable tension is his word count of the programme description: it is a rhetorical move, not evidence about biosafety, and it is worth asking whether a programme can be judged on what its paperwork failed to mention.

2.5-6 · FURTHER · COURSE TIME 10 MIN

Behind the cancellation of a $125 million US virus hunting program

Kelsey Piper · 2023

Read the original paper → · the course assigns: whole piece

> **Kelsey Piper reports that the US government cancelled DEEP VZN, a $125 million virus-hunting programme, and argues the decision was right on two counts: the fieldwork was exposing researchers to real danger, and success would have meant publishing the blueprints of viruses nobody currently knows how to use.**

Average summary

The news is the cancellation. DEEP VZN was a **$125 million** USAID programme to track down viruses that might cause the next pandemic. Piper reports, citing the British Medical Journal, that post-COVID worries about biosafety and biosecurity prompted the US government to reevaluate and ultimately cancel it.

She frames the case for the programme as simple: find the next pandemic before it finds us. The case against it had two parts, one old and one newer. The old objection was usefulness: there are millions of viruses out there and it is hard to guess which will threaten people, and leaders in the field complained in a 2018 commentary in Nature that viral discovery

was of little practical value. The newer objection was that the work creates risk of its own.

The first risk is accidents. In remote-region discovery work, she writes, high-risk exposures became practically routine. Her example is a Thai team doing US-backed virus hunting who were repeatedly bitten by bats, reported earlier that year by the Washington Post. Thiravat Hemachudha, who supervised the expeditions, concluded the work was too dangerous to continue, told his American funders he was done, and said it had produced no benefits for Thailand while putting his researchers at enormous risk. She also quotes Johns Hopkins professor Steven Salzberg, who in 2021 asked why we fund scientists to collect bats infected with deadly viruses and bring them back into the midst of cities.

The second risk is the one she says cannot be mitigated, because it arrives when the work goes exactly as planned. If discovery found a virus as contagious and as deadly as the 1918 influenza, but, unlike that virus today, one nobody has any preexisting immunity to, publishing it might help us build vaccines and might also hand a blueprint to anyone wanting mass death. She grounds that in 1995, when the Japanese doomsday cult Aum Shinrikyo carried out a sarin attack on the Tokyo subway along with other chemical and biological attacks including anthrax; the man who made the

sarin was a Kyoto University-trained virologist who could not get a truly deadly bioweapon, because, she writes, we do not know of any yet. MIT biologist Kevin Esvelt's version is that terrorists and nihilists are currently blocked by not knowing which viruses would do the job, so characterising candidates from nature does their work for them.

The last section is about the field's reaction. Esvelt says many virologists feel attacked and the whole field demonised, and insists 99.9 percent of virological research is lifesaving, with this one practice being massively net negative. Piper's own position is that cancellation should be the first step toward better prevention work rather than a retreat: studying viruses already crossing into humans at human-wildlife interfaces remains crucial.

Why it's assigned

Day 2 is the Prevent layer and this page is Deny, which asks what should not be researched or published. This is the shortest honest account of the course's own guiding question about virus hunting: what the programmes were for, why experts came to think the benefit was thin, and which specific risk finally ended this one. It also shows what denial costs inside a scientific community, because the people who lost the funding were mostly doing lifesaving work.

Words to know

Virus hunting, or viral discovery Going out to find and characterise previously unknown viruses in wildlife. The practice the piece is about, as distinct from virology in general.

DEEP VZN The USAID programme that was cancelled. Piper describes it as aimed at discovering viruses in wildlife that could threaten humans.

Biosafety and biosecurity Preventing accidents, and preventing deliberate misuse. The piece turns on the difference: the first risk can be reduced with planning, the second cannot, because it comes from the results themselves.

Human-wildlife interface The places where people and animals meet closely enough for infections to cross. Piper's constructive point is that viruses already crossing there are worth studying, unlike viruses that have never reached humans.

Characterise To test a virus in the laboratory to work out what it can do. Esvelt's objection is to characterising viruses in order to credibly identify which ones could cause a pandemic.

Numbers in this reading

$125 million: the size of the USAID programme to track down viruses that might cause the next pandemic.

A 2018 commentary in Nature, in which leaders in the field called viral discovery "of little practical value".

Millions of viruses are out there, which is her shorthand for why guessing which ones matter is so hard.

1995: the Aum Shinrikyo sarin attack on the Tokyo subway, alongside other chemical and biological attacks including anthrax, by a group whose sarin was made by a Kyoto University-trained virologist.

1918: the influenza pandemic she uses as her benchmark for contagiousness and deadliness. Her hypothetical is worse, because unlike the 1918 virus today it would be one nobody has any preexisting immunity to.

99.9 percent of virological research is lifesaving, in Esvelt's words, with the identification of pandemic-capable viruses as the exception he calls massively net negative.

2021: when Steven Salzberg made his "leave the bats alone" argument.

A decade: how long Hemachudha's team spent on the work before deciding to stop.

What the author is unsure about

Piper writes as a columnist, and most of the piece is argument rather than reporting. The one piece of hard

news, that worries about the work prompted the US government to reevaluate and cancel the programme, is attributed to reporting by the British Medical Journal rather than to her own.

She is careful about where certainty ends. The scenario of discovering a 1918-grade virus is put as a hypothetical, with "say that you can". She concedes the misuse possibility "might sound speculative" and then argues from the Aum Shinrikyo precedent instead, with Syra Madad's line that history has shown there are groups that will try.

The strongest honest admission is a gap in knowledge treated as protective: a bioweapon of that kind has not been available to anyone, because we do not know of any such virus yet. The whole argument rests on that "yet", and on Esvelt's claim that discovery work is what would end it.

She also grants the opposite case real weight. Virology has done tremendous good and will continue to do so; most virologists have done nothing dangerous; accident risk specifically could be reduced with comprehensive global planning and coordination, which is why she separates it from the misuse risk that cannot be.

For the discussion

The sharpest line in the piece is Esvelt's, and it is a general principle rather than a point about this

programme: if the thing that currently stops a determined person is not knowing which virus would work, then publishing a ranked list of candidates is doing their work for them.

The tension worth raising is who pays for prevention: Hemachudha's team spent a decade taking personal risk, got no benefit for Thailand out of it, and the programme's cancellation was still experienced by the field as condemnation.

A question for the group: Piper draws the line between viruses already crossing into humans, which she thinks we must study, and viruses that have never left remote wildlife, which she thinks we should leave alone. Is that line administrable, or does it collapse the moment a surveillance programme has to decide what it is allowed to sample?

2.5-7 · FURTHER · COURSE TIME 115 MIN

Recipes for Future Plagues

The After On Podcast (host: Rob Reid) · 2022

Read the original paper → · the course assigns: the publisher's show notes only; the episode page carries no transcript and the audio was not transcribed

> **A roughly 115-minute conversation in which MIT's Kevin Esvelt argues that a USAID virus-hunting programme was on course to find unknown pandemic-grade viruses and publish**

> **their genomes, and the host, convinced enough to co-release the episode on two shows, spends the second half arguing the other side.**

Average summary

This summary is written from the publisher's show notes, not from a transcript, because the episode page carries none and the audio was not transcribed. So what follows is the host's own published description of the conversation and its stakes, not the guest's words. The episode is now **Forever On episode 58**, Kevin Esvelt, dated **28 February 2022**, since the After On archive moved to foreveron.com. The same conversation was co-released on Sam Harris's Making Sense as Recipes for Future Plagues, which is the title the course uses. A third-party transcript of the Making Sense release exists online for anyone who needs exact wording.

The subject is **DEEP VZN**, then a new programme inside USAID, the agency that distributes most American foreign aid. According to Esvelt, an evolutionary engineer at MIT, the programme aimed to discover new pandemic-grade viruses and then broadcast their genomes to the world. The number the note carries is his estimate that around **30,000** people in dozens of countries could then easily assemble them. The host's framing of the consequence is that this could easily and rapidly lead

to the deaths of millions, or even hundreds of millions of people, while stating that he is certain the people behind the programme have entirely good intentions.

The note explains the unusual co-release. Both shows judged the subject extraordinarily important and wanted it out fast; because the host had done research connected to the subject and already knew the guest, he conducted the interview and both shows distributed it.

The host also anticipates disbelief, saying he struggled to believe his own ears at first, and arguing that listening to Esvelt makes clear this is not a pandemic-era conspiracy theory even for a listener who ends up unpersuaded that the danger is as large as Esvelt believes. He states that he would not have conducted the interview if he thought the guest had a partisan agenda, that he does not know Esvelt's party affiliation and does not care, and that every point he has heard Esvelt make on the subject has been rooted in science rather than politics.

He discloses his own view. Some of the programme's plans could help fight pandemics, but after many hours of research and conversations with people in academia, philanthropy, security and entrepreneurship, he believes most of its agenda is appallingly misguided, and he says hiding that would be dishonest.

The structure is worth knowing before listening. In the second half the host presents all the best arguments he has heard or thought of in favour of DEEP VZN, on the grounds that the subject is too important for listeners not to hear the other side, and he says he wants to convey why some very smart and ethical people sincerely think the programme is a good idea without that making them any less smart or ethical.

Why it's assigned

This is the primary-source argument behind page 2.5's third topic, virus hunting programmes, and the page's guiding question about what their benefits were and why they were shut down. It is the conversation that put DEEP VZN in front of a general audience in early 2022, before the Senate hold and the cancellation that readings 2.5-5, 2.5-6 and 2.5-8 cover, which makes it the Deny option's clearest case of regulating knowledge generation: not what gets published about a pathogen, but whether to go looking for one at all. A follow-up episode exists for the aftermath, Forever On episode 62, DEEP VZN Revisited, from 12 March 2026.

Words to know

- **DEEP VZN**: Discovery and Exploration of Emerging Pathogens, Viral Zoonoses, the USAID programme at the centre of the episode. It is the subject of three other readings on this page.

- **Virus hunting**: searching wildlife for previously unknown viruses that might infect people. The dispute is over whether finding and characterising them helps defence more than it helps attackers.
- **Zoonosis**: a disease that passes from animals to humans. It is the rationale for virus hunting, since most new human pathogens arrive that way.
- **Information hazard**: knowledge that causes harm by existing in the open. The genome of an unknown pandemic-grade virus is the host's example of one.
- **Co-release**: publishing the same episode on two shows at once. Here it is evidence of how urgent both hosts considered the subject rather than a production detail.

Numbers in this reading

- Around 30,000 people in dozens of countries who, by Esvelt's estimate as the note states it, could easily assemble such viruses once the genomes were published.
- Roughly 115 minutes of running time, with the second half given to the arguments in favour of the programme.
- 28 February 2022: the release date, as Forever On episode 58, co-released on Making Sense.
- Millions, or even hundreds of millions of deaths: the host's framing of the potential consequence, stated as his reading of the stakes rather than as an estimate.

What the author is unsure about

The note is explicit that this is an argument, not a settled account. The host says he struggled to believe it at first, and he writes for a listener who may finish the episode unpersuaded that the danger is as large as Esvelt believes. He insists the people running the programme have entirely good intentions, and that smart, ethical people sincerely support it. His own judgement, that most of the agenda is appallingly misguided, is offered as a disclosed opinion formed from research and conversations, and he concedes that some of the programme's plans could help fight pandemics. The 30,000 figure is Esvelt's estimate as the host reports it. And because only the show notes were available, nothing here reflects the evidence, qualifications or counterarguments actually voiced during the conversation.

For the discussion

- The useful framing to bring is that this is the earliest of the four DEEP VZN readings, published while the programme was still live, so it shows what the objection looked like as a prediction rather than as hindsight after the cancellation.
- The host's choice to spend the second half on the strongest case for the programme is the model for how this unit's deep dive should be argued, and worth imitating out loud: the virus hunting case

rests on real surveillance benefits, not on carelessness.

- A question worth asking: the objection here is to publishing the genomes rather than to the searching, so would DEEP VZN have been acceptable with restricted data sharing, and who would have held the data?

2.5-8 · FURTHER · COURSE TIME 5 MIN

Senate Foreign Relations Committee letter on USAID's DEEP VZN program

U.S. Senate Foreign Relations Committee · 2023

Read the original paper → · the course assigns: the whole letter

> **A single senator tells USAID's administrator that after two years of briefings he still will not release the money for DEEP VZN, because he does not believe the agency can control its research partners and does not believe hunting novel viruses is a development agency's job.**

Average summary

This is a one-page letter, not an article, and it is short enough that this summary is shorter than the course's other average summaries. It is dated **May 16, 2023**, written on Committee on Foreign Relations letterhead

to Samantha Power, the Administrator of the US Agency for International Development, and signed by James E. Risch as Ranking Member. It is written throughout in the first person: my office, my hold.

The subject is USAID's work to identify zoonotic disease threats, specifically the funds under congressional notification #148 of April 2021 for Discovery & Exploration of Emerging Pathogens, Viral Zoonoses, which is DEEP VZN. Risch writes that he remains unconvinced this is an effective use of US foreign assistance funding and therefore cannot clear his hold.

He documents the engagement to date to show that the objection is not new: more than two years of work by USAID on concerns raised by his office and others, including staff briefings on October 26 and December 10, 2021, February 22, 2022, and January 12 and March 27, 2023, plus his written inquiries of November 23, 2021 and February 2, 2022. He explicitly notes and appreciates USAID's recent efforts to address serious lapses in its management of the **$208 million** PREDICT programme, which ran from 2009 to 2019, by building more stringent safeguards into the award for its proposed successor.

Then the two objections. First, oversight: he remains deeply concerned that USAID does not yet exercise the level of oversight and control over its prime and sub-prime implementing partners that life sciences

research of dual use concern surely requires, and he calls it highly unlikely that it could reach that level over the coming days, weeks, months or years. Second, institutional fit: he is unconvinced that hunting novel viruses would or should ever fall within the core competency of the US government's lead development agency.

His conclusion is flat. He cannot support moving forward with field work or any further investment in DEEP VZN under this notification as currently designed, even with the enhanced safeguards described in Power's letter and accompanying slides of April 24, 2023.

Why it's assigned

Day 2 is the Prevent layer and this page is Deny, the layer that decides what gets researched. This letter is the mechanism by which the Deny layer actually operated in the DEEP VZN case: not a scientific finding or a regulation, but one legislator declining to release funds. Read it next to the Piper and Reid pieces to see the difference between the public argument about the programme and the paperwork that stopped it.

Words to know

Congressional notification, CN The document an agency sends its oversight committees before committing funds, which is what "CN #148 (April

2021)" refers to. In practice a committee leader can place a hold and the agency waits (not from this reading).

Hold An informal block on proceeding, which is why the letter's operative sentence is that he cannot clear on his hold rather than that the programme is cancelled.

Ranking Member The senior member of the committee's minority party (not from this reading). It matters here because the letter is institutional in form and personal in substance.

Dual use concern Research that could serve both beneficial and harmful ends. Risch uses the phrase "life sciences research of dual use concern" to set the standard of control he thinks USAID has not met.

Prime and sub-prime implementing partners The organisation awarded the money and the ones it subcontracts to. His worry is about the chain, not just the grantee.

Core competency What an organisation is actually built to do. His second objection is institutional rather than scientific: he argues virus hunting does not belong in a development agency at all.

Numbers in this reading

$208 million: the PREDICT programme, 2009 to 2019, whose management the letter describes as having serious lapses. DEEP VZN is called its proposed successor.

CN #148, April 2021: the congressional notification carrying the funds in dispute.

More than two years: how long USAID had been working on the concerns raised by his office and others.

Five staff briefings: October 26 and December 10, 2021, February 22, 2022, January 12 and March 27, 2023. Plus two written inquiries, November 23, 2021 and February 2, 2022.

April 24, 2023: the date of the letter and slides from USAID describing enhanced safeguards, which he says do not change his position.

May 16, 2023: the date of this letter.

What the author is unsure about

Very little is hedged, which is itself the point: this is an assertion of a position, not an analysis.

The one explicit probability judgement is his own forecast, and he labels it: even if USAID could attain the required level of control over its partners in the coming days, weeks, months or years, which he calls highly unlikely, he would still be unconvinced the work

belongs at a development agency. So the second objection does not depend on the first being right.

What he does not do is also worth noting. He gives no evidence for the oversight claim beyond his judgement, does not describe what the serious lapses in PREDICT were, does not say what the enhanced safeguards contained, and makes no biosafety or biosecurity argument about the science itself, as distinct from the agency's capacity to supervise it. There is no claim that an accident occurred.

The letter is signed by one member and speaks in the first person, so it is not a committee finding.

For the discussion

The decisive objection here is institutional, not scientific: not that virus hunting is too dangerous, but that it is not what a development agency is for, which is a route to denial that needs no risk assessment at all.

His oversight worry is the one that generalises beyond this programme: the money reaches prime partners who subcontract to others, and the letter's claim is that nobody in that chain was being supervised to the standard dual use research requires.

Worth raising as a tension: the letter credits USAID for adding more stringent safeguards after PREDICT's problems and then refuses the money anyway, which

raises the question of what evidence could ever have cleared the hold.

2.5-9 · CORE · COURSE TIME 25 MIN

Why Do Exceptionally Dangerous Gain-of-Function Experiments in Influenza?

Marc Lipsitch · 2018

Read the original paper → · the course assigns: whole chapter

> **Marc Lipsitch argues that the small class of influenza experiments expected to create a pathogen that is both highly deadly and readily transmissible between humans carries a risk that lands on the whole world, and that the questions only those experiments can answer are too narrow to be worth it.**

Average summary

This is a chapter in a laboratory manual of influenza methods, written by an epidemiologist who says the title is a serious question, not a rhetorical one: what kinds of experiments should influenza virologists choose to do? Lipsitch states up front that his view is contrary to that of many leading influenza virologists.

The class of work he means is **gain-of-function research of concern**, or GoFRoC, as the US National Science Advisory Board for Biosecurity defines it, also

described as research likely to create a **potential pandemic pathogen (PPP)**: experiments reasonably expected to produce a novel strain that combines high virulence for humans with high transmissibility between humans, especially one that existing population immunity would not touch because it is antigenically new.

His risk argument is about who carries the risk. Normally scientific merit is judged first and biosafety conditions are set afterwards to let the work proceed safely, and for US NIH grants biosafety is reviewed only after the science is scored. He says that system has worked relatively well, with known deaths from laboratory exposure extremely rare, around one per year globally. It breaks down for PPP work, because an accident there would not stop at the worker or the building. Most lab pathogens are limited by something, such as not spreading person to person, or already being everywhere, or being held down by vaccination. A potential pandemic pathogen is defined by meeting none of those limits.

He gives the contested numbers plainly. His own published estimate puts a single year of work in a single lab on a novel transmissible influenza at between **1 in 1,000 and 1 in 10,000** odds of accidentally starting a pandemic. Ron Fouchier's reply put it more than one million times lower. Both are small, both rest on assumptions because the data are

poor, but multiplied by the size of a pandemic, Lipsitch says the risk is unacceptable even a thousandfold below his own figure.

Then the value side, which is most of the chapter. He accepts there are questions only gain-of-function work can answer, but shows they are narrow: such experiments cannot say whether a set of mutations is necessary, or the smallest sufficient set, or whether it would arise in nature, or whether it transfers to another viral background, or whether what spreads between ferrets would spread between people. And he argues the public health wins claimed for this work, over H7N9 in Cambodia and over the recreated 1918 virus, were all available from safer methods first.

Why it's assigned

Day 2 is the Prevent layer, and this page is the Deny option: keeping dangerous knowledge from being made in the first place. This chapter is the course's answer to its own guiding question about how individual scientists judge the risk-benefit ratio of their own work, and it answers it uncomfortably: the judgement is made by the people with the most to gain, biosafety is handled after the fact, and the risk falls on people who never consented and do not know the work is happening. It also gives the honest history of who has tried to govern this class of research, and how little was settled.

In depth

The question, and two views he rejects

Lipsitch, a Harvard epidemiologist, rejects two positions: that in the right laboratories no influenza gain-of-function experiment is exceptionally dangerous, and that experiments likely to produce potential pandemic pathogens are worth whatever risk they carry. His terms are interchangeable: exceptionally dangerous gain-of-function experiments, gain-of-function research of concern (GoFRoC) as the National Science Advisory Board for Biosecurity defines it, and research likely to create a PPP. The paradigm case, which the chapter's one figure diagrams, is passaging modified H5N1 through ferrets to select for droplet transmissibility: start from an avian-adapted virus, introduce change, select, repeat.

A brief history of the debate

The first effort to make a PPP was arguably Tumpey and colleagues' recreation of the 1918 H1N1 strain from viral RNA preserved in frozen corpses. Whether that was wise was asked, but only inside the US Department of Health and Human Services, particularly the NIH, and the work went ahead.

At a conference in Malta in September 2011, Ron Fouchier presented ferret experiments that had produced a virus capable of small-droplet

transmission between ferrets, which many, though not all, read as a proxy for human-to-human spread. Yoshihiro Kawaoka's laboratory soon reported a related set. The controversy was about publication, and about biosecurity specifically: whether naming the mutations would hand bioterrorists a recipe. The US government, which had funded both studies, recommended publishing in full, and they were. Virologists briefly announced, then lifted, a self-imposed moratorium.

Others raised a different risk: a laboratory accident releasing something combining high virulence and antigenic novelty with human transmissibility. Klotz and Sylvester coined the term potential pandemic pathogen; the first major discussion meeting he knows of was at the Royal Society of London in 2012. After mishaps at US high-containment federal laboratories, criticism hardened, and a July 2014 meeting he co-organised produced the Cambridge Working Group, which called for the work to be curtailed pending a formal risk-benefit assessment; a countervailing group, Scientists for Science, formed against it. In October 2014 the White House Office of Science and Technology Policy and the NIH announced a funding pause, and an NSABB-led assessment carried out by a private firm, Gryphon Scientific. The pause initially touched about two dozen NIH grants on influenza and coronaviruses, was

narrowed to about half that, and sat inside a portfolio of about 250 active grants on those viruses.

The process surfaced problems nobody had highlighted: no framework exists for evaluating research risks to people who are not research participants, biosafety data from laboratories are very poorly available, and risk-benefit calculations are uncertain as a result. OSTP announced policy guidance in January 2017, and nothing was finally resolved.

How scientists choose, and where biosafety sits

Researchers pick experiments by curiosity, novelty, importance, speed, cost and elegance; biosafety is normally not on that list. The scientific judgement comes first, and biosafety conditions are then set so the work can be done safely. For US NIH grants, biosafety is reviewed only after scientific merit is scored, and is not supposed to affect that score. He grants that this has worked relatively well, with known fatalities from laboratory exposure extremely rare, around one per year globally.

Why this class of work is different

The arrangement fails when an accident reaches past the building to a much larger, even global, population. Accident risk is conventionally probability times expected magnitude of harm, and for most pathogens the second term is small because something limits it:

anthrax does not spread person to person, filoviruses would be contained in a developed-world setting, gut flora are already everywhere, measles is held down by vaccination, and flaviviruses lack a vector in cool climates. A potential pandemic pathogen meets none of these: widespread susceptibility, high virulence, little or no current human infection, and an efficient route of transmission where the work is done. So a probability that is acceptable at a cost of one or two cases is not acceptable at thousands or millions.

His published estimate, built from laboratory-acquired infections in US BSL3 laboratories, puts a single year of work in a single laboratory on a novel influenza readily transmissible between humans at between a 1/1000 and a 1/10,000 chance of accidentally sparking a pandemic. Fouchier's reply placed it more than one million times lower. His point is not the number: both are small, both rest on assumptions because data are limited, but multiplied by the magnitude of a pandemic, his figure or anything between the two is clearly unacceptable, and would stay unacceptable to most people a thousandfold lower.

Of the major mishaps at US government labs in recent years, nearly all involved taking the agent out of high containment into a lower-containment lab because it was believed inert. Containment cannot stop someone deliberately removing material thought

to be safe, and vaccinating or monitoring lab staff does nothing about exposure elsewhere. Accidents in even the most secure labs are inevitable; what is unique here is that one could lead to sustained transmission and global spread. That risk is different in kind, so it should count, like great expense, as a reason not to run the experiment at all, rather than something settled afterwards.

What only these experiments can answer

Only a gain-of-function experiment can answer one form of question: starting from a particular viral genotype, is a particular set of genetic changes sufficient to produce a trait the starting genotype lacked? You cannot measure a trait you have not created.

Then the questions it cannot answer. Whether that set of changes is necessary: unanswerable in principle, because you can never try every alternative set. Whether it is the smallest sufficient set: unanswerable in practice beyond about three mutations, because the roughly 10 to the power of 12 combinations cannot all be tested, least of all for a trait like transmission, assessed in several animals per sequence. Whether the mutations would arise in nature: no, because natural viruses face competing pressures and forces such as population bottlenecks, and almost none share the starting genotype. Whether they would work in another viral background: only if

epistasis were rare in influenza, and it is pervasive, since the very mutations that gave H5N1 hemagglutinin human receptor specificity here failed in a slightly different background.

Whether a ferret-transmissible virus would transmit between humans can never be known, barring deliberate human infection, which would be unethical, or an accident, which would be catastrophic. Ferret transmissibility is a very good but imperfect predictor, and at least one prominent influenza virologist has argued these strains probably would not spread in people. Taken seriously, that undercuts using the mutations this work finds to gauge the risk from viruses in nature.

What matters for prevention, and the safer routes to it

The practical goal is preventing and responding to novel pandemics, which partly depends on predicting which strains in animal reservoirs are threatening. If a strain from birds or pigs carries traits predicting human transmission, that can justify culling infected animals or making a strain-specific vaccine. Both depend on prioritising, because culling every infected animal is impractical and vaccines are currently practical only strain by strain. Both sit inside a portfolio that does not depend on guessing the strain: better surveillance, and stockpiles of antivirals, protective equipment and ventilators.

His table, modified from an earlier paper of his, lists ten approaches that inform all this without creating a potential pandemic pathogen, from molecular modelling and single-protein binding studies through sequence-database comparisons to universal vaccine research, each with its own advantage in throughput or generality. Even gain-of-function work can be done without pandemic-potential strains: in attenuated strains, in strains of lower virulence, or in strains the population is already immune to.

Prioritisation will stay imperfect, and here is why: there have been only four influenza pandemics since the turn of the twentieth century, so models of human transmissibility rest on essentially four independent data points. 2009 showed the cost: the strain that emerged arguably failed all three then-accepted criteria of human adaptation, including the lysine at PB2 amino acid 627 thought necessary, and the criteria had to be rewritten. No amount of research, he argues, will produce a reliable predictive model, because the data needed to validate one are the pandemics themselves.

What GoFRoC adds that safer work did not

Authors from the US CDC wrote in 2015 that countermeasures against H7N9 strains in Cambodia that year could be prioritised only thanks to gain-of-function experiments. Lipsitch calls that exaggerated: every mutation they flagged had already been

pinpointed in published sequence analysis of existing strains, or in experiments that were not GoFRoC.

Participants in the 1918 recreation claimed major public health benefits after 2009. He answers that their examples, such as recognising that elderly people alive in 1918 were probably protected in 2009, in no way depended on rebuilding the virus: comparing the two sequences would have raised the hypothesis, and serological tests on people born before 1918 could and did test it.

Conclusion

Prediction, prevention and mitigation are necessarily inexact, he concludes, because we have so few pandemics to study. This class of experiment can add to the evidence base but cannot qualitatively change it, and its contribution to applied public health has been far more modest than claimed. Counting the unique risk as a factor in whether to do the work at all exposes how narrow the questions are that it alone can answer. He ends by urging virologists toward methods that are safe and, he says, more rewarding scientifically.

Words to know

Gain-of-function Changing an organism so it acquires an ability it did not have. The term covers a huge, mostly unremarkable range of work, which is

why Lipsitch insists on the narrow sub-class he is actually arguing about.

GoFRoC, gain-of-function research of concern The US government's label, defined by the National Science Advisory Board for Biosecurity, for experiments reasonably expected to create a strain combining high virulence for humans with high human transmissibility.

Potential pandemic pathogen, PPP The term Klotz and Sylvester coined for such a strain: an infectious agent capable of both high virulence and efficient transmission between humans. It is the unit his whole risk argument works in.

Antigenic novelty Being different enough from anything people have met that existing immunity does not recognise it. It is what makes a virulent, transmissible strain a population-level problem rather than an individual one.

Passage Growing a virus through a series of hosts or cultures, which lets variants that do better in that setting take over. Selection through passage is how the paradigm experiments produce the trait they are looking for.

Epistasis When the effect of one genetic change depends on the other changes present. Because epistasis is pervasive in influenza, a result obtained in one viral background may not transfer to another,

which is one of his main limits on what these experiments prove.

Loss-of-function, and gain-plus-loss-of-function Removing an ability, and then removing one and selecting for its return. He offers the second as a way to study the same biology without ever creating a pandemic-capable strain.

Numbers in this reading

Between 1/1000 and 1/10,000: his published estimate of the chance that a single year of work in a single laboratory on a novel human-transmissible influenza accidentally sparks a pandemic, built from laboratory-acquired infections in US BSL3 laboratories.

More than one million times lower: where Fouchier's reply placed that probability.

1000-fold lower than his estimate: the point at which he says the expected risk would still be unacceptable to most people.

Around one per year globally: known fatalities from laboratory exposure to pathogens, which he uses to show the ordinary system has worked.

About two dozen NIH grants initially affected by the October 2014 funding pause, soon narrowed to about half that number, out of an NIH portfolio of about 250 active grants on influenza and coronaviruses.

Four: influenza pandemics since the turn of the twentieth century, and therefore the number of independent data points any model of pandemic potential rests on.

Roughly 10 to the power of 12: the combinations of three or more mutations that would have to be tested to show a set is the smallest sufficient one, which is why he calls that question unanswerable in practice.

PB2 amino acid 627: the site where lysine had been thought necessary for human adaptation, and which the 2009 pandemic strain lacked.

Five years: how long the debate had been intense at the time of writing. 1918, 2004, September 2011, 2012, July 2014, October 2014, 2015 and January 2017 are the dated steps in his history.

What the author is unsure about

He is clear that this is an argument, not a finding, and that it runs against the view of many leading influenza virologists.

On the central probability, he says plainly that he will not try to settle the dispute with Fouchier, that both estimates are small, and that both depend on assumptions because the data are limited. His case is built to survive being wrong by a factor of a thousand rather than to prove a particular number.

He states that biosafety data from laboratories in the United States and elsewhere are very poorly available, and that risk-benefit calculations are uncertain because of it.

He accepts that some questions can be answered only by gain-of-function experiments, and some only by experiments that create a potential pandemic pathogen. His claim is that the extra value is modest, not that there is none.

He says we will never know whether a ferret-transmissible strain would transmit between humans, and reports that at least one prominent virologist thinks it would not.

He insists prediction will stay imperfect, and goes further: no amount of research can make the model reliable, because validating it would require more pandemics.

On the history, he flags the limits of his own account, which is US-centred, and says he does not know whether China held a formal debate or whether Western press coverage simply missed it. He also notes that at the time of writing there was no final resolution of the policy question, and that records of the NSABB meetings were not available online.

For the discussion

His strongest structural point is not about influenza at all: in the normal grant system, scientific merit is scored first and biosafety is arranged afterwards, and for this one class of work that order is backwards, which is a governance claim anyone could test against their own field.

The tension worth raising is consent. He says there is no framework for assessing research risk to people who are not research participants, and here those people are everyone, which is a gap in research ethics rather than in biosafety engineering.

A good question for the group: if his and Fouchier's estimates differ by over a million times and neither can be checked because the biosafety data do not exist, who should get the benefit of the doubt, and what would have to be measured before anyone could settle it?

2.6-1 · CORE · COURSE TIME 40 MIN

Preliminary Taxonomy of AI-Bio Misuse Mitigations

Frontier Model Forum · 2025

Read the original paper → · the course assigns: whole piece

> **The Frontier Model Forum sorts the safeguards meant to stop someone using a frontier AI model to help build a bioweapon**

> **into five functions, gives each one its biological use and its limits, and argues that none of them holds alone, so technical and societal layers have to be stacked.**

Average summary

This is an issue brief, not a study. The **Frontier Model Forum** (FMF) wrote it from discussions with experts inside the Forum and in the wider biosafety and biosecurity communities, and calls it preliminary on purpose.

The problem it starts from: frontier AI could speed up good biological research, and the same dual-use nature could let a malicious actor use a legitimately accessed model to create a biological threat, such as a bioweapon. Under the Frontier AI Safety Commitments, all FMF member firms have written mitigation measures (safeguards) into their frontier AI frameworks, to kick in once capability assessments suggest a model has reached an "enabling capability threshold". But those frameworks name general safeguards, such as harmlessness training, harm refusal techniques, and input and output monitoring, without saying which of them actually suit biological risk. This brief is the attempt to lay that out.

Scope matters. "Frontier AI" here means mainly large language models, though the safeguards may also help with AI-enabled Biological Tools (BTs), which are

models trained on biological data. Misuse means only severe, large-scale catastrophic scenarios, such as creating and deploying a bioweapon. The brief deliberately skips two things: protecting the safeguards themselves from being broken (jailbreak prevention, secure access protocols), and organisational measures like safety culture and risk governance.

The taxonomy has **five functions**. Capability limitation changes the weights or the training process so the model never holds the dangerous knowledge. Behavioural alignment shapes what the model will say. Detection and intervention watches inputs and outputs. Access control decides who gets in and what they can reach. Supporting ecosystem mitigations are things developers do to help other people's defences work, like sharing threat intelligence with DNA synthesis companies. Safeguards are also sorted by where they sit: model-level, system-level, societal-level.

Every single safeguard in the five tables comes with a limitations column, and that is the real content. Filtering struggles because the same knowledge builds vaccines. Unlearning may be reversible with relatively modest effort. Refusal training changes surface behaviour, not capability, so targeted fine-tuning can undo it. Output filters may be bypassed by splitting a request into benign-looking substeps or

across accounts, and may produce a high rate of false positives that block legitimate scientific queries. Geogating can be bypassed with a VPN.

The brief's conclusion is defence in depth: layer imperfect guardrails whose failure modes are uncorrelated, so when they fail they fail independently.

Why it's assigned

Unit 2 is the **Delay** layer, and this page is its Deny sub-layer: keeping threat actors away from the knowledge and infrastructure they would need. This brief is the catalogue for that sub-layer at the level of the model itself, and it touches all five of the page's topics: pre-training filtering, fine-tuning, output filtering, API and access control, and know-your-customer verification, which is its nearest thing to proof of personhood. Each technique arrives with its cost and its failure mode written next to it. It is also the direct answer to the page's guiding question about which interventions deny robustly while keeping models useful for legitimate researchers, because the brief refuses to rank them and says why: measuring resilience to novel adversarial approaches is still an open research area.

In depth

What the brief is, and what it leaves out

Frontier model developers have committed to researching, implementing, validating and sharing mitigation measures, also called **safeguards**. This preliminary taxonomy of them was built from discussions with experts inside the Frontier Model Forum and in the broader biosafety and biosecurity communities.

Under the Frontier AI Safety Commitments, every FMF member firm has written mitigations into its frontier AI framework, triggered once capability assessments suggest a model has hit an **enabling capability threshold**: an ability that could enable extreme harms if the model were deployed without extra safeguards. Those frameworks name harmlessness training, harm refusal techniques, and input and output monitoring, but never say which suit biological risk. Footnotes point to Anthropic's ASL-3 Deployment Standard and OpenAI's Preparedness Framework.

Frontier AI here means mainly large language models, though the safeguards may also help with **AI-enabled Biological Tools (BTs)**, tools trained on biological data with the potential to facilitate development of hazardous biological agents. Misuse covers only severe, large-scale catastrophic scenarios. Societal safeguards are the ones outside the model: physical controls, supply chain security, regulatory compliance, coordination. Left out: protecting the safeguards and the models themselves from compromise, such as

jailbreak prevention, and organisational measures such as safety culture and risk governance.

The hedges come before the tables: safeguards are an active research area, promising techniques may be replaced, efficacy varies with the model and with how it is released, and the brief makes no claim about an ideal combination, because measuring resilience to novel adversarial approaches is itself an active research area. The list is "not intended to be prescriptive".

Five functions, three places

Capability limitation alters weights or training so the model never holds the knowledge: common for LLMs, increasingly common for BTs via data filtering. **Behavioural alignment** shapes responses so the capability is not elicited: common for LLMs, only theorised for BTs, because what counts as a dangerous output there is harder to define. **Detection and intervention** automatically spots inputs and outputs that may lead to undesired behaviour, again only theorised for BTs. **Access control** governs who uses the model and how it talks to external systems, widely used for both, with precedents in human genomics. **Supporting ecosystem mitigations** are where developers hand governments, organisations and civil society the tools to run their own defences.

Safeguards also sit **model-level** (training, fine-tuning, alignment), **system-level** (the deployment layer, filtering without touching parameters), or **societal-level**.

Table 1: capability limitation

Data filtering strips training data that could produce dual-use or harmful capabilities, using automated classifiers, keyword filters and models trained to spot subtler patterns: in biology, pathogen engineering protocols and weaponisation procedures. The same knowledge about biological sequences and pathogen research is what lets a model help with vaccine development and disease understanding, and capabilities of concern can emerge from combinations of seemingly benign data through unexpected inferences.

Targeted unlearning, experimental, pulls specific knowledge out after training, for example virology while keeping general biology. The brief says it may be reversible with relatively modest effort, has been shown to lack robustness, can be recovered through specific prompting or targeted fine-tuning with small datasets, and that models may regenerate removed knowledge from adjacent information still present.

Model distillation, experimental, builds narrow models, for example one that excels at medical diagnosis but lacks what bioweapons development

needs; it is unclear how well this stops capabilities being reconstructed, and many use cases mean many models and higher cost.

False learning, experimental, trains the model on deliberately fabricated but plausible-sounding wrong information about dangerous procedures, so a misuser gets instructions that make the attempt ineffective or unsafe for them; side effects on reliability are unpredictable and it could mislead legitimate researchers.

Table 2: behavioural alignment

Refusal training is three techniques: supervised fine-tuning on curated refusals plus helpful but harmless answers; reinforcement learning from human feedback, using human preferences between outputs as the reward signal; and adversarial training, marked experimental, on examples crafted to make the model fail, found through human and automated red-teaming and analysis of real-world jailbreaks. In biology: refusing requests about building bioweapons, acquiring dangerous pathogens or bypassing biosecurity protocols.

The limits answer the page's second guiding question. Safety training modifies surface-level behaviours without altering the underlying capabilities, so targeted fine-tuning can undo it, in biology by fine-tuning on material such as virology. Jailbreaks bypass

alignment post-training. And translating human values into training objectives brings reward hacking, alignment faking and goal misgeneralisation.

Reinforcement learning from AI-assisted feedback (RLAIF) uses AI to generate feedback from predefined principles or a constitution, a scalable alternative to purely human feedback; the brief names Anthropic's Constitutional AI and OpenAI's deliberative alignment. It depends heavily on how comprehensive those principles are, and remains vulnerable to sophisticated bypass attempts.

Table 3: monitoring and intervention

Automated input and output monitoring prompts language models to act as classifiers in real time, watching for dangerous pathogens, toxins or key steps in bioweapon development. It can be bypassed by splitting a request into benign-looking substeps or across accounts, and may produce a high rate of false positives that block legitimate scientific queries.

Custom-trained classifier models score inputs and outputs against explicit rules, with Anthropic's Constitutional Classifiers as the example; the classifier may have vulnerabilities of its own or interpret the constitution incorrectly, and it adds latency.

Manual and automated abuse monitoring watches patterns over time, catching someone systematically

assembling what they need even when no single prompt trips a filter; defining "abusive" without violating privacy or flagging ordinary research is hard. **Context-aware access control**, experimental, opens certain capabilities only to verified researchers or institutions. **Tar pitting or throttling**, experimental, detects misuse then degrades the model without telling the user, through wrong information or slower token generation; sophisticated users may notice, and degraded output could severely mislead legitimate users.

Table 4: access control

Staged deployment widens access gradually, from small groups of research partners under strict monitoring agreements outward as controls are validated; monitoring gets harder as access expands. **Sanctions screening and geogating** block users in sanctioned countries by location and IP, aimed at states suspected of pursuing illicit bioweapons programmes; VPNs defeat it, and sanctions lists often will not perfectly correlate with actual threats. **User verification protocols**, experimental, require real identity verification, know your customer, which creates accountability but adds friction, raises privacy concerns, can be spoofed with stolen identities and may exclude legitimate anonymous users. **Access to BTs**, experimental, stops an LLM acting as an

interface to specialised biological tools, though which uses are permissible is hard to define.

Table 5: ecosystem mitigations

Nucleic acid synthesis screening is DNA and RNA suppliers checking orders against curated databases of pathogenic sequences and verifying customers' institutional affiliation and stated research purpose. The brief calls it a critical chokepoint, because it stops a hazardous design on a screen becoming physical material, and says AI developers could help by sharing threat intelligence and building better screening tools. It relies on how comprehensive the databases are, and is currently not legally mandated, so not all providers screen.

User verification protocols for other services covers databases of sensitive biological materials and sellers of specialised equipment; credentials can be stolen and standards vary worldwide. **Insider threat mitigations** screen people given access to high-containment facilities, dangerous pathogens or sensitive biotechnologies, but vetting may not predict future intent and may miss collaborators, visitors or cyber infiltration aimed at insiders. **Oversight of automated bio-platforms** means licensing, audits or regulation for commercial "cloud labs" and other automated platforms; regulation lags the technology, and illicit actors may seek unregulated platforms or build their own. **Export controls** restrict cross-border

transfer of pathogens, toxins, genetic elements, dual-use equipment and sensitive technical data; intangible transfer of expertise and designs is exceptionally difficult to control, and controls only touch what crosses a border. **Secure information-sharing networks** carry threat intelligence, including AI misuse techniques observed in the wild; building trust across sectors and borders is a significant hurdle.

Why technical safeguards alone will not do

Three reasons. First, no single safeguard covers all risk vectors, so the answer is several imperfect guardrails with uncorrelated failure modes that fail independently of each other. Second, proliferation: as models get cheaper and more efficient, the resources needed to reach a given capability level fall, so capabilities spread, and bespoke safeguards on proprietary models may become obsolete once other developers train comparable models with none. Third, current techniques have limitations and their effectiveness varies widely, and refusal training in particular may be less robust for open-weight models or models reachable through fine-tuning APIs (Qi et al., ICLR 2025). The conclusion: a coordinated stance across pre-deployment and post-deployment phases, with no single point of failure.

Emerging practices, and the conclusion

Configurations vary, but refusal training and automated input and output monitoring are the commonly used techniques. Effective risk reduction usually takes several safeguards layered together, aimed at specific, plausible threat scenarios in specific deployment contexts, informed by structured risk assessments. Safeguards must then be verified as actually reducing risk to an acceptable level, tested alone or in combination, weighing cost, transparency, performance, research and privacy (NIST AI 800-1, "Managing Misuse Risk for Dual-Use Foundation Models", 2024).

One trade-off is left open: monitoring may complement output refusals or substitute for them, because refusing outright may push an actor to diversify their attack strategy and make the misuse less visible. That is the reasoning behind tar pitting. The brief closes on three future directions: safeguards robust against evolving threats, including agentic AI systems that use specialised biological tools; standardised evaluation methodologies; and better understanding of what safeguards cost in model performance.

Words to know

Enabling capability threshold. The point at which a model's measured abilities could enable extreme harms if it were deployed with no extra safeguards. It is the trigger that switches the mitigations on, so

where the threshold sits decides when anything happens.

AI-enabled Biological Tools (BTs). Models trained on biological data that can give insights, predictions and designs about biological systems, with the potential to help develop hazardous agents. They matter because most safeguards in this brief were built for language models, and for BTs several are only theorised.

Defence in depth. Stacking several imperfect guardrails whose failure modes are uncorrelated, so a failure in one does not cause a failure in the next. This is the brief's whole answer, and it is the same Swiss cheese idea the unit opens with.

Targeted unlearning. Removing specific knowledge from a model after it has been trained, instead of retraining it. Worth knowing because the brief says it has been shown to lack robustness and may be reversible with relatively modest effort.

Jailbreak. An adversarial prompt that gets a model to do what its safety training says it should refuse. It is why behavioural alignment is treated as a surface layer rather than a real barrier.

Tar pitting or throttling. Detecting misuse and then degrading the response, in speed or quality, without the user knowing they have been blocked. The design

goal is that the actor keeps wasting time instead of switching tactics.

Geogating. Blocking access by checking a user's location or IP address against sanctions lists. Useful shorthand for the whole class of controls a VPN can defeat.

Nucleic acid synthesis screening. DNA and RNA suppliers checking orders against databases of pathogenic sequences and verifying the customer. The brief calls it the chokepoint between a digital design and physical material, and notes it is not currently legally mandated.

Numbers in this reading

Five functional categories of safeguard: capability limitation, behavioural alignment, detection and intervention, access control, supporting ecosystem mitigations.

Three modes of application: model-level, system-level, societal-level.

Five tables, holding four capability-limitation safeguards, two behavioural-alignment safeguards, five monitoring and intervention safeguards, four access-control safeguards and six ecosystem safeguards.

Three principal reasons given for needing both technical and societal safeguards: uncorrelated failure modes, proliferation as models get cheaper, and the measured limitations of current techniques.

Four emerging practices listed as foundational.

Three named future research directions in the conclusion.

The brief gives no effectiveness percentages, no cost figures and no evaluation scores anywhere, which is itself the point: it says measuring safeguard resilience is still an active research area.

What the author is unsure about

Almost everything, stated openly. The taxonomy is "preliminary" and "not intended to be prescriptive". AI safeguards are an active area of research and techniques that currently seem promising may be replaced. The appropriateness and efficacy of any measure may vary with the model and with how it is released. Many of the mitigations merit additional research into whether they work.

Most importantly, the brief refuses to say which combination is best, because measuring resilience to various and novel adversarial approaches is an active research area. It also leaves open whether monitoring should complement or replace refusals, and calls for

standardised evaluation methodologies that do not yet exist.

On individual techniques it is specific about what is not known: it remains unclear how effectively model distillation prevents harmful capabilities being reconstructed, false learning may have unpredictable side effects, and unlearning has been shown to lack robustness. Several safeguards are marked as experimental or proposed rather than deployed, and several are noted as appearing in some member firms' frameworks but not all.

For the discussion

Every safeguard in this brief arrives with a limitations column, and for the model-level ones the limitation is usually the same: the technique changes the surface while the capability stays in the weights. So the honest answer to "which denies most robustly" may be the one the brief gives: none alone, and the combination has not been measured yet.

Tar pitting is the one worth arguing about, because a safeguard designed so the user cannot tell they have been blocked is also a safeguard the user cannot appeal, and the brief admits degraded outputs could severely mislead legitimate users if applied incorrectly.

She submitted a public comment to the U.S. Food and Drug Administration on generative AI, in her

individual capacity, asking for monitoring after approval, version tracing and a way to roll back, which is the same post-deployment half of the "coordinated stance across pre-deployment and post-deployment phases" this brief ends on.

2.6-2 · CORE · COURSE TIME 10 MIN

AIxBio Research Hub

2025

Read the original paper → · the course assigns: the whole site, four pages, read in the browser on 22 Oct 2026

> **This is not an article but a working library: a curated, continuously updated index of the research, people and organisations at the intersection of AI and biology, built so a newcomer can get oriented fast and then follow the field.**

Average summary

Summarized from a detailed structured account of the site, not its original text, so nothing here is a direct quotation. The account covers all four of the site's pages, read on 22 October 2026. The footer carries a 2026 copyright line for the AIxBio Research Hub, and the site gives no author or host organisation anywhere on it.

The **AIxBio Research Hub** describes itself as a platform for researchers, professionals and policymakers to find and stay up to date with the latest research on the implications and risks at the intersection of artificial intelligence and biology. It claims to put that whole research landscape in one place, and labels the collection curated and updated continuously.

The home page is a four-card index, numbered, each card with an OPEN button. **01 Browse and filter** goes to the Research Database, a structured, searchable catalogue of papers, reports and preprints across the AI x biology risk landscape. **02 Start here** goes to Recommended Resources, a hand-picked set of primers, explainers and key readings for getting oriented quickly. **03 Follow the field** goes to Organisations and Newsletters, the labs, institutes and newsletters shaping the conversation. **04 Explore visually** leaves the site for a shared interactive citation map hosted on Litmaps, for tracing how ideas connect across the literature.

The recommended resources page is the one to read first. It says the selection is for people new to the AIxBio field and builds on the list developed by ERA for their AIxBio Research Fellowship. The page says six sections; five are in the capture: Introductions to the field, Risk Assessments and Evaluations, Policy and Governance, Technical Resources, and Research

Agendas. The same handful of organisations supply almost everything: the Centre for Long-Term Resilience, the Nuclear Threat Initiative, RAND, SecureBio, Epoch AI, the National Academies, the Frontier Model Forum and the U.S. Government. One name here is already in the course's own pile, RAND's Evaluating Safety and Security in Biological Research, and another is a Frontier Model Forum brief that is the sister of this page's main reading, on evaluations rather than mitigations.

The organisations and newsletters page lists **nine newsletters** and thirteen organisations (twelve were captured). The newsletters run from quarterly briefings to a biweekly policy playbook, and include the Pandora Report from George Mason University's Schar School biodefense programme and Health Security Decoded from the Johns Hopkins Center for Health Security. The organisations split into think tanks and research centres (CLTR, Johns Hopkins CHS, NTI | bio, RAND CAST, SecureBio, CSET, the National Security Commission on Emerging Biotechnology) and funders (Coefficient Giving, the AI Safety Fund, SentinelBio, Longview Philanthropy).

The database holds **288 entries** with search, sorting and filters by item type, tag, institution and year. The newest entries run to October 2026, and the oldest single item is from 2011.

Why it's assigned

Unit 2 is the **Delay** layer, and this page is the Deny deep dive: model-level safeguards. The course note calls this hub a platform for finding and staying up to date with research on the implications and risks at the AI and biology intersection, which makes it the one reading on the page that is a tool rather than an argument. For a deep dive that has to be researched from scratch, it is where the search starts, and its Start Here list is the fastest route to the five named topics (pre-training filtering, fine-tuning and weight locking, output filtering, API control, proof of personhood).

In depth

What this is, and what the capture includes

This is written from a capture of the whole site made on 22 October 2026. The site has four pages plus one external map, and all four were read. Not in it: the Litmaps citation map, a third-party app that renders as a canvas rather than text, and the 250-plus older database entries beyond the recent run. Entries came through as title, item type, author and date, without abstracts.

The masthead reads AIxBio Research Hub with four navigation links: Database, Resources, Organisations, Litmaps. Its positioning line pairs AI with biology above the words research hub. Nothing on the site names an author, an editor or a host institution.

The home page: four doors

The home page is a numbered four-card index, each card with a kicker and an OPEN button.

01 BROWSE & FILTER, Research Database. A structured, searchable catalogue of papers, reports and preprints across the AI x biology risk landscape.

02 START HERE, Recommended Resources. Primers, explainers and key readings for getting oriented quickly.

03 FOLLOW THE FIELD, Organisations & Newsletters. The labs, institutes and newsletters shaping the conversation, and the ones worth subscribing to.

04 EXPLORE VISUALLY, Litmaps. An interactive citation map for tracing how ideas connect, and an outbound link to a shared map on app.litmaps.com.

Start here: the orientation list

The page says these resources are suggested for people new to the AIxBio field, and that the selection builds on the list developed by **ERA** for their AIxBio Research Fellowship. It says six sections; five are in the capture.

Introductions to the field. Understanding AI-Facilitated Biological Weapon Development, Centre for Long-Term Resilience, an overview of how AI

capabilities intersect with biological risks. The Convergence of Artificial Intelligence and the Life Sciences, Nuclear Threat Initiative, called a foundational paper on AI-bio convergence and dual-use challenges. The near-term impact of AI on biological misuse, CLTR. Capability-Based Risk Assessment for AI-Enabled Biological Tools, CLTR. Developing Guardrails for AI Biodesign Tools, NTI.

*Risk Assessments and Evaluation*s, frameworks and methodologies for assessing AI-bio risks. Global Risk Index for AI-enabled Biological Tools, CLTR and RAND. Why we recommend risk assessments over evaluations for AI-enabled biological tools, CLTR. The Virology Capability Test (VCT), SecureBio. Assessing the Impacts of Technology Maturity and Diffusion on Malicious Biological Agent Development Capabilities, RAND. Do the biorisk evaluations of AI labs actually measure the risk of developing bioweapons?, Epoch AI. Toward Comprehensive Benchmarking of the Biological Knowledge of Frontier Large Language Models, RAND.

Policy and Governance, key policy documents and recommendations. Biosecurity Governance Across Uncertain Artificial Intelligence Futures, RAND. How the UK Government should address the misuse risk from AI-enabled biological tools, CLTR. Biological Tools and the EU AI Act, CLTR. America's AI Action Plan, U.S. Government.

Technical Resources, deep dives into capabilities and methods. The Age of AI in the Life Sciences, National Academies of Sciences, Engineering, and Medicine. Evaluating Safety and Security in Biological Research, RAND. SecureBio AI x Biosecurity Resources. Issue Brief: Preliminary Taxonomy of AI-Bio Safety Evaluations, Frontier Model Forum. Secure Bio Blog.

Research Agendas, strategic priorities. One item: the NTI AI-Bio Research Agenda.

Worth noticing: the Frontier Model Forum item here is a sister brief to the main reading on this page, on evaluations rather than mitigations.

Follow the field: nine newsletters

The **AIxBIO Briefing** is quarterly, covering key research, policy and risks at the intersection of AI and biosecurity. **Securing the Interface** is written by Richard Moulange, described as a Cambridge-trained AI researcher turned security policy advisor, with analysis of AI's security implications and early warning on emerging vulnerabilities; the blurb says he advised the UK AI Security Institute in its earliest days and contributed internationally through the NTI Global AIxBio Forum. The **Pandora Report** is a regular newsletter on biodefense produced by students and faculty of the George Mason University Schar School of Policy and Government Biodefense program. **Health Security Decoded** is a monthly

roundup from the Johns Hopkins Center for Health Security, tracking news, events, research and policy across biosecurity and biodefense, preparedness and response, public health, national security and emerging diseases. The **Nuclear Threat Initiative** newsletter carries updates on NTI's work reducing nuclear and biological threats. **GCBR Organization Updates** collects updates from organisations working to reduce global catastrophic biological risks; it is an independent project led by Tessa Alexanian and Anemone Franz outside their organisational affiliations, published every few months. **Global Biodefense** covers pathogens and preparedness for public health emergencies. The **Pandemic Action Playbook** is biweekly, from the Pandemic Action Network, carrying policy developments, insights, actions, events, opinions and resources. The **Council on Strategic Risks** newsletter covers interconnected and emerging risks, technology and biosecurity.

Follow the field: the organisations and the funders

The heading says thirteen; twelve entries were captured. The research and policy side: the **Centre for Long-Term Resilience (CLTR)**, a UK nonprofit think tank developing policy tools for monitoring and mitigating extreme threats, including at the AI-biotechnology intersection. The **Johns Hopkins Center for Health Security (CHS)**, a nonprofit think tank inside Johns Hopkins working on biological and

health security, with a large amount of AIxBio work. **NTI | bio**, the biosecurity division of the Nuclear Threat Initiative. **RAND Center on AI, Security, and Technology (RAND CAST)**, studying how high-consequence dual-use technologies change the global threat environment. **SecureBio**, a biosecurity nonprofit founded by Kevin Esvelt. The **Council for Strategic Risk (CSR)**, a non-partisan policy institute on systemic security risks. The **Center for Security and Emerging Technology (CSET)**, in Georgetown's Walsh School of Foreign Service, giving decision-makers data-driven analysis of the security implications of emerging technologies. The **National Security Commission on Emerging Biotechnology**, examining how emerging biotechnology will shape national defense activities, including the Department of Defense's.

The money side. **Coefficient Giving** is described as one of the largest funders of both AI safety and biosecurity, supporting policy, technical research, institutions at the AI-bio intersection, and talent development. The **AI Safety Fund (AISF)** is a non-governmental grantmaker supporting technical alignment and misuse prevention, with growing attention to dual-use risks of AI in biological research. **SentinelBio** is a nonprofit philanthropic fund focused on eradicating pandemics, including responsible development of biosecurity tools and AI-bio convergence. **Longview Philanthropy**, through its

Emerging Challenges Fund, aims to fill critical funding gaps at the nexus of AI, biosecurity and nuclear policy, with rapid support for emerging organisations.

The database: 288 entries and what they are about

The research page has a search box, a sort control (Best match, Newest first, Oldest first, Title A-Z), a count reading **288 ENTRIES**, and a Clear All Filters control. Each result card shows title, item type, authors, publication venue where applicable, date, an optional tag row, a Read Abstract toggle and a View link out to the source.

The filter counts shown on 22 October 2026 tell you what kind of field this is. By item type, blog posts (62) and web pages (60) outnumber journal articles (53), with 53 reports and 46 preprints: this literature lives in blogs, think-tank reports and preprints rather than in journals. By tag, Capabilities and Threat Assessment tie at 69 each, ahead of Risk assessments (58) and Evaluations (56), with Mitigations at 40, so assessment outweighs mitigation. By institution, RAND leads with 13 and CLTR has 7. By year, 2025 has 96, 2026 and 2024 have 71 each, 2023 has 30, and almost the entire literature was written in the last three years.

What the newest entries are arguing about

The newest items in the default listing, read as a group, show four live arguments.

Whether the risk is real and how large. Two blog posts by Christopher Rodriguez, both dated 2026-07-28, land on opposite sides: "Why bioterror is actually rare" and "Bioterrorism is actually a problem right now". "Why AI-assisted bioweapons won't kill us all" (Abi Olvera, 2026-07-21) and "Bioweapons as the Optimist's Exception" (O. H. Scharfman, 2026-07-24) argue the sceptical case, while Elizabeth Sherwood-Randall's "AI and the New Age of Bioweapons" ran in Foreign Affairs on 2026-08-18 and "AI Chatbots Know How to Make Deadly Biological Weapons. Some Will Teach You." (Wells and Ramkumar) is a newspaper article dated 2026-07-26.

Whether the benchmarks measure anything. "AI Biosecurity Benchmarks and Real-World Risk" (Stephen Turner, preprint, 2026-08-27), "Interpreting Dual-Use Biology Benchmarks for Frontier AI: Measurement, Saturation, and Generational Analysis" (Grant Ellison and colleagues, 2026-08-06), and "A Threshold Exceedance Framework for CBRN Uplift Evaluation in Frontier Language Models" (arXiv, eleven authors, 2026-07-13).

Agents using biological tools, the newest worry, with four 2026 entries: "Can LLM Agents Select and Engage with Biological Tools?" (2026-06-25), "Testing Large Language Model Agents on the Use of

Biological Tools for Nucleic Acid Synthesis Screening Evasion" (2026-08-11), "Restricting AI Agent Use of Biological Tools" (2026-07-20), and "Frontier AI Agents and Biological Tools" (Frontier Model Forum, 2026-07-31).

Governance and what safeguards cost. The site's newest entry is "Managing Biological Risks from General-Purpose AI Models under the EU AI Act", a GovAI report dated 2026-10-19 with eleven authors including Cassidy Nelson and John Halstead. Alongside it: "Building a Defense-in-Depth Biosecurity Strategy for the AI Era" (2026-08-18), "A Biosafety Agenda for AI-Bio Governance" (Bhuvana Sudarshan and Rebecca Hersman, 2026-08-25), and the one that cuts the other way, "When AI Biosafety Safeguards Block Public Health" (Larry Kerr, 2026-08-06).

Two more worth knowing by name. "Developing a standard definition for sequences of concern", a twenty-five-author article in Frontiers in Bioengineering and Biotechnology (2026-06-04), includes Tessa Alexanian, Kevin Esvelt, James Diggans and Nicole E. Wheeler, and is the screening-side foundation behind the synthesis reading on the other deep dive. And OpenAI's own "Biodefense in the Intelligence Age" is dated 2026-07-22.

Words to know

Preprint. A paper published by its authors before, or without, peer review. It matters here because 46 of the 288 entries are preprints and 60 are web pages: this field's thinking moves faster than journals do.

Citation map. A picture of which papers cite which, so you can see clusters and the lines between them instead of a list. The hub's fourth door is one of these, hosted on Litmaps.

Dual-use research. Work that can be used to protect people or to harm them, with no clean line between the two. It is a tag on 26 entries and the reason this whole field exists.

GCBR, global catastrophic biological risk. A biological event large enough to threaten civilisation rather than one country or one hospital system. One of the listed newsletters is named after the organisations working on it.

Facet. A filter that also tells you how many items it would return, the way the preprint filter carries its count of 46. Reading the facets is the fastest way to see a field's shape without reading a single paper.

Numbers in this reading

288 entries in the research database, the count shown on the page on 22 October 2026.

Item types: Blog post 62, Webpage 60, Journal article 53, Report 53, Preprint 46, Magazine article 4, Newspaper article 4, Book 3, Book section 1.

Tags, top ten of 43 (the other 33 sat behind a show-more control): Capabilities 69, Threat Assessment 69, Risk assessments 58, Evaluations 56, Policy 54, Mitigations 40, Biological Tools 39, Dual-use research 26, LLMs 26, Framework 24.

Institutions, eight shown of 19 (11 collapsed): RAND 13, CLTR 7, GovAI 3, NTI 3, AI Safety Summit 1, Center for AI Safety 1, CSET 1, CNAS 1.

Years: 2026 has 71, 2025 has 96, 2024 has 71, 2023 has 30, 2022 has 2, 2011 has 1.

Nine newsletters listed. Thirteen organisations claimed in the heading, twelve captured.

Four pages on the site, plus one external citation map.

What the author is unsure about

Nothing, because the site makes no argument. It is a directory, and the only claim it makes about itself is that the collection is curated and updated continuously.

The open questions are editorial, and the site does not answer them. It names no author, editor or host organisation. It does not say who decides what gets

in, or how often the collection is actually updated. The one piece of provenance it gives is that the Start Here list builds on a list developed by ERA for their AIxBio Research Fellowship, so the orientation reading is inherited rather than independently assembled.

Two things to hold lightly. The filter counts are a snapshot from one day, 22 October 2026, so they will move. And a hub like this tells you what a field is publishing, not what is true: a blog post and a Science paper sit in the same list, with the item type as the only signal.

For the discussion

RAND leads the institution filter with 13 entries and the Centre for Long-Term Resilience has 7, more than anyone else listed, and the Start Here list leans on those same two names, so a few think tanks are shaping how this field frames itself.

The tag counts show Capabilities at 69 and Threat Assessment at 69 against Mitigations at 40, so the field is still spending more effort describing the danger than building the denial layer this deep dive is about.

The one entry pointing the other way, "When AI Biosafety Safeguards Block Public Health" (Larry Kerr, 2026-08-06), is the question worth asking the group: who pays the cost when a safeguard refuses the

wrong person, and does anyone in this literature measure it?

2.6-3 · FURTHER

Deep Ignorance: Filtering Pretraining Data Builds Tamper-Resistant Safeguards into Open-Weight LLMs

Kyle O'Brien, Stephen Casper et al. · 2025

Read the original paper → · the course assigns: the project site's own summary of the paper, not the paper itself

> **If you strip the risky material out of an open-weight model's training data before you train it, the model scores near random chance on biorisk knowledge, keeps its general ability, and resists later attempts to fine-tune that knowledge back in, though nothing the team tested survived a combined attack.**

Average summary

This is the project site for the paper, which is its own short summary of the work, not the paper itself. The site's own figure caption and filtering description are included; the full results live in the paper, arXiv:2508.06601. The site says the paper was accepted to **ICLR 2026** (the International Conference on Learning Representations, a machine learning conference: not from this reading).

Ten authors, from **EleutherAI**, the **UK AISI** (the UK's AI Security Institute), the University of Oxford and the ERA Fellowship. Kyle O'Brien and Stephen Casper are marked equal contributors; the others are Quentin Anthony, Tomek Korbak, Robert Kirk, Xander Davies, Ishan Mishra, Geoffrey Irving, Yarin Gal and Stella Biderman.

The problem they start from. As LLMs get more capable, a risk is that they learn unsafe knowledge during training. Today's safeguards suppress that knowledge in post-training, usually through refusal training, which means the unsafe knowledge stays in the model's weights. Open-weight models, the ones anyone can download and modify locally, bring real benefits in transparency, research and the deconcentration of power, but they are vulnerable to tampering attacks that strip the safety training out. As the site puts it, it is straightforward to train an open-weight model never to refuse unsafe requests, and that matters more now that open-weight models are starting to rival the best closed-weight ones. The site quotes the International AI Safety Report (Bengio et al., 2025): once weights are available for public download, there is no way to implement a wholesale rollback of all existing copies.

Their question, which they call intuitive yet understudied: can you stop an LLM learning unsafe technical capabilities by filtering out enough of the

relevant pretraining data before training starts? They trained multiple **6.9B** LLMs from scratch, one on an unfiltered dataset and others on filtered versions with biorisk knowledge removed, using biothreat-proxy knowledge as the representative example.

Three results. **Knowledge prevention**: the filtered models do significantly worse on the biorisk knowledge evaluations, nearly at random chance, with no notable regressions in general knowledge. **Tamper-resistance**: fine-tuning the filtered models on 300M tokens of high-quality biorisk documents does improve their performance, but leaves it well below the no-filtering baseline, which makes filtering significantly more tamper-resistant than current safeguards. **Defence in depth**: filtering cannot stop a model using harmful knowledge handed to it in the prompt, Circuit-Breaking techniques offer a complementary defence, and none of the defences they tested resisted staged attacks that combine fine-tuning with in-context retrieval.

The filtering itself is cheap: less than a 1% increase in training compute. The team released the models, optimizer states and intermediate checkpoints publicly.

Why it's assigned

Unit 2 is the **Delay** layer, and this page is its Deny deep dive. Pre-training filtering is the first topic the

course page lists, and this is the one reading on the page that tests it directly, by training models with and without the filter, rather than cataloguing it. It also answers the page's third guiding question, what has failed and why, from the other end: refusal training is the approach that fails for open weights, because it never removes the capability, and this paper's whole premise is that you have to not teach it in the first place.

Words to know

Open-weight model. A model whose parameters anyone can download and change on their own machine. The site's point is that you cannot recall it: once the weights are out, there is no rollback.

Tampering attack. Fine-tuning a downloaded model to undo its safety training, for example training it never to refuse. This is the attack the paper is defending against, and it is described as straightforward.

Pretraining data filtering. Removing material from the training corpus before training begins, rather than teaching the finished model to decline. The difference is whether the capability exists at all.

Blocklist, then classifier. The two-stage pipeline: every document is checked against a list of prohibited terms, and only documents containing a blocked term, such as "pathogen(s)", go to a fine-tuned text

classifier that scores their content and excludes those above a threshold. In practice the vast majority of documents pass at the blocklist stage, which is why the cost stays low.

In-context. Knowledge supplied in the prompt rather than learned in training. It is the hole in this defence: filtering stops the model knowing, not the model reading.

Circuit breaking and latent adversarial training. Two post-training defences the team layered on top of filtering; their strongest released combination is labelled "maximum defense".

Numbers in this reading

6.9B parameters: the size of each model they trained from scratch, unfiltered and filtered.

Nearly random chance: how the filtered models score on the biorisk knowledge evaluations.

300M tokens of high-quality biorisk-related documents: the adversarial fine-tuning attack they ran against the filtered models, and the amount of attack their safeguards resist.

Less than a 1% increase in training compute (FLOPS): what the filtering cost.

8.42% of documents removed: the strong filter, which is blocklist only, in the released artifacts table.

500B tokens: the size of the released unfiltered pretraining checkpoint.

arXiv:2508.06601, 2025: the preprint, accepted to ICLR 2026.

What the author is unsure about

The authors are careful about the size of their claim. They say the results "suggest" that data filtering "may be" a simple way to prevent dangerous capabilities, and call rigorous pretraining data filtering "promising", not solved.

They state three limits plainly. Filtered models fine-tuned on biorisk data do improve, so this is resistance, not immunity. Filtering cannot prevent a model leveraging harmful knowledge provided in-context. And none of the defences they tested, including circuit breaking on top of filtering, resisted staged attacks that combine fine-tuning and in-context retrieval.

They also name the gaps they want others to fill: filtering other types of knowledge beyond biorisk, and developing scaling trends, which is to say nobody yet knows whether this holds for models much larger than 6.9B. One more thing to hold: this page is the team's own summary of their own work, and the evaluations behind "nearly at random chance" are in the paper, not here.

For the discussion

The cleanest framing in the whole unit is here: refusal training teaches a model to decline, filtering means there is nothing to decline, and for weights that can be downloaded and edited only the second one survives contact with an attacker.

Less than a 1% increase in training compute is the number to raise, because it removes the usual objection that denial-by-design is too expensive, and leaves the real objection, which is deciding what to cut out of a corpus.

The honest limit is worth saying out loud too: filtering stops the model knowing, not the model reading, and the team's own staged attack, fine-tuning plus in-context retrieval, broke every defence they tried.