



PART 1

AI, Terrorism, and WMD: The Risk Is in the Barriers It Lowers

The useful question isn't whether AI can build a weapon. It's which barriers between intent and capability AI can lower, for whom, and how fast.

AUTHOR

Chris Serdinak

AFFILIATION

Co-Founder, Helix Stratus Group

PUBLISHED

September 2026

For more than two decades I worked terrorism cases in the United States and around the world. The work taught me to take new capabilities seriously without confusing what is theoretically possible with what an adversary can actually pull off.

There is real reason for concern about threat actors using artificial intelligence. But most of the public debate about AI, terrorism, and weapons of mass destruction is stuck on the wrong question. One camp asks whether a model can design and build a weapon on its own, and when the answer is no, calls the risk hype. The other camp watches a model produce something disturbing and concludes catastrophe is close. Both are staring at the finish line, when those of us who work this problem should be studying the course.

CENTRAL QUESTION

Which barriers between intent and real-world capability can AI lower, for whom, by how much, and which barriers still hold?

That framing avoids both complacency and panic. It also gives policymakers, labs, technology companies, intelligence professionals, and first responders something they can actually measure.

1. THE EVIDENCE HAS MOVED

In early 2024 the evidence was reassuring, and appropriately cautious. A RAND red-team study found no statistically significant difference between biological-attack plans produced with the internet alone and plans produced with help from large language models. An OpenAI study of 100 biology experts and students found at most a mild uplift, which did not hold up as statistically significant once the researchers corrected for multiple comparisons. Both teams stressed that access to information is not the same as building something in the physical world.

Those results were a reason for discipline, not a reason to stop watching. By the end of 2025, the UK AI Security Institute was describing a different picture. Its *Frontier AI Trends Report* found that leading models had exceeded PhD-level biology experts on its knowledge tests, with chemistry quickly catching up. Models could write protocols judged feasible for selected lab work and were getting better at troubleshooting. In one study, non-experts using frontier models had 4.7 times higher odds of writing a feasible protocol for a viral-recovery task than people using the internet alone.

EARLY 2024

RAND and OpenAI studies: no statistically significant uplift from LLMs over internet-only research.

END OF 2025

UK AISI: models exceed PhD-level biology experts on knowledge tests; non-experts show 4.7× higher odds of a feasible viral-recovery protocol.

None of that means a terrorist can ask a chatbot for a weapon and get one. What it shows is narrower, and still important: parts of the expertise, planning, and troubleshooting barriers are getting easier to cross.

The 2024 and 2025 findings don't necessarily contradict each other. They tested different models, tasks, and people, and above all the technology changed in between. In AI terms, a 2024 study isn't two years old; it may be several model generations behind what's available today.

The *International AI Safety Report 2026* calls this the “evidence dilemma”: capabilities improve quickly while reliable evidence about the risks arrives slowly. A serious human-subjects study has to be designed, approved, funded, run, reviewed, and published, and by the time it appears the model it tested may have been replaced several times. That doesn't make formal research less valuable. It means periodic studies can't be our only warning system.



2. OPEN-WEIGHT MODELS CHANGE THE EQUATION

Commercial frontier models are only half the story. The other half is the explosion of downloadable models that anyone can run on their own hardware. People often call them “open-source,” but “open-weight” is more accurate: the trained model is released so it can be downloaded, run locally, modified, fine-tuned, and built into other systems.

That distinction matters for security. A model reached through a commercial provider can be monitored, rate-limited, updated, or shut off, and the provider can add classifiers, refusals, account controls, and abuse monitoring. Once capable weights are released, none of that is enforceable. The model can be run offline, stripped of its safeguards, trained on specialized data, and wired to other tools without the developer ever knowing.

Hardware requirements vary widely. Large general-purpose models still need serious memory, accelerators, and power. Smaller models, though, can be compressed or fine-tuned for one narrow job, and a specialty model doesn't have to beat a frontier system at everything to matter. It only has to do its one task well. NVIDIA researchers have argued that small language models are already capable and cheap enough for many specialized agentic tasks, and many of them run on a single GPU. Counting data centers won't tell us where the risk is. The model that matters may run on hardware a small business, a lab, or an individual already owns.

The spread also compounds. Each released model can be copied, compressed, fine-tuned, repackaged, and built into thousands of downstream tools, and those tools produce derivatives of their own.

In July 2026 the UK AI Security Institute reported that the best open-weight models it tested on cyber tasks performed about as well as closed frontier models released four to seven months earlier. AISI was careful to note that it tested cyber only and that no inferences about other capabilities should be drawn. Still, the pattern deserves attention: the gap between a capability appearing at the controlled frontier and becoming freely available without those controls may now be measured in months.

Self-hosting changes the intelligence problem, too. There may be no suspicious API request, no commercial account, no provider alert, no central log. The activity that matters most may happen on private hardware running a model that no longer talks to its maker, and an annual threat assessment can't keep pace with that.

3. OPEN MODELS ARE AN AMERICAN ADVANTAGE, AND A RESPONSIBILITY

None of this makes open-weight AI a threat in itself. It is essential to American innovation, research, competition, and economic growth. Open models let businesses customize systems, protect their own data, avoid lock-in to a single provider, and serve markets the largest companies will never reach. They let researchers study how models behave, and they give small American companies a way to compete. We shouldn't give that up through broad restrictions.

Supporting open innovation doesn't mean every capability should be released in every form regardless of consequence, though. I'd propose a set of capability tripwires for models that show a credible ability to help design, optimize, troubleshoot, or automate WMD-relevant work. The triggers should be demonstrated behavior, not model size, political pressure, or speculation:

BOX 1 • PROPOSED CAPABILITY TRIPWIRES	
T1	A validated jump in the model's ability to produce consequential technical designs
T2	The ability to walk users through repeated experimental failure
T3	Working integration with lab instruments or autonomous systems
T4	A measurable drop in the expertise needed to complete a high-consequence task
T5	Safeguards that can be removed without reducing the underlying capability
T6	Any of the above on hardware cheap enough for private use

Crossing a tripwire should trigger closer evaluation, joint government-industry review, staged release decisions, incident sharing, and stronger protection of the physical chokepoints that remain. It should not mean automatic prohibition.

4. THE BARRIER CHAIN

A technical paper is not a weapon, and a detailed answer from a model is not a successful experiment. But information shouldn't be dismissed just because it isn't enough on its own. Real-world capability is built by crossing a chain of barriers:

TABLE 1 • FROM INTENT TO CAPABILITY: TEN BARRIERS

NO.	BARRIER	DESCRIPTION
01	Intent	Willingness to pursue mass harm
02	Expertise	Command of several specialized fields
03	Information	Finding, judging, and combining scattered knowledge
04	Tacit knowledge	What practitioners learn only through failure and repetition
05	Materials	Controlled, scarce, hazardous, or traceable inputs
06	Equipment	Suitable facilities and tools
07	Troubleshooting	Recognizing failure, diagnosing it, adapting
08	Logistics	People, money, time, and coordination
09	Security	Avoiding detection and intervention
10	Deployment	Turning a working capability into an effective attack

These barriers aren’t equal, and they look very different across chemical, biological, radiological, and nuclear threats. AI can’t manufacture scarce material or repeal physics, and in some domains physical access and industrial capacity will remain the dominant constraints.

Risk doesn’t require every barrier to fall, however. It rises when two or three critical barriers get cheaper or easier, especially for someone who has already cleared the others. That’s why a single label like “AI-enabled WMD risk” is too blunt for serious analysis. The better unit of measure is the individual barrier: what changed, for which user, under what conditions, and what still stands in the way.

5. AI IS A TEACHER, AND THE OPERATOR MATTERS

Calling AI a faster search engine undersells it. A search engine retrieves documents. A capable model can synthesize them, translate technical language, read images, write code, spot missing steps, critique a plan, and keep troubleshooting through round after round of failure.

It can also teach. A good instructor notices confusion, rephrases, offers an analogy, answers the follow-up, and breaks a process into pieces. Models now do a version of that privately, around the clock, and at enormous scale. For students, entrepreneurs, and small labs, that’s transformative. For a malicious or reckless actor, the same patience and clarity help them learn a field’s vocabulary, work through unfamiliar material, and get past the intimidation that used to keep people out.

Models still hallucinate, skip critical details, and sound confident where an expert would hesitate. In high-consequence work those flaws matter, and they can lead an unskilled user straight into failure. They also give defenders an opening, because known failure patterns can be built into evaluations, safeguards, detection, and responder training. Defenders should study where models fail as closely as where they succeed.

The relationship that matters most, though, is between the model and the person using it. A skilled operator knows how to frame the problem, ask precise follow-ups, throw out bad answers, and connect digital guidance to physical equipment. A novice takes the first answer; an expert interrogates it. OpenAI's 2024 study offered an early hint of this, since the average accuracy gain was larger for experts than for students, although the result wasn't statistically conclusive.

My read is that AI will make a capable operator faster well before it makes an amateur capable.

So the threat actor to worry about isn't always a lone person starting from zero. It may be a small network with partial expertise, an educated extremist, an insider, or a group with access to legitimate infrastructure. What the model adds has to be measured against where the actor starts.

6. WHAT AN EARLY-WARNING SYSTEM SHOULD WATCH

Trying to infer malicious intent from ordinary science is unreliable and dangerous. Most research, hiring, and procurement in sensitive fields is legitimate, and association isn't evidence of anything. A better warning system watches the barriers instead:

BOX 2 • BARRIER-BASED WARNING INDICATORS

1. Has a model shown a new capability in a credible evaluation, and how fast is it improving?
2. Is that capability confined to a controlled environment, or can it be downloaded, self-hosted, and stripped of safeguards?
3. What hardware does it need, and is that requirement falling?
4. Which barrier does it lower (teaching, design, troubleshooting, automation, cost, or access), and what does the user still need?
5. Are the physical requirements getting cheaper or more widely available?
6. Are several weak signals moving together across models, lab automation, procurement, funding, or hiring?

No single model release, paper, or purchase answers those questions. The value comes from connecting signals over time, with enough context to tell a real pattern from normal scientific and commercial activity.

The same framework tells us where to invest. If AI lowers an information barrier while material controls stay strong, reinforce screening at the material boundary. If it improves troubleshooting, strengthen lab governance and insider-risk reporting. If capability moves onto local hardware, invest in field detection and responder readiness. In each case the job is to find where our safety still rests on an assumption that is no longer true.

Controls should be targeted the same way. Blanket restrictions can look decisive while missing the barrier that actually matters, and broad surveillance of science produces false positives, erodes trust, and punishes ordinary research. Relying only on voluntary model safeguards fails in the other direction, leaving labs and local responders unready for capabilities that spread beyond any one provider. Good policy makes the dangerous pathways harder without making legitimate science impossible.

7. WARNING SYSTEMS MUST KEEP THE EVIDENCE

Speed can’t come at the cost of rigor. AI-generated threat reporting can cause the very failure it’s meant to prevent: a system summarizes a claim, drops its caveats, repeats it across reports, and slowly turns a tentative finding into accepted fact.

Every consequential judgment should keep its original source, date, version, method, and limitations, and should say plainly which of three things it is:

TABLE 2 · CATEGORIES OF ANALYTIC JUDGMENT

CATEGORY	STANDARD
Demonstrated capability	Backed by a credible evaluation, reproducible evidence, or physical validation
Analytical assessment	A reasoned judgment linking several facts, with its assumptions stated
Speculation	A plausible scenario that hasn’t been demonstrated

Those categories should never blur, and an automated severity score should never stand in for expert judgment. High-consequence alerts need review by people who understand the science, the intelligence context, adversary behavior, civil liberties, and the cost of both false alarms and missed warnings.

At Helix Stratus Group, the company I founded, we're building a project aimed at this gap: tracking how technological barriers move, preserving the underlying evidence, spotting convergence across signals that otherwise look unrelated, and putting consequential judgments in front of qualified human reviewers. We'll share more as it develops.

8. CONCLUSION

AI won't eliminate physics, logistics, scarcity, or human error. It can change who is able to cross individual barriers, how fast they move, and where our defenses are most likely to break. That is the threat worth measuring, and the advantage will go to whoever sees the barriers moving and acts before an adversary proves they've already crossed them.



NEXT · PART 2 OF 2

The lab is leaving the institution, and the first person to walk into it probably won't be a federal agent. It will be a patrol officer, a firefighter, or a paramedic. What they need, and who should build it.