Concept Paper

Dangerous Knowledge and Medical AI: Why Technical Containment Is Necessary but Insufficient for Managing Dual-Use Risk

D. John Doyle *

Posted Date: 18 September 2026

Keywords: dual-use research; AI safety; biosecurity; insider risk; medical AI governance; graded risk

 doi:10.20944/preprints202609.1616.v1

Concept Paper

Dangerous Knowledge and Medical AI: Why Technical Containment Is Necessary but Insufficient for Managing Dual-Use Risk

D. John Doyle

Emeritus Professor of Anesthesiology, Cleveland Clinic Lerner College of Medicine; djdoyle@hotmail.com or danieljohndoyle.substack.com or danieljohndoyle.com

Abstract

Artificial intelligence systems deployed across medicine—drug discovery, diagnostic imaging, genomic analysis, epidemiological modeling, and clinical decision support—are trained on domain knowledge that carries both legitimate therapeutic value and dual-use potential. Prevailing AI safety practice relies on technical containment: refusal training, output classification, red-teaming, and access controls. This paper advances a bounded thesis: technical containment is a necessary but insufficient layer for managing dual-use risk in high-consequence biomedical domains, because reliable recognition of dangerous outputs requires expertise that is itself hazardous, concentrating insider risk within the very teams tasked with safety. This is not a claim that misuse is inevitable or that containment is futile; it is a claim about the limits of any single layer. The paper adopts a graded-risk model in which computational capability, tacit knowledge, materials, and infrastructure jointly determine real-world uplift, and it argues that these barriers differ substantially across domains and tools. It surveys dual-use research governance, AI safety methods, biosecurity institutions, and medical-AI regulation to locate the gaps, and proposes an operationalized, layered governance response—including explicit deployment-gating criteria for the highest-risk applications. The paper does not present novel empirical findings and identifies the empirical questions on which its practical implications depend.

Keywords: dual-use research; AI safety; biosecurity; insider risk; medical AI governance; graded risk

1. Introduction

Machine-learning systems trained on biomedical literature, chemical libraries, medical images, and genomic data can support target identification, image interpretation, and multimodal precision medicine [1–3]. Evidence of real-world clinical benefit remains uneven and contingent on validation [4], but the scientific case for continued research is strong.

Many of these capabilities pose a classic dual-use problem: methods built for benefit can be deliberately inverted toward harm [5,6]. This dilemma predates AI. What AI can change is scale—accelerating search, synthesis, and iteration, and extending technical assistance to a wider range of users—though the magnitude of that change is contested and domain-dependent [7].

The AI safety community has responded with technical and operational controls: refusal training, output classifiers, red-teaming, access restriction, monitoring, and staged deployment [8]. These reduce low-sophistication misuse. They rest, however, on an assumption that deserves scrutiny in high-consequence science: that harmful requests and outputs can be specified and recognized reliably across novel scientific contexts.

This paper defends a deliberately bounded claim. Technical containment cannot be the sole or primary mechanism for managing dual-use risk in high-consequence biomedical domains, because competent recognition of dangerous outputs requires domain expertise that is itself hazardous. This concentrates insider risk within safety teams and creates an adaptive dynamic in which

 doi:10.20944/preprints202609.1616.v1

knowledgeable adversaries can probe and route around known safeguards. The paper does not argue that containment is worthless, that misuse is inevitable, or that any medical-AI capability is categorically uncontainable. Those stronger claims are not supported by current evidence, and conflating the sufficiency argument with an impossibility argument has weakened prior treatments of this topic.

Three conditions make the question timely: frontier systems increasingly perform expert-level scientific reasoning, though benchmark performance does not establish real-world misuse and physical-world uplift remains uncertain [9] (vendor and preprint capability reports should be read with caution given limited external auditability); medical-product regulation has matured but centers on safety, effectiveness, bias, and cybersecurity rather than deliberate biosecurity misuse;[10] and frontier capability is concentrated among a few developers whose internal evaluations are only partly visible externally.

2. A Graded-Risk Framing

Because the strength of dual-use concern varies widely, this paper adopts a graded-risk model throughout. Real-world uplift is a joint function of computational capability, tacit knowledge, physical materials, laboratory infrastructure, and operational competence. In silico proposals are not synthesizable agents; a sequence is not a pathogen; a toxic-scoring molecule is not a weapon. The 2002 chemical synthesis of poliovirus from published sequence data required substantial expertise and facilities [11], and red-team assessments of AI-assisted biological planning have found meaningful but bounded uplift [12].

This framing has two consequences. First, not every medical-AI tool is equally dual-use; general-purpose models and specialized biological design tools warrant different treatment [7]. Second, the isomorphism between defensive and offensive knowledge—developed below—is best understood as a matter of degree, strongest where recognition of a hazardous output genuinely requires the same specialized judgment needed to produce it, and weaker where material and tacit barriers dominate.

3. Dual-Use Governance and Its Limits

Formal dual-use research governance rests on a sound principle: scientifically valuable work may still require prospective risk assessment because information hazards can exist independently of intent [13]. The US framework has been in flux, and readers should treat the current federal baseline as the most recent policy in force rather than any single prior document; the relevant point for this paper is structural, not tied to a specific instrument.

Life-sciences oversight, however strengthened, is a poor fit for broadly deployed AI. Its objects are funded projects, agents, experiments, and institutions. Hazardous capability in AI is distributed across weights, training data, external tools, fine-tuning, and user interactions; open-weight systems can be copied and modified outside any developer's monitoring environment [14]. International instruments—the Biological Weapons Convention, WHO guidance, national biosecurity strategies and biosafety standards—operate through states, funders, and facilities, whereas model weights and API access cross borders rapidly and locally run models may leave little audit trail. Oversight at the point of training does not guarantee oversight at the point of use.

4. Technical Guardrails: What They Can and Cannot Do

Alignment methods distinguish outer alignment (does the objective capture intent?) from inner alignment (does the system robustly pursue it?) [15]. For medical AI the specification problem is immediate: a prohibition such as "do not facilitate biological-weapon development" must generalize across unfamiliar mechanisms, indirect requests, and legitimate neighboring research.

Value-alignment approaches such as Constitutional AI improve average behavior but depend on the completeness of their principles and evaluations [16]. Refusal behavior induced by instruction tuning and preference optimization [17] remains vulnerable to jailbreaks and adversarial attacks

 doi:10.20944/preprints202609.1616.v1

[18,19]. Output filtering and guard models catch recognizable patterns [20], but scientific outputs are context-dependent: a structure benign in one setting may be hazardous in another. Interpretability contributes evidence about model behavior but does not by itself yield judgments about hazardous biomedical capability, and post hoc explanations can mislead [21,22].

The unifying limitation is the specification problem: one cannot reliably define what is dangerous without the domain knowledge to recognize it. Evaluating whether a generated molecule is toxic requires medicinal chemistry, toxicology, and pharmacology; assessing a sequence requires genomic and virological expertise; judging an epidemiological model requires quantitative epidemiology. In each case, the evaluator qualified to defend is, to a substantial degree, qualified to offend.

5. The Core Problem: Knowledge-Based Insider Risk

To prevent dangerous outputs, an organization needs people who can recognize them—and those people thereby hold hazardous knowledge. Unlike an ordinary insider who steals credentials, this risk can be created by the safety process itself: competent evaluation of hazardous outputs requires enough expertise to distinguish meaningful results from implausible text and to understand how components combine.

Two clarifications keep this claim honest. First, its magnitude is organization-specific and should not be asserted without internal data; developers can compartmentalize expertise, restrict access, separate red teams from deployment, and log sensitive activity. Established insider-threat guidance already emphasizes governance, role separation, access control, and continuous assessment over screening alone. Second, the concern is a matter of degree consistent with the graded-risk model: it is acute where recognition and production knowledge genuinely coincide, and attenuated where tacit and material barriers remain decisive.

An adaptive-adversary dynamic compounds this. If a developer flags molecules binding known toxin targets, an adversary with domain knowledge can infer the heuristic from system behavior and route around it—requesting molecules that achieve harmful ends through unflagged mechanisms, or mining databases for known-but-unflagged compounds. This is a plausible failure mode, not a demonstrated one; no controlled demonstration of sophisticated evasion of medical-AI guardrails is presented here, and characterizing such evasion as already "demonstrated" would overstate the evidence.

The clearest illustration of objective inversion is the widely discussed exercise in which a generative drug-discovery workflow was redirected toward predicted toxicity, generating tens of thousands of candidate structures—including a known nerve agent—within hours, without altering the underlying platform [6]. The candidates were neither synthesized nor validated; the exercise established direction of risk, not feasibility of harm.

Is this isomorphism unique to biomedicine? Only partly. The paper's own software-security analogy cuts against a categorical reading: security teams routinely hire former attackers and accept the associated insider risk. The honest position is that biomedical defense knowledge overlaps offense knowledge more heavily, and with higher consequence, than in most other domains—not that it is uniquely and completely identical.

6. Regulatory Frameworks and Their Gaps

FDA materials on AI-enabled medical products address safety, effectiveness, credibility, transparency, bias, cybersecurity, predetermined change control, and lifecycle management [10,23,24]. None establishes a dedicated requirement to assess whether a product could be deliberately repurposed for toxin or pathogen design—a scope observation, not a charge of regulatory failure. The EU AI Act applies risk-based obligations to high-risk systems, including many medical devices, and regulates general-purpose models with systemic risk [25], but does not create a biomedical dual-use review or an insider-threat program. Institutional Review Boards protect human

 doi:10.20944/preprints202609.1616.v1

research participants [26] and are not constituted as biosecurity bodies. The net result is a mismatch: device rules, human-subjects rules, life-sciences dual-use policy, and AI risk frameworks each govern part of the chain from training through evaluation, deployment, fine-tuning, and physical execution—none governs all of it. The practical implication is institutional layering, not abandonment of any framework.

7. The Capability–Safety Tradeoff, Weighed

Openness (capability, access, competition, transparency) and safety (restriction, monitoring, centralization) cannot be simultaneously maximized. Prior treatments assert this tension but rarely weigh it. A responsible analysis should be explicit that the expected benefits of medical AI are large and near-certain across many applications, while the highest-consequence misuse pathways are lower-probability but potentially catastrophic and deeply uncertain. That structure—frequent, distributed benefit versus rare, high-variance harm—argues against both blanket restriction and blanket openness, and in favor of capability-specific gating: most medical AI should be developed and deployed openly, while a small tail of applications warrants heightened control.

International coordination is incomplete but not futile. Strategic and economic competition create incentives against unilateral restraint [27,28], and the Biological Weapons Convention lacks the verification regime of its chemical-weapons counterpart. Historically, nonproliferation regimes slow, shape, and stigmatize rather than eliminate. The realistic question is how much risk reduction a coalition can achieve, at what cost to beneficial research, and with what verification and incident response.

8. Toward Operationalized, Layered Governance

Technical measures should be described as risk reduction, not containment. Beyond them, four elements deserve specification.

Organizational controls. Least-privilege access, separation of duties, two-person controls for the most sensitive evaluations, tamper-resistant logging, protected reporting channels, periodic reauthorization, and independent review reduce—but do not eliminate—insider risk, and impose real costs in speed and reproducibility. Crucially, these controls do not escape the concentration problem by assertion: they mitigate it by ensuring no single individual holds both complete hazardous knowledge and unsupervised action capability, and by making sensitive activity observable and reversible. That is the specific mechanism by which oversight bodies avoid simply relocating the concentration trap.

Deployment-gating criteria. The claim that some applications may be "too dangerous to deploy" should be operationalized rather than gestured at. A workable gate would weigh: (a) the degree of defense–offense knowledge overlap; (b) the marginal uplift the system provides beyond publicly available resources, established empirically; (c) the residual material and tacit barriers to physical execution; (d) the reversibility and detectability of misuse; and (e) the counterfactual availability of the capability elsewhere. Applications scoring high on overlap and uplift and low on residual barriers and detectability are candidates for restricted release, structured access, or non-deployment—while the large majority of medical AI, scoring the other way, proceeds openly.

Detection and response. Even where prevention is imperfect, organizations can monitor anomalous access, preserve audit trails, rehearse incident response, and coordinate with public-health and law-enforcement authorities. Monitoring must be proportionate, privacy-conscious, and paired with human adjudication to manage false positives, and it will not observe locally run systems.

External oversight. An independent committee with access to confidential evaluations, authority to require remediation, conflict-of-interest rules, protected dissent channels, and a duty to publish non-sensitive summaries would add expertise and accountability. Membership should span biosecurity, clinical medicine, laboratory science, AI evaluation, ethics, security, and international policy.

 doi:10.20944/preprints202609.1616.v1

9. Conclusions

Medical AI offers substantial, near-certain benefits across most applications, and its development should continue. In a bounded set of high-consequence domains, however, it also carries genuine dual-use risk, because recognizing hazardous outputs requires expertise that overlaps heavily with the expertise needed to produce them. What can be stated with confidence is narrower than impossibility: technical measures alone cannot be assumed to contain dangerous knowledge reliably in these domains; expert review can itself create insider and concentration risks; international coordination is incomplete; and current governance frameworks each cover only part of the chain. Transparency about these limits is more credible than treating refusal behavior or output filtering as proof that dual-use risk has been solved.

The path forward is layered governance with capability-specific gating, honest cost-benefit reasoning, and explicit criteria for the rare cases warranting restriction—paired with the empirical work on which these judgments depend. The most important open questions are quantitative: how much current systems increase the capability of novices versus experts; which combinations of access, tools, tacit knowledge, and infrastructure create meaningful uplift; whether concentrated or distributed expert review yields lower net risk; and which organizational controls most reduce insider misuse without crippling beneficial research. These require controlled evaluation, red-team studies, incident data, and physical-world validation—not inference from benchmark scores.

AI Disclosure: Generative AI was used for researching, brainstorming, drafting and revision. Given the subject matter, the author notes the obvious reflexivity: a manuscript about dual-use AI risk was itself produced with AI assistance, and its claims should be judged on their evidence rather than on their provenance.

Appendix A. Three Cautionary Cases

Each case isolates a different link in the sociotechnical chain, and none establishes that present-day medical AI makes catastrophic misuse easy.

Objective inversion. A generative drug-discovery platform redirected toward predicted toxicity produced roughly 40,000 candidate structures, including a known nerve agent, within hours [6]. Candidates were neither synthesized nor validated. The lesson: a safety claim resting on a platform's benevolent purpose is fragile, because the same search machinery serves any objective.

Tool-enabled autonomy. A system coupling large language models to search, code execution, laboratory software, and robotics could plan syntheses and operate hardware with limited human direction [29]. No harmful synthesis was reported; the significance is architectural—the unit of safety is the model-plus-tools system, not isolated text output. Governance must evaluate tool permissions, action limits, logging, human authorization, and interruptibility.

Digital-to-physical translation. Chemical synthesis of poliovirus from public sequence data recovered infectious virus, but required specialized expertise, materials, and facilities [11]. Information can guide reconstruction, yet material execution remains a substantial, distinct barrier—reinforcing the graded-risk model and the need to coordinate model-layer controls with sequence screening, biosafety, and procurement oversight.

Together these cases support the paper's bounded thesis—technical filters on isolated outputs cannot carry the whole burden of safety—without implying that misuse is easy or inevitable.

References

1. Vamathevan J, Clark D, Czodrowski P, et al. Applications of machine learning in drug discovery and development. Nat Rev Drug Discov. 2019;18(6):463-477.
2. Esteva A, Robicquet A, Ramsundar B, et al. A guide to deep learning in healthcare. Nat Med. 2019;25(1):24-29.
3. Acosta JN, Falcone GJ, Rajpurkar P, Topol EJ. Multimodal biomedical AI. Nat Med. 2022;28(9):1773-1784.

 doi:10.20944/preprints202609.1616.v1

4. Liu X, Faes L, Kale AU, et al. A comparison of deep learning performance against health-care professionals in detecting diseases from medical imaging: a systematic review and meta-analysis. Lancet Digit Health. 2019;1(6):e271-e297.
5. National Research Council. Biotechnology Research in an Age of Terrorism. Washington, DC: National Academies Press; 2004.
6. Urbina F, Lentzos F, Invernizzi C, Ekins S. Dual use of artificial-intelligence-powered drug discovery. Nat Mach Intell. 2022;4:189-191.
7. Sandbrink JB. Artificial intelligence and biological misuse: differentiating risks of language models and biological design tools. arXiv. 2023;arXiv:2306.13952.
8. National Institute of Standards and Technology. Artificial Intelligence Risk Management Framework: Generative Artificial Intelligence Profile. NIST AI 600-1. Gaithersburg, MD: NIST; 2024.
9. Madani A, Krause B, Greene ER, et al. Large language models generate functional protein sequences across diverse families. Nat Biotechnol. 2023;41(8):1099-1106.
10. US Food and Drug Administration. Artificial Intelligence and Medical Products: How CBER, CDER, CDRH, and OCP Are Working Together. 2024.
11. Cello J, Paul AV, Wimmer E. Chemical synthesis of poliovirus cDNA: generation of infectious virus in the absence of natural template. Science. 2002;297(5583):1016-1018.
12. Mouton CA, Lucas C, Guest E. The Operational Risks of AI in Large-Scale Biological Attacks: Results of a Red-Team Study. Santa Monica, CA: RAND Corporation; 2024.
13. World Health Organization. Global Guidance Framework for the Responsible Use of the Life Sciences: Mitigating Biorisks and Governing Dual-Use Research. Geneva: WHO; 2022.
14. Carlini N, Tramer F, Wallace E, et al. Extracting training data from large language models. In: 30th USENIX Security Symposium. USENIX Association; 2021:2633-2650.
15. Hubinger E, van Merwijk C, Mikulik V, Skalse J, Garrabrant S. Risks from learned optimization in advanced machine learning systems. arXiv. 2019;arXiv:1906.01820.
16. Bai Y, Kadavath S, Kundu S, et al. Constitutional AI: harmlessness from AI feedback. arXiv. 2022;arXiv:2212.08073.
17. Ouyang L, Wu J, Jiang X, et al. Training language models to follow instructions with human feedback. arXiv. 2022;arXiv:2203.02155.
18. Wei A, Haghtalab N, Steinhardt J. Jailbroken: how does LLM safety training fail? Adv Neural Inf Process Syst. 2023;36:80079-80110.
19. Zou A, Wang Z, Carlini N, Nasr M, Kolter JZ, Fredrikson M. Universal and transferable adversarial attacks on aligned language models. arXiv. 2023;arXiv:2307.15043.
20. Inan H, Upasani K, Chi J, et al. Llama Guard: LLM-based input-output safeguard for human-AI conversations. arXiv. 2023;arXiv:2312.06674.
21. Lipton ZC. The mythos of model interpretability. Queue. 2018;16(3):31-57.
22. Rudin C. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. Nat Mach Intell. 2019;1:206-215.
23. US Food and Drug Administration. Marketing Submission Recommendations for a Predetermined Change Control Plan for Artificial Intelligence-Enabled Device Software Functions; and Artificial Intelligence-Enabled Device Software Functions: Lifecycle Management and Marketing Submission Recommendations. Guidance/Draft Guidance for Industry and FDA Staff. 2025.
24. International Medical Device Regulators Forum. Good Machine Learning Practice for Medical Device Development: Guiding Principles. 2025.
25. Regulation (EU) 2024/1689 of the European Parliament and of the Council of 13 June 2024 laying down harmonised rules on artificial intelligence. Off J Eur Union. 2024;L 2024/1689.
26. Federal Policy for the Protection of Human Subjects (Common Rule). 45 CFR part 46. Revised 2018; effective 2019.
27. Horowitz MC. Artificial intelligence, international competition, and the balance of power. Tex Natl Secur Rev. 2018;1(3):36-57.

 doi:10.20944/preprints202609.1616.v1

28. National Security Commission on Artificial Intelligence. Final Report. Arlington, VA: NSCAI; 2021.
29. Boiko DA, MacKnight R, Kline B, Gomes G. Autonomous chemical research with large language models. Nature. 2023;624(7992):570-578.