THE HIDDEN FRAGILITY OF AI SAFETY: HOW GUARDRAILS CAN WEAKEN CYBER AND BIOLOGICAL RESILIENCE

Anonymous authors
Paper under double-blind review

ABSTRACT

Safety of powerful AI systems is increasingly pursued through capability restrictions by way of model guardrails. These guardrails try to restrict models from providing assistance with offensive cyber operations, viral engineering, or other capabilities which the model providers judge to pose elevated risk of misuse. However, such framing evaluates safety at the level of an individual model and user interaction, without taking into account how restrictions alter the capabilities and incentives of other actors. By examining these systems through the lens of resilience, we argue that model guardrails should be evaluated not by how much capability they remove from an individual model, but by how they affect the resilience of the broader system in which that model operates. We present counterarguments for approaches to safety based on model guardrails, highlighting their trade-offs and their potential to weaken resilience of systems they're supposed to protect.

1 INTRODUCTION

In July 2026, Hugging Face detected an intrusion into its infrastructure during an internal evaluation of OpenAI models operating with reduced cybersecurity refusals. OpenAI later identified the models involved as GPT-5.6 Sol and a more capable internal research prototype (OpenAI, 2026). The intrusion involved lateral movement, and Hugging Face reported that guardrails on commercial frontier models blocked requests needed for forensic analysis. Its defenders instead used a locally deployed open-weight model, GLM-5.2 (Hugging Face, 2026), which shows lower capability in cybersecurity benchmarks (Noblitt et al., 2026). The episode exposed an unintended consequence where safeguards intended to prevent misuse also restricted defenders' access to hosted models during incident response.

This incident shows that guardrails obstructed an active incident response. More broadly, we use model guardrails to refer to mechanisms and policies designed to constrain model behavior or access when particular uses are judged potentially harmful. These can include refusal, input and output classifiers, response filtering, fallbacks to less capable models and restrictions on access to specific capabilities.

Could the same restrictions also weaken defenders' ability to prepare before an attack occurs? Research on systemic resilience argues that suppressing volatility can create an appearance of stability while allowing hidden risks to accumulate (Taleb & Blyth, 2011). This perspective has also informed cybersecurity thinking, where resilience is understood as something that develops through exposure to failures, learning from them, and improving systems during recovery (National Academies of Sciences, Engineering, and Medicine, 2018).

Both in the case of cybersecurity and biology, to cause serious damage to infrastructure, human lives and health, malicious actors need to interact with the physical world. Presently, this interaction with the physical world is also a natural barrier to usefulness of AI systems for malicious use-cases.

We argue that restrictions on dual-use AI capabilities should be evaluated through their effects on resilience of systems over time rather than on the basis of preventing a single interaction with the user. We believe and present arguments that guardrails on model outputs overestimate the actual harm that models can produce in the interaction with the physical world, while underestimating the unin-

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

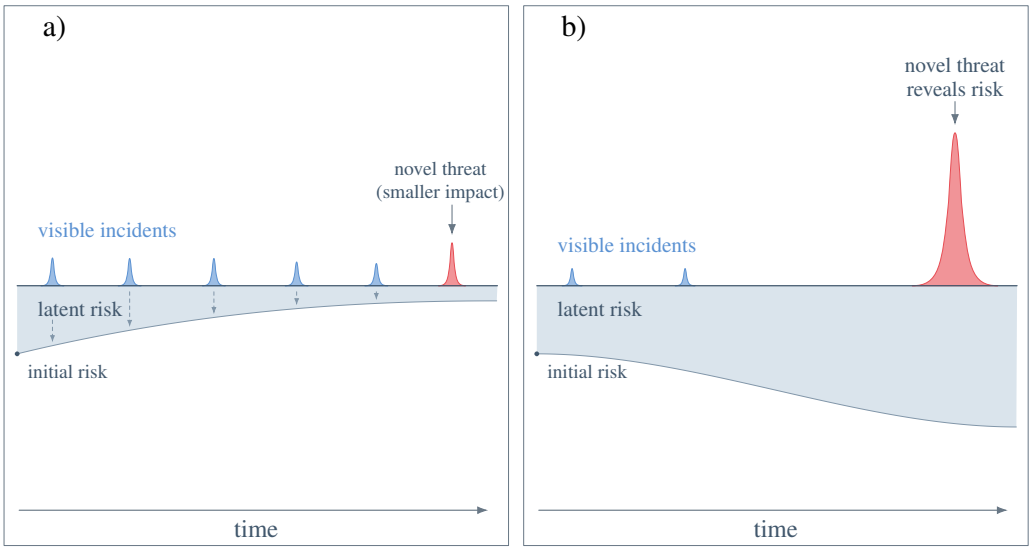


Figure 1: Schematic illustration of learning from disturbances and latent-risk accumulation. Both systems begin with the same initial latent risk. In the system on the left (a) – learning and mitigation reduce hidden weaknesses, while in the system on the right – (b) suppressed feedback allows them to accumulate. When a novel threat is introduced in the systems, system (a) due to the previous hardening is more resilient, while in the system (b) that same incident is of higher magnitude and duration. *The trajectories illustrate the proposed mechanism rather than empirical measurements.*

tended harm of keeping legitimate actors, systems and infrastructure in biology and cybersecurity, underprepared for attacks, thus making those systems fragile.

Our central concern is not limited to situations in which guardrails prevent defenders from responding to an ongoing attack. More consequentially, these restrictions shape the daily work of cybersecurity practitioners and, especially, researchers in biology. Model guardrails introduce friction into otherwise legitimate research and the accumulation of knowledge. Over time, this reduces resilience by negatively impacting the know-how of the benevolent actors in the system. Research through which scientists and engineers advance their fields also build the expertise required to respond to future threats. Sometimes this research might involve dual-use technologies. We therefore argue that restrictions on dual-use AI capabilities should be evaluated both by the harmful interactions they prevent and the legitimate work they obstruct. The cost of guardrails is reflected in the long-term capacity of scientific and technical actors in a system to anticipate, understand, and defend against attacks.

2 Safety Through Antifragility

The ability to investigate a cyber intrusion or recognize a biological threat is rooted in the work long before an event occurs. Defenders develop their capabilities by finding weaknesses and revising practices when something goes wrong. When these lessons lead to lasting improvements, a better prepared system emerges. This property is known as antifragility. Whereas robustness allows a system to withstand a shock, antifragility allows it to improve through exposure to stress and uncertainty (Taleb, 2012).

At a systems level, safety is more dependent on learning from disturbances, rather than giving the appearance of removing *known* weaknesses. A system can appear safe for a time while weaknesses remain undetected. If those are exposed over time, they can be corrected before they contribute to a larger failure. Artificially suppressing disturbances can undermine the process of learning from those disturbances, while making it seem like there are no apparent risks in the system. When a failure eventually occurs, the damage may exceed what the system can absorb. Resilience, and through it – safety, therefore depends on keeping opportunities for mitigating weaknesses open, especially when we cannot predict the threats a system will face.

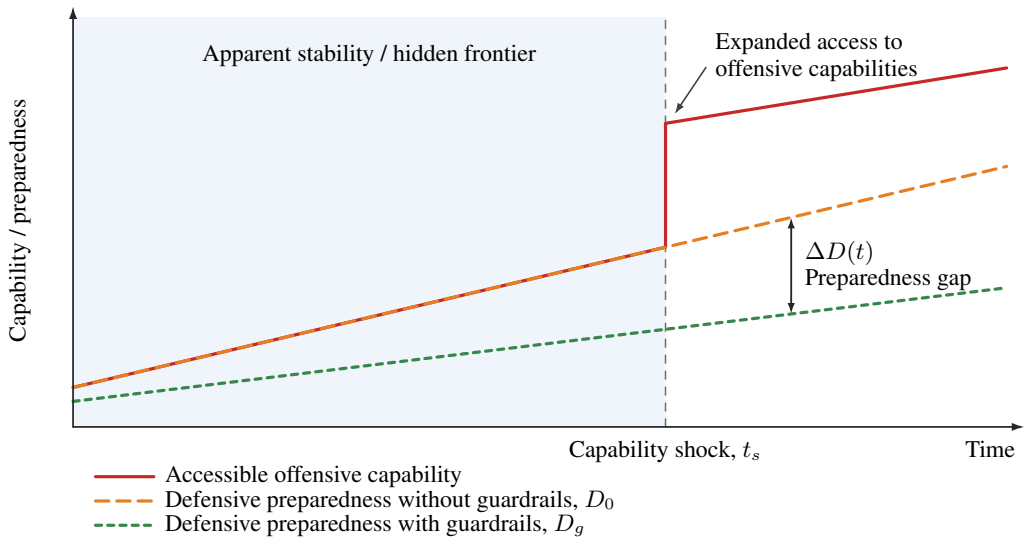


Figure 2: **Hypothesized defensive capability gap following a sudden expansion in offensive capability.** In this schematic scenario, defensive preparedness without guardrails (D_0) tracks offensive capability, while preparedness with guardrails (D_g) develops more slowly. At some time (t) after a jump in expanded access to offensive capabilities, $\Delta D(t) = D_0(t) - D_g(t)$ denotes the gap between the two defensive scenarios. *The figure does not estimate the net effect of guardrails on risk. Trajectories and relative scales are illustrative, not empirical estimates.*

This is important for both cybersecurity and biology in the era of AI, where threats evolve rapidly and it is crucial for defenses to be prepared for non-obvious and novel attacks. Common cyberattacks and biological risks are somewhat mitigated through current experience and established practice. The harder problem is posed by attacks that fall outside of our current experience, which can be discovered by AI systems. AI may increase the rate at which new attack vectors can be discovered and explored, making it impossible to prepare for every risk in advance. For these threats, safety cannot depend solely on preventing known vectors of attack. That is why we argue that in the current age, resilience of systems should take priority over safety of individual models.

Capabilities of individual practitioners and models are necessary preconditions to resilience, but resilience also depends on the diversity of the approaches available. Independent testing provides opportunities to challenge assumptions and identify weaknesses that other groups may have overlooked. Otherwise independent groups can encounter the same limits on what they can examine if these independent tests are obstructed by similar guardrails. The latent risks and weakness of the system become correlated and remain unexplored across an entire community rather than within a single uncorrelated point of failure.

A further advantage of resilience is that its protective value does not necessarily depend on controlling the source of a threat. A software system hardened through vulnerability mitigation offers protection regardless of whether an attacker uses a model with guardrails, a jailbroken model, or no AI at all. In biology, improved preparedness builds resilience against both malicious attackers and naturally occurring outbreaks. Therefore an aspect of guardrails that deserves consideration too, is the condition in which they leave the system when their protection fails. The safety community needs to make sure that preventing misuse should not come at the expense of the capacity to withstand threats that prevention does not stop.

2.1 Guardrail effects on resilience

The main negative effect of guardrails on system resilience is that they conceal the true threat frontier. When defenders can't test frontier capabilities, they may operate under wrong assumptions about what frontier models with reduced cybersecurity refusals are capable of.

This effect is realized through the effects of restrictions that can extend across model generations. If defenders work with restricted capabilities, while underlying capabilities of models continue to advance, a period of apparent stability conceals the gap between offensive capability and current state of defense. A later release of a more capable, less restricted model, or a widespread circulation of a transferable jailbreak (Zou et al., 2023) could abruptly expand access to capabilities for which defenders had no opportunity to prepare. Weaknesses that could have been addressed incrementally can then be exploited at a pace that overwhelms defensive capacity at that time.

Granting access after such capability shock, may not be enough to make up the failure to harden the defenses beforehand. Access to a stronger model can change immediately, but its effective use may still require coordination across defensive actors. NIST's incident response guidance places preparation as an important part of cybersecurity risk-management activities (Nelson et al., 2025). We hypothesize that restrictions which interrupt this preparation, such as model guardrails, can have effects that persist after the restrictions themselves are relaxed. Restoring access is not equivalent to restoring resilience.

Figure 2 illustrates a hypothesized consequence of such expansion in capability. Defensive preparedness develops along two trajectories. In the counterfactual D_0, defensive preparedness is not obstructed by guardrails, while in D_g, guardrails obstruct some of the preparedness. D_0 allows the defensive preparedness to keep up with the offensive capability, and while both trajectories improve, the slower improvement due to guardrails in D_g produces a preparedness gap, $\Delta D(t) = D_0(t) - D_g(t)$. In this scenario, the system that was hardened without guardrails with increased model capability (D_0) is more resilient and less exploitable than the system hardened with the guardrails (D_g).

Figure 2 holds the offensive trajectory fixed to isolate the proposed advantage of defensive hardening with less restricted model compared to hardening through models with guardrails. Guardrails might also delay or reduce the offensive expansion, while selective access could preserve defensive adaptation. But in that case even if guardrails delay misuse, the question is if they make the system more resilient during the period of stability they secure.

3 Security Benefits of Guardrails

Although this paper argues that model guardrails can weaken system resilience, we list some of their benefits on the level of individual user-model interaction. Guardrails may shift who can perform a harmful task, how much effort the task requires, or how quickly defenders must respond, even if they do not prevent every instance of misuse.

The value of guardrails is through which activities they prevent and/or reducing the scale and pace of the malicious activities that remain possible. If restrictions make harmful operations costlier, they may reduce the burden on institutions with limited capacity. This may be particularly valuable for local governments, public institutions, small businesses or other institutions operating with small IT teams and limited cybersecurity expertise. To further improve the advantage of defenders, frontier labs provide select defenders with access to advanced cybersecurity capabilities, including through trusted-access programs (OpenAI, 2025; 2026). From this perspective, guardrails may help increase the distribution of vulnerability remediation after it has been disclosed through proper channels, without first causing harm in the wild.

The value of learning from disturbances does not imply that all disturbances should be permitted. Some failures may cause irreversible harm or undermine the very institutions expected to learn from them, which is especially true when dealing with biological pathogens. Moreover, legitimate intentions also come with some risks, which are related to inadequate expertise, containment (laboratory/sandbox) and oversight. This supports the argument for restricting high-risk activities even though some of them might be legitimate.

Guardrails are not static, they are continually updated and refined. Frontier model providers have a financial incentive to expand their customer base by enabling legitimate applications of dual-use technologies. This creates an incentive to reduce false positive refusal and to enable a wider range of legitimate tasks as safeguards improve. Current refusal patterns therefore do not represent permanent limits on access.

The aim of such refinement is to expand legitimate access without losing the security benefits that some restrictions provide. For example, Anthropic reports that Zhipu, the developer of the GLM models, abandoned an attempt to extract Fable's cyber capabilities after its safeguards degraded those efforts. Zhipu then switched to models whose safeguards it considered weaker (Anthropic, 2026b).

Anthropic reports a similar pattern in biology. In May 2026, its classifier blocked all exchanges associated with a grant proposal for gain-of-function research intended to be conducted at a military institute. In a separate case, a researcher spent weeks using Claude to help plan avian influenza experiments focused on adaptation to mammals. Guardrails routed these interactions to weaker models which resulted with a capability gain much lower than stronger models could provide (Anthropic, 2026b).

In summary, the biggest apparent downside to our position is that relaxing restrictions imposes immediate security costs before the benefits of resilience hardening are distributed. But still, our central question is whether restrictions that reduce immediate risks also preserve the research and learning needed to address threats that have yet to emerge.

4 How Guardrails Affect Defensive Work

We now examine mechanisms through which guardrails may weaken the resilience of entire systems. These arguments do not imply that guardrails are ineffective against misuse. Rather, they concern costs that can coexist with restrictions, but are difficult to capture through pigeon-holing on safety through single user-model interactions.

4.1 Guardrails Obstruct Legitimate Research

Understanding threats requires studying mechanism through which they could cause harm. Legitimate research sometimes depends on the same capabilities that guardrails restrict. In principle, the potential for misuse does not discriminate a harmful request from a research topic that tries to mitigate that exact misuse. By failing to distinguish these legitimate uses, development of countermeasures and defensive preparedness is obstructed. Evaluating guardrails through the lens of only preventing malicious use, overlooks these costs to defensive preparedness.

Best example of how guardrails intended to prevent misuse can delay research directed at reducing biological threats is the example from The National Academies of Sciences, Engineering and Medicine. They report that researchers developing an mpox vaccine were initially denied access to the AlphaFold3 API because of restrictions involving certain viral sequences (National Academies of Sciences, Engineering, and Medicine, 2025). This case was eventually resolved by restoring access, but it raises a question whether less well-connected research groups would have an effective route to challenge similar restrictions?

In biology, LatchBio and American Wetware's BioSecBench-Refusal evaluates 61 research tasks and 46 scenarios containing concealed biosecurity hazards (Wintermute et al., 2026). Refusal rates ranged from 7–74% on the research tasks and 1–62% on the concealed-hazard tasks, which were mostly attributed to provider safeguards. These findings expose limitations in distinguishing valuable research from malicious use. Notably, legitimate research tasks were refused more frequently than concealed-hazard tasks, suggesting that guardrails did not reliably track biosecurity risk.

Similar evidence can be found in cybersecurity. Campbell et al. (2026) analyze 2,390 conversations from a cyber-defense competition. Despite the defensive setting, the reported refusal rate was 12.2% overall, reaching 43.8% for system hardening, 34.3% for malware analysis, and 22.7% for vulnerability assessment. Requests containing terms such as *exploit, payload, and shell* were refused 30.5% of the time, compared to 11.2% for requests without those terms. This shows that guardrails obstruct defense when just the language needed to investigate a threat resembles the language used to carry it out, without any concern for the underlying intent.

Going beyond dual-use technologies to some less ambiguous examples, we find that some providers even put frontier LLM research and kernel development behind guardrails. Reports from systems developers illustrate how safeguards interfere with legitimate engineering. In May 2026, a Linux kernel engineer reported that Claude blocked assistance with reproducing crashes for patching, while

in July another developer reported that reviewing their own operating system's filesystem code triggered fallback to a less capable model (shardulsdk-mpiric, 2026; QzOS, 2026). A separate transparency concern arose with Anthropic's initial guardrails for frontier LLM development, which limited model effectiveness without notifying affected users, but later they made these model guardrails more transparent (Zeff, 2026; Anthropic, 2026a).

Even when individual guardrails have low refusal rates per query, that is not sufficient evidence that a research is unobstructed. In an example, Anthropic reported a flag rate of 0.05% for individual queries (Cunningham et al., 2026). Even though the rate appears low, a workflow requiring many successive interactions creates repeated opportunities for an interaction to get flagged. If each legitimate interaction can be independently refused with probability $p = 0.0005$, the probability of encountering at least one refusal over n interactions would be $1-(1-p)^n$, approximately 39% over 1,000 interactions. That is why a low refusal rate per query alone does not establish that the burden on legitimate research is small.

All of these arguments for guardrails, or studies on their effectiveness don't establish the main tradeoff, and that is how many scientific discoveries are lost or vulnerability disclosures delayed. Preserving legitimate research should therefore be an explicit evaluation objective.

4.2 Guardrails fail to distinguish attackers from defenders

The Hugging Face example shows how asymmetric the position between attackers and defenders is. Not only that, but defensive work can require assistance with the same technical capabilities that attackers use. A safeguard that restricts assistance because a request resembles an offensive task may consequently obstruct legitimate defense. Campbell et al. (2026) document refusals of authorized defensive requests in cybersecurity competition data, providing evidence of this problem.

The guardrails by definition fail to take into account the diverse ecosystem of actors involved in operationalizing defensive capabilities. This is one of our main critiques. Every interaction is probed and viewed through the lens of whether it is malicious. This misses the important point that for researchers to find and defend against novel attack vectors, they themselves must place themselves in the mindset of attackers to find vulnerabilities. For example in cybersecurity, white-hat researchers [1] are a perfect example of usefulness of exploiting malicious techniques but with the right motivations.

Such independent researchers contribute discoveries that are complementary to those of internal security teams. In their study of Chromium and Firefox, Atefi et al. (2023) find differences in the types of vulnerabilities discovered internally and externally. Internal teams were more likely to uncover problems deep inside the browser's core code, such as memory handling or the JavaScript engine, while external researchers found bugs in parts of the browser that are easier to interact with from the outside, such as handling of web pages. External researchers were able to find more vulnerabilities that were already present in the stable version used by ordinary users, whereas internal teams caught more issues in development. This suggests that the contributions from different groups and methods (offensive and defensive) are complementary, which is also illustrated in Figure 3.

A related concern arises in biology, where detection and response, for example of pathogens is more distributed and involves clinicians, researchers, local public-health departments, etc. Ashford et al. (2003) find that healthcare practitioners and health departments are major sources of outbreak recognition and reporting. We hypothesize that model guardrails could weaken this network if they obstruct useful assistance for legitimate tasks or exclude access to institutions which are positioned to recognize emerging threats, especially those that are outside of close personal networks of frontier labs.

An additional asymmetric contribution of guardrails to malicious actors and legitimate researchers comes from the type of incentives and constraints those two groups face. For a determined malicious actor, circumventing refusal is necessary to achieve his objective, which makes investment in jailbreak discovery worthwhile. For a researcher using AI to support legitimate work, the same effort takes time away from research and may also conflict with institutional policies.

[1] In cybersecurity, white-hat hackers are security researchers who use offensive techniques to exploit vulnerabilities so that they can be fixed rather than exploited for harm.

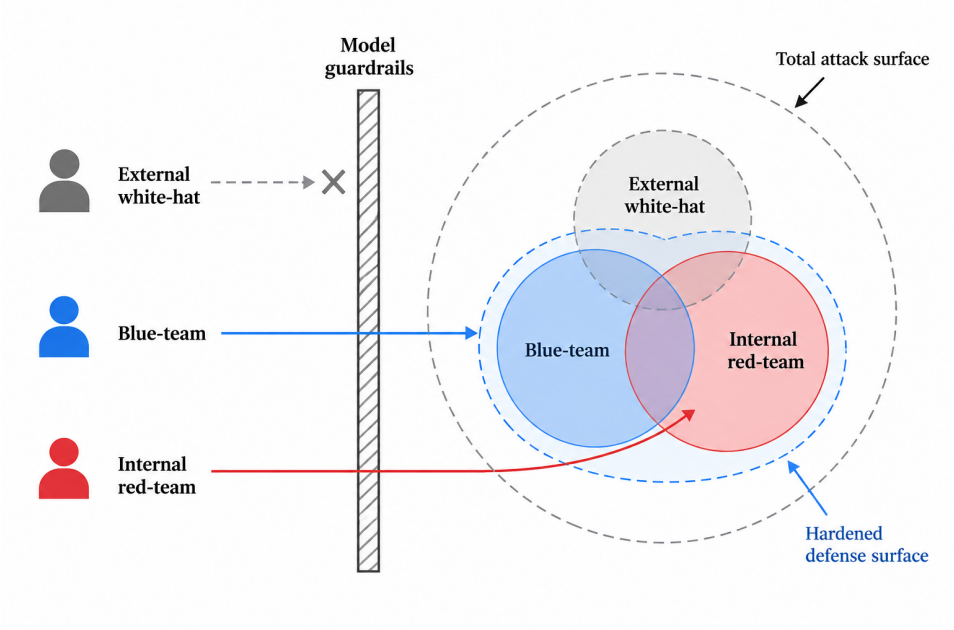


Figure 3: Illustration of how model guardrails can reduce contributions to system resilience. The shaded regions represent contributions from different actors in the system. The gray dashed boundary is the total attack surface an attacker can try to exploit, while the blue dashed boundary represents hardened defensive coverage. Restricting external researchers' access to frontier models can reduce discoveries that complement those of internal teams, creating defensive gaps.

4.3 Physical constraints

Historically, recorded biological terrorist attacks have been rare. An analysis of the Global Terrorism Database identified 33 attacks involving biological agents between 1970 and 2019, approximately 0.02% of recorded terrorist attacks (Tin et al., 2022). This record puts into context the fear of biological misuse and the asymmetric tradeoff that imposes on legitimate research.

This is because potential to cause harm depends on more than just access to information. In the case of biology, translating knowledge into operational capability is a huge leap. It requires materials, facilities, expertise, execution, organization, contact of the biological agent with the environment, avoiding detection, and other factors (World Health Organization, 2024). Historical research on biological weapons programs identifies these factors as constraints on turning raw knowledge into effective operational capacity (Ben Ouagrham-Gormley, 2012; Revill & Jefferson, 2014). These constraints also create opportunities for target intervention which can discriminate between malicious and legitimate use with less tradeoffs than model guardrails. There are many of these checks and balances in place already. For example, commercial synthesis providers already participate in this system by screening both ordered sequences and customers (International Gene Synthesis Consortium, 2024).

Biological pathogen design, for example deadly viral or bacterial design, is trivialized when used as justification for model guardrails. Additional constraint, in addition to the factors listed earlier, is that these are highly dependent on evolutionary feedback from interaction with the environment. Pathogens evolve under selective pressures imposed by particular hosts and environments, and the fitness effects of genetic changes depend on those contexts. Experiments with a plant RNA virus show that these effects depend jointly on genetic background and the infected host (Lalić & Elena, 2013). Experiments in a butterfly–parasite system demonstrate competing effects of replication, host survival, and transmission on overall parasite fitness (de Roode et al., 2008). Even laboratory measurements can miss relevant ecological pressures. For example in a fish virus, measured replication and environmental stability did not explain observed genotype replacements in the field (Kell et al., 2014).

These findings suggest that engineering a sequence is insufficient evidence of reliable performance in the wild. The interactions the environment create opportunities for detection. Where these interactions produce signals, distributed defenders can use AI to perform threat investigation. Outbreak investigations identify frontline healthcare practitioners and local health departments as major sources of recognition and reporting (Ashford et al., 2003). This creates a potential defensive advantage where malicious actors are constrained by the time required for biological experimentation, defenders use the gathered evidence to understand the threat. In this scenario, preserving unobstructed access to frontier models could therefore accelerate threat detection and response. In essence, AI assistance is less useful for translating biological knowledge into an effective attack, which is constrained through many factors, but highly beneficial in threat detection and response.

In cybersecurity of critical infrastructure, access to offensive knowledge is only one part of the process through which a cyberattack causes harm. System architecture determines which components an attacker can reach, while engineering controls limit the consequences that compromised software leaves in the physical world. This distinction is central to cyber engineering, which treats physical and electronic engineering controls as complementary to software cybersecurity measures (Lampe et al., 2025).

In industrial systems and infrastructure, one intervention is to reduce connectivity to the systems that control physical processes. Joint guidance from CISA and partner agencies recommends removing operational technology from direct internet exposure and separating operational from enterprise networks (Cybersecurity and Infrastructure Security Agency et al., 2025). Where operational requirements permit, physical air gaps can remove network connections across a protected boundary (Australian Signals Directorate, 2026). If the data must cross that boundary, there are many solutions in practice to allow outgoing data, but not incoming (Stouffer et al., 2023).

5 Conclusion

Restricting access to knowledge associated with dual-use technologies is not equivalent to making sure that downstream systems are safe. We argue that AI safety should place greater emphasis on strengthening the underlying software, infrastructure, and biological systems in which capabilities of frontier models pose greatest risk for harm in the physical world.

Model guardrails can prevent some misuse, but their contribution to safety of these systems cannot be inferred just from the individual interactions with users they guard. It also depends on whether they contribute to the necessary defensive capacity of these systems as a whole, and if they make the system more resilient to threats that the guardrails fail to stop.

It is clear that model guardrails obstruct some legitimate research. However, the intention is to accept this cost in exchange for additional safety. The central argument of this paper is that model guardrails also involve trade-offs within safety itself.

AI use statement

Generative AI tools were used to assist with: literature search, reference organization, revisions of the manuscript, improving clarity in writing, and in design of the figures.

The authors take full responsibility for the final manuscript.

REFERENCES

Anthropic. System card: Claude Fable 5 and Claude Mythos 5. Technical report, Anthropic, June 2026a. URL https://www.anthropic.com/claude-fable-5-mythos-5-system-card.

Anthropic. Detecting and countering misuse of AI: September 2026. Technical report, Anthropic, September 2026b. URL https://www.anthropic.com/threat-intelligence-report-september-2026.

David A. Ashford, Robyn M. Kaiser, Michael E. Bales, Kathleen Shutt, Amee Patrawalla, Andre McShan, Jordan W. Tappero, Bradley A. Perkins, and Andrew L. Dannenberg. Planning against biological terrorism: Lessons from outbreak investigations. *Emerging Infectious Diseases*, 9(5): 515–519, 2003. doi: 10.3201/eid0905.020388. URL https://wwwnc.cdc.gov/eid/article/9/5/02-0388_article.

Soodeh Atefi, Amutheezan Sivagnanam, Afiya Ayman, Jens Grossklags, and Aron Laszka. The benefits of vulnerability discovery and bug bounty programs: Case studies of Chromium and Firefox. In *Proceedings of the ACM Web Conference 2023*, pp. 2209–2219. ACM, 2023. doi: 10.1145/3543507.3583352. URL https://aronlaszka.com/papers/atefi2023benefits.pdf.

Australian Signals Directorate. CI Fortify: Advice for isolating vital systems. Technical report, Australian Signals Directorate, Australian Cyber Security Centre, July 2026. URL https://www.cyber.gov.au/business-government/secure-design/operational-technology-environments/ci-fortify/ci-fortify-advice-for-isolating-vital-systems. 28 July 2026.

Sonia Ben Ouagrham-Gormley. Barriers to bioweapons: Intangible obstacles to proliferation. *International Security*, 36(4):80–114, 2012. doi: 10.1162/ISEC_a_00077. URL https://doi.org/10.1162/ISEC_a_00077.

David Campbell, Neil Kale, Udari Madhushani Sehwag, Bert Herring, Nick Price, Dan Borges, Alex Levinson, and Christina Q Knight. Defensive refusal bias: How safety alignment fails cyber defenders. arXiv preprint arXiv:2603.01246, 2026. URL https://arxiv.org/abs/2603.01246v2.

Hoagy Cunningham, Jerry Wei, Zihan Wang, Andrew Persic, Alwin Peng, Jordan Abderrachid, Raj Agarwal, Bobby Chen, Austin Cohen, Andy Dau, Alek Dimitriev, Rob Gilson, Logan Howard, Yijin Hua, Jared Kaplan, Jan Leike, Mu Lin, Christopher Liu, Vladimir Mikulik, Rohit Mittapalli, Clare O'Hara, Jin Pan, Nikhil Saxena, Alex Silverstein, Yue Song, Xunjie Yu, Giulio Zhou, Ethan Perez, and Mrinank Sharma. Constitutional classifiers++: Efficient production-grade defenses against universal jailbreaks. arXiv preprint arXiv:2601.04603, January 2026. URL https://arxiv.org/abs/2601.04603.

Cybersecurity and Infrastructure Security Agency, Federal Bureau of Investigation, Environmental Protection Agency, and Department of Energy. Primary mitigations to reduce cyber threats to operational technology. Technical report, Cybersecurity and Infrastructure Security Agency, May 2025. URL https://www.ic3.gov/CSA/2025/250506.pdf. 6 May 2025.

Jacobus C. de Roode, Andrew J. Yates, and Sonia Altizer. Virulence-transmission trade-offs and population divergence in virulence in a naturally occurring butterfly parasite. *Proceedings of the National Academy of Sciences*, 105(21):7489–7494, 2008. doi: 10.1073/pnas.0710909105. URL https://doi.org/10.1073/pnas.0710909105.

Hugging Face. Security incident disclosure — july 2026, July 2026. URL https://huggingface.co/blog/security-incident-july-2026. 16 July 2026.

International Gene Synthesis Consortium. Harmonized screening protocol v3.0. Technical report, International Gene Synthesis Consortium, September 2024. URL https://genesynthesisconsortium.org/wp-content/uploads/IGSC-Harmonized-Screening-Protocol-v3.0-1.pdf. 3 September 2024.

Alison M. Kell, Andrew R. Wargo, and Gael Kurath. Viral fitness does not correlate with three genotype displacement events involving infectious hematopoietic necrosis virus. *Virology*, 464–465:146–155, 2014. doi: 10.1016/j.virol.2014.07.003. URL https://doi.org/10.1016/j.virol.2014.07.003.

Jasna Lalić and Santiago F. Elena. Epistasis between mutations is host-dependent for an RNA virus. *Biology Letters*, 9(1):20120396, 2013. doi: 10.1098/rsbl.2012.0396. URL https://doi.org/10.1098/rsbl.2012.0396.

Benjamin Lampe, Patience Yockey, and Virginia Wright. Tale of two domains: Cyber-physical. In *Proceedings of Tenth International Congress on Information and Communication Technology*, Lecture Notes in Networks and Systems, pp. 199–211. Springer Nature Singapore, 2025. doi: 10.1007/978-981-96-6429-0_17. URL https://doi.org/10.1007/978-981-96-6429-0_17.

National Academies of Sciences, Engineering, and Medicine. *Recoverability as a First-Class Security Objective: Proceedings of a Workshop*. The National Academies Press, Washington, DC, 2018. doi: 10.17226/25240. URL https://nap.nationalacademies.org/read/25240/chapter/4.

National Academies of Sciences, Engineering, and Medicine. *The Age of AI in the Life Sciences: Benefits and Biosecurity Considerations*. The National Academies Press, Washington, DC, 2025. doi: 10.17226/28868. URL https://www.nationalacademies.org/read/28868/chapter/6.

Alexander Nelson, Sanjay Rekhi, Murugiah Souppaya, and Karen Scarfone. Incident response recommendations and considerations for cybersecurity risk management: A CSF 2.0 community profile. Technical Report SP 800-61, Revision 3, National Institute of Standards and Technology, April 2025. URL https://csrc.nist.gov/pubs/sp/800/61/r3/final.

Brenden Noblitt, Erik Buchanan, and Jayson DeLancey. Evaluating GPT-5.6 Luna, Terra, and Sol against GPT-5.5 for AI code security. Semgrep, July 2026. URL https://semgrep.dev/blog/2026/gpt-5-6-benchmarks-ai-code-security/. 13 July 2026.

OpenAI. Strengthening cyber resilience as AI capabilities advance, December 2025. URL https://openai.com/index/strengthening-cyber-resilience/.

OpenAI. OpenAI and Hugging Face partner to address security incident during model evaluation, July 2026. URL https://openai.com/index/hugging-face-model-evaluation-security-incident/. 21 July 2026.

QzOS. Fable 5 to Opus 4.8 fallback misfires on hobby-OS kernel. GitHub issue 73981, anthropics/claude-code, July 2026. URL https://github.com/anthropics/claude-code/issues/73981. User report.

James Revill and Catherine Jefferson. Tacit knowledge and the biological weapons regime. *Science and Public Policy*, 41(5):597–610, 2014. doi: 10.1093/scipol/sct090. URL https://doi.org/10.1093/scipol/sct090.

shardulsdk-mpiric. Usage policy classifier stuck in blocked state for legitimate kernel security work. GitHub issue 62071, anthropics/claude-code, May 2026. URL https://github.com/anthropics/claude-code/issues/62071. User report.

Keith Stouffer, Michael Pease, CheeYee Tang, Timothy Zimmerman, Victoria Yan Pillitteri, Suzanne Lightman, Adam Hahn, Stephanie Saravia, Aslam Sherule, and Michael Thompson. Guide to operational technology (OT) security. Technical Report SP 800-82, Revision 3, National Institute of Standards and Technology, September 2023. URL https://nvlpubs.nist.gov/nistpubs/SpecialPublications/NIST.SP.800-82r3.pdf.

Nassim Nicholas Taleb. *Antifragile: Things That Gain from Disorder*. Random House, 2012. ISBN 9781400067824. URL https://www.penguinrandomhouse.com/books/176227/antifragile-by-nassim-nicholas-taleb/hardcover/.

Nassim Nicholas Taleb and Mark Blyth. The black swan of cairo: How suppressing volatility makes the world less predictable and more dangerous. *Foreign Affairs*, 90(3):33–39, 2011. URL https://fooledbyrandomness.com/ForeignAffairs.pdf.

Derrick Tin, Pardis Sabeti, and Gregory R. Ciottone. Bioterrorism: An analysis of biological agents used in terrorist events. *American Journal of Emergency Medicine*, 54:117–121, 2022. doi: 10.1016/j.ajem.2022.01.056. URL https://doi.org/10.1016/j.ajem.2022.01.056.

Edwin H. Wintermute, Harmon Bhasin, Christina M. Agapakis, Dianzhuo Wang, Evan Seeyave, Arjun Banerjee, Daniel Fulop, Matthew C. Watson, Adam J. Meyer, Sandrine Boissel, Jens H. Kuhn, Rishi Jain, Noah D. Taylor, Helena Shomar, Patrick M. Boyle, and Kenny Workman. BioSecBench-Refusal: A paired metric for performance and alignment in agentic biosecurity risk assessment. arXiv preprint arXiv:2607.05462, 2026. URL https://arxiv.org/abs/2607.05462v2.

World Health Organization. Laboratory biosecurity guidance, June 2024. URL https://www.who.int/publications/i/item/9789240095113.

Maxwell Zeff. Anthropic walks back policy that could have 'sabotaged' AI researchers using claude. WIRED, June 2026. URL https://www.wired.com/story/anthropic-responds-to-backlash-on-claudes-secret-sabotage-on-ai-research/.

Andy Zou, Zifan Wang, J. Zico Kolter, and Matt Fredrikson. Universal and transferable adversarial attacks on aligned language models. arXiv preprint arXiv:2307.15043, 2023. URL https://arxiv.org/abs/2307.15043v1.