ISSUE BRIEF

FRONTIER MODEL FORUM

Issue Brief: Preliminary Taxonomy of Frontier AI-Bio Misuse Mitigations

DATE
July 30, 2025

Frontier AI presents transformative opportunities within the biological sciences, including the potential to rapidly accelerate beneficial research discoveries and development. However, the dual-use nature of these technologies may also introduce novel risks. One potential harm involves the misuse of legitimately accessed frontier AI systems by malicious actors to create biological threats, such as a bioweapon.[1] As frontier AI capabilities advance, it is crucial to develop robust risk management practices that enable society to harness the benefits of AI in biology while proactively managing its most severe potential risks.

In light of this challenge, frontier model developers have committed to researching, implementing, validating, and sharing mitigation measures (also known as safeguards) to prevent the misuse of their models. This issue brief presents a preliminary taxonomy of safeguards designed to reduce the risk of biological misuse stemming from access to frontier AI models. Drawing from discussions with experts within the Frontier Model Forum (FMF) and the broader biosafety and biosecurity communities, this brief outlines the current landscape of AI-bio misuse safeguards, identifies potential future approaches to mitigations, and underscores the importance of implementing societal-level measures as a complement to technical safeguards.

MISUSE MITIGATIONS

Following the Frontier AI Safety Commitments, all FMF member firms have outlined mitigation measures in their Frontier AI Frameworks. These frameworks commit to implementing safeguards designed to prevent misuse of a model's capabilities in ways that could cause large-scale harm.[2] FMF firms have outlined various methods and approaches to manage and mitigate risks once frontier capability assessments suggest that the models have reached "enabling capability thresholds" — abilities that could potentially enable extreme harms if the model is deployed without additional safeguards.[3] While these frameworks outline various general safeguards to prevent misuse, such as harmlessness training, harm refusal techniques, and input and output monitoring, they do not specify how or which of these techniques are well suited to mitigate biological risks.[4]

This brief aims to outline examples of technical misuse safeguards and societal measures designed to prevent biological risks stemming from the deliberate misuse of frontier AI. Here, frontier AI refers primarily to Large Language Models (LLMs), though the safeguards outlined below may also be useful for specialized models such as AI-enabled Biological Tools (BTs) — AI tools trained on biological data using machine learning techniques, such as deep neural networks, with the potential to facilitate the development of hazardous biological agents.[5] This includes tools built to provide insights, predictions, and designs (e.g. BDTs) related to biological systems.[6] Given the focus of Frontier AI Frameworks on the most severe, large-scale risks posed by the most advanced AI models, misuse is scoped to include only severe and large-scale catastrophic scenarios, such as the creation and deployment of a bioweapon.

Technical misuse safeguards are technical interventions that can be implemented by frontier AI developers and downstream developers to prevent users from eliciting information, actions, or assistance from AI models or systems for harmful use cases.[7] Societal safeguards are measures implemented outside the AI model and its direct deployment environment, typically involving physical world controls, supply chain security, regulatory compliance, or inter-organizational coordination. For brevity, of the large set of societal safeguards that exist, this brief only discusses those that focus on preventing catastrophic outcomes and that AI developers may play a role in strengthening through measures like information sharing, reporting, or supporting defensive systems and research. This brief does not address safeguard approaches aimed at protecting the safeguards or models themselves from compromise (e.g. jailbreak prevention, secure access protocols).

While specific approaches to implementing safeguards may vary, developers (including model developers and downstream developers) generally follow a holistic approach with technical safeguards, incorporating multiple layers of defense to prevent misuse.[8] Following this principle, this issue brief is intended to capture a snapshot of the current and potential future suite of technical and societal safeguards available for model developers and societal actors. That said, AI safeguards are an active area of research and techniques that currently seem promising may be replaced as the research progresses. In practice, the appropriateness and efficacy of any specific mitigation measure may vary depending on the nature of the model and the manner in which it is made available to the public. Many of the mitigations discussed here merit additional research to examine their effectiveness. As such, this brief also does not make claims about an ideal combination of safeguard techniques based on their effectiveness, as measuring their resilience to various and novel adversarial approaches is an active research area.

AI-BIO MISUSE MITIGATIONS

AI-bio misuse safeguards can be categorized by their function. For example, they may be used for:

- **Capability Limitation**: Approaches that alter the model's weights or training process to prevent models from possessing knowledge or abilities that could enable harm in the biological domain. These approaches are common for LLMs and are becoming increasingly common for BTs, particularly via data filtering.

- **Behavioral Alignment**: Approaches that seek to prevent a model's potentially dangerous capabilities from being elicited by shaping the model's responses to human requests and its autonomous decision-making processes. While behavioural alignment techniques are common for LLMs, to date they have only been theorized for BTs, given a different set of challenges regarding understanding what constitutes a "dangerous output".

- **Detection and Intervention**: Approaches that rely on automated methods to detect model usage (e.g., inputs and outputs) that may give rise to undesired behavior. These mitigations are common for LLMs, but have only been theorized for BTs.

- **Access Control**: Approaches that govern who can use a model, what capabilities they can access, and how the model can interact with external systems. These methods establish boundaries that determine the conditions under which model capabilities can be utilized. This category of safeguards are used widely for both LLMs and BTs, and has precedents in many other related domains (e.g. human genomics).

- **Supporting Ecosystem Mitigations**: Approaches where developers provide information, tools, and capabilities that enable other actors – governments, organizations, and civil society – to implement effective defenses against AI-enabled threats. Rather than directly controlling societal defenses, developers contribute by sharing resources that strengthen the broader defensive ecosystem.

Safeguards may also be categorized by their mode of application.[9]

- **Model-level**: Techniques applied during model training, fine-tuning, or alignment that directly modify the model's parameters and underlying behavior patterns to prevent harmful outputs.

- **System-level**: Techniques implemented in the deployment environment or application layer that monitor, filter, or restrict model inputs/outputs without modifying the model's internal parameters.

- **Societal-level**: Measures implemented outside the AI model and its direct deployment environment, typically involving physical world controls, supply chain security, regulatory compliance, or inter-organizational coordination.

The preliminary taxonomy in Tables 1 through 5 below outline potential misuse safeguards, including safeguards that exist outside the domain or control of AI model developers (or "societal safeguards"). The list of safeguards are meant to offer an overview of prospective mitigations, but is not intended to be prescriptive given that many of these techniques are still being explored, have limitations, and configurations of specific techniques are often case specific and may not be applicable to all risk scenarios.

TECHNICAL AND SOCIETAL AI-BIO MISUSE SAFEGUARDS

Both technical and societal safeguards are critical to mitigating frontier AI biosecurity risks, for three principal reasons. First, mitigating complex human-mediated risks often requires a multi-faceted approach, implementing several layers of imperfect guardrails with uncorrelated failure modes to reduce the risk of an unacceptable outcome. No single safeguard covers all risk vectors; rather, each addresses a specific area or risk vector. A combination of mitigations can create a more robust defense posture by having the safeguards address various parts of the potential attack chain, ensuring that if the safeguards fail, they fail independently of each other. Technical safeguards present an important first line of defense for combatting model misuse, but a combination of technical and societal mitigations is critical to reducing the risk of catastrophic outcomes. Finally, given the complexity of accurately assessing the full array of risks stemming from a given model capability operating within a network of enabling tools, a defense-in-depth approach serves as a hedge against inaccuracies in risk measurement and gaps in risk mitigation.

Second, as models become cheaper and more efficient, the number of resources needed to train or obtain access to a model with a given capability level falls, which in turn causes capabilities to proliferate over time.[10] While bespoke technical safeguards for proprietary models may help to decrease the risks of advanced scientific capabilities being misused to enable harmful outcomes, they may eventually become obsolete if other developers train new models with similar capability levels and no safeguards in place.[11] In these cases, societal safeguards are especially critical to mitigate the risk of misuse.

Third, although researchers are making substantial progress in developing technical safeguards, current techniques have limitations and their effectiveness can vary widely. For instance, techniques like refusal training may be less robust for open-weight models or models accessible via fine-tuning APIs.[12] These limitations underscore the need for a defense-in-depth approach incorporating guardrails

Table 1: Capability Limitations Mitigations for AI-Bio

Safeguard	Description	Bio Application	Limitations
Data Filtering**	Removing content from training datasets that could lead to dual-use or potentially harmful capabilities. Developers can use several methods, such as: automated classifiers to identify and remove content related to bioweapons development, detailed attack methodologies, or other high-risk areas; keyword-based filters to exclude documents containing specific terminology or instructions of concern; and machine learning models trained to recognize subtle patterns in content that might contribute to capabilities that could enable harmful outcomes.[13]	Data filtering may be used to remove detailed pathogen engineering protocols, weaponization procedures, and/or other biological knowledge that could enable harmful outcomes.	Scientific and technical knowledge often has dual-use applications. Knowledge about biological sequences and pathogen research enables models to assist with vaccine development and disease understanding, but these same capabilities could potentially help malicious actors modify dangerous pathogens. It can be challenging to comprehensively identify and remove all potentially dangerous information without negatively impacting the model's beneficial biological knowledge. Finally, capabilities of concern can also emerge from combinations of seemingly benign training data through unexpected inferences.
Targeted Unlearning** †	Attempts to selectively remove specific capabilities that could enable harmful outcomes from models after initial training, offering a more precise alternative to full retraining. Possible approaches include fine-tuning on datasets to overwrite specific knowledge while preserving general capabilities, or modifying how models internally structure and access particular information.	Targeted unlearning may be used to remove detailed knowledge of dangerous biological procedures or potentially harmful domains (e.g., virology), while maintaining general biological understanding.	Unlearning methods may be reversible with relatively modest effort. They have been shown to lack robustness, as knowledge can sometimes be recovered through specific prompting techniques or model manipulation, such as targeted fine-tuning with small datasets. Models may also regenerate removed knowledge by inferring from adjacent information that remains accessible.
Model Distillation†	Create specialized versions of frontier models with capabilities limited to specific domains.	Model distillation may be used to create a model that excels at medical diagnosis while lacking the knowledge needed for biological weapons development.	While the capability limitations may be more fundamental than post-hoc safety training, it remains unclear how effectively this approach prevents harmful capabilities from being reconstructed. Additionally, multiple specialized models would be needed to cover various use cases, increasing development and maintenance costs.
False Learning†	Training the model on deliberately fabricated but plausible-sounding incorrect information related to dangerous procedures (e.g., bioweapon synthesis), aiming to mislead potential misusers.	The model could provide subtly flawed instructions for synthesizing a dangerous pathogen, rendering the attempt ineffective or unsafe for the user.	Introducing incorrect information about biological sequences, structure, or functional data may have unpredictable side effects on the model's overall reliability and usefulness, and could potentially mislead legitimate researchers.

** Indicates consensus i.e. the technique is mentioned explicitly in all FMF member firm frontier AI frameworks.*

*** Indicates a safeguard mentioned in one or several FMF member firm frontier AI frameworks, but not all.*

† Indicates experimental/proposed techniques.

Table 2: Behavioral Alignment Mitigations for AI-Bio

Safeguard	Description	Bio Application	Limitations
Refusal Training*†	**SFT***: Developers curate datasets of desired model behaviors and fine-tune models to match these examples. Datasets include refusal examples (such as declining to provide instructions on making bioweapons) and helpful, yet harmless responses. SFT directly teaches models specific behavioral patterns through imitation learning. **RLHF***: Developers use human preferences between different model outputs to questions as a reward signal. They then use reinforcement learning to optimize models for these reward signals. **Adversarial Training**†: Enhancing model robustness by training it on examples specifically crafted to deceive it or cause failures, thereby teaching it to resist such manipulations. Developers can improve processes to search for failure modes through both human and automated red-teaming methods, and analyses of real-world jailbreaks.	Refusal training may include training the model to recognize and refuse requests for information on building bioweapons, acquiring dangerous pathogens, or bypassing biosecurity protocols, or to provide appropriate warnings about biosafety. It may also be used to help identify and patch vulnerabilities that could be exploited to extract biological information that may be used for malicious purposes.	Safety training methods modify surface-level behaviors without altering underlying model capabilities, which can be undone through targeted fine-tuning. In the biological domain, this may be done by fine-tuning a model with specific harmful information, such as virology. Adversarial prompts (“jailbreaks”) can bypass alignment post-training. Behavioral alignment requires translating human values into training objectives, but this translation introduces challenges such as reward hacking, alignment faking, or goal misgeneralization.
Reinforcement Learning from AI-assisted Feedback (RLAIF)**	Uses AI systems to generate training feedback based on predefined principles or constitutions, and has gained traction as a scalable alternative to purely human-generated feedback.	Methods like Anthropic's Constitutional AI, OpenAI’s deliberative alignment, and the broader category of AI-assisted feedback, including RLAIF, can be used to guide the model to refuse harmful requests, such as for creating a bioweapon.	The effectiveness of this safeguard depends heavily on the comprehensiveness and interpretation of the predefined principles, and it can still be vulnerable to sophisticated bypass attempts.

Table 3: Monitoring and Intervention Mitigations for AI-Bio

Safeguard	Description	Bio Application	Limitations
Automated Input/Output Monitoring Systems and Response Protocols*	Language models can be prompted to act as classifiers on inputs to and outputs of the model. They act as surveillance systems that analyze interactions in real time, using both simple methods like keyword detection and more sophisticated semantic analysis to identify potentially dangerous queries or responses.	This may include monitoring for keywords and patterns related to dangerous pathogens or toxins, or any key steps required for the development of bioweapons.	Filters may be bypassed if malicious prompts are separated into seemingly benign substeps or distributed across different accounts. Filters may also produce a high rate of false positives, blocking legitimate scientific queries.
Custom-trained Classifier Models**	Language models can be trained via fine-tuning to classify inputs and outputs from the model, producing probability scores that indicate whether content contains potential harmful materials.	Developers may guide training using explicit rules defining permissible and restricted content, such as Anthropic's Constitutional Classifiers. They may specifically screen for dangerous biological procedures, weaponization information, or concerning dual-use research.	The classifier itself might have vulnerabilities or interpret the constitution incorrectly, and adds latency to the response generation.
Manual and Automated Abuse Monitoring**	Monitoring user interaction patterns over time to detect suspicious activity indicative of attempts to misuse the model, such as repeated probing for sensitive information or trying to circumvent safety filters.	This can identify users systematically trying to gather information needed for bioweapon development, even if individual prompts don't trigger input filters.	Defining and detecting "abusive" patterns accurately without violating user privacy or flagging benign research behaviour is challenging.
Context-Aware Access Control[†]	Restricts access to certain model capabilities based on user authentication and context.	May be useful in cases when certain queries should only be accessible to verified researchers or institutions.	Defining appropriate contexts and verifying user credentials reliably can be complex and may create barriers for legitimate users.
Tar Pitting / Throttling[†]	Detects harmful use and then intentionally restricts model capabilities in a way that slows down the user without the restriction being detected by the user. For example, the model might offer useless or wrong information, or slow down token generation rates. Compared to a refusal where the user knows the model is stopping progress, this approach is designed so users are unaware of the intervention.	If a user is persistently trying to elicit sensitive bioweapon information, the system could subtly degrade the quality or speed of responses on that topic to frustrate their efforts without triggering obvious refusals.	Sophisticated users may eventually detect the manipulation, and degraded outputs could severely mislead legitimate users if applied incorrectly.

Table 4: Access Control Mitigations for AI-Bio

Safeguard	Description	Bio Application	Limitations
Staged Deployment**	Developers may implement staged rollouts for new models, starting with highly controlled environments and gradually expanding access as controls are validated. Initial deployment could involve small groups of external users or research partners operating under strict monitoring agreements, allowing developers to observe model behavior and identify potential risks before broader release.	Staged deployment may include giving initial access to verified researchers to carry out research, followed by expanded access to wider groups over time.	Monitoring usage becomes more difficult as access expands.
Sanctions Screening / Geogating**	Preventing users from countries under international sanctions from accessing frontier AI models by checking user locations and IP addresses against lists of sanctioned countries and regions and blocking access to frontier AI systems for users in those locations.	Sanctions screening or geogating attempts to prevent access to powerful models by actors in states known or suspected to be pursuing illicit bioweapons programs.	Users may be able to bypass geogating using VPNs or other anonymization techniques, and sanctions lists often will not perfectly correlate with actual threats.
User Verification Protocols[†]	Requiring users to verify their real-world identity (Know Your Customer) before granting access to certain model capabilities.	Linking model usage to a verified identity creates accountability and may deter users from attempting illicit activities like seeking bioweapon designs.	User verification processes introduce friction, raise privacy concerns, can be spoofed with stolen identities, and may exclude legitimate anonymous users.
Access to BTs[†]	Controlling or monitoring how the LLM interacts with or provides information related to specialized BTs, potentially blocking requests that try to automate dangerous designs using such tools.	Prevent the LLM from being used as an interface to generate potentially hazardous information via BTs.	Defining permissible use cases for biological tool use, and the conditions under which they would be permissible, is difficult given the many possible use cases.

Table 5: Ecosystem Mitigations for AI-Bio

Safeguard	Description	Bio Application	Limitations
Nucleic Acid Synthesis Screening	Companies selling DNA/RNA synthesis services screen orders against curated databases of pathogenic sequences to identify and block potentially dangerous requests before synthesis occurs and verify the identity and legitimacy (e.g., institutional affiliation, stated research purpose) of their customers. May also include peptide synthesis screening.	Nucleic acid synthesis screen can act as a critical chokepoint preventing the direct conversion of hazardous genetic sequences into tangible biological material, effectively disrupting a critical step in physically manufacturing a malicious design created digitally. AI developers could support these efforts by sharing threat intelligence on AI-enabled misuse scenarios and collaborating on the development of advanced AI tools to improve the accuracy and efficiency of sequence screening and customer verification processes for DNA/RNA synthesis companies.	Synthesis screening relies on the comprehensiveness of databases, and may not catch all malicious requests. In addition, screening is currently not legally mandated and therefore not all synthesis providers screen their orders.
User Verification Protocols for other Services	Companies compiling databases of sensitive biological materials, or selling specialized equipment verify the identity and legitimacy of their customers.	User verification protocols can impede malicious actors from procuring the essential materials needed to build a bioweapon. AI developers could support these protocols by collaborating on identity verification protocols	Identity verification processes can potentially be spoofed using stolen credentials or shell corporations. They also introduce bureaucratic hurdles, potentially increasing the cost of access for legitimate researchers. Global implementation standards and enforcement also vary significantly.
Insider Threat Mitigations (e.g. vetting)	Measures to reduce the risk of insider threats, including formal screening processes for individuals granted access to high-containment facilities, dangerous pathogens, or particularly sensitive biotechnologies.	Aims to reduce "insider threats" risks by preventing individuals with legitimate access from stealing materials, leaking sensitive information, or conducting unauthorized hazardous experiments.	Vetting procedures may not reliably predict future malicious intent and can raise privacy concerns for personnel. They may also fail to capture risks from external collaborators, visitors, or cyber infiltration targeting insiders.

Table 5: Ecosystem Mitigations for AI-Bio (Cont'd)

Safeguard	Description	Bio Application	Limitations
Oversight of Automated Bio-Platforms	Establishing licensing requirements, mandatory safety/security audits, or regulatory oversight for commercial "cloud labs" and other highly automated biological design, synthesis, and testing platforms.	Ensures that organizations providing powerful, potentially AI-integrated bioengineering services implement robust internal safeguards (e.g., sequence screening, user verification), proportionate to the advanced capabilities offered.	Regulation often lags behind rapid technological development in automated biology, and defining appropriate, globally consistent oversight mechanisms can be complex. Illicit actors may seek unregulated platforms or attempt to build their own capabilities.
Export Controls on Dangerous Biological Goods & Technology	National governments implement regulations restricting the cross-border transfer of specific pathogens, toxins, genetic elements, critical dual-use equipment, and related technical data deemed sensitive for proliferation.	Export controls aim to prevent states or non-state actors from acquiring materials, equipment, or technologies relevant to bioweapon development from foreign sources. This limits international proliferation pathways for both physical goods and critical know-how.	The effectiveness depends heavily on international cooperation and harmonized control lists, as illicit procurement networks actively work to circumvent these national controls. Furthermore, controlling the transfer of intangible technology (expertise, complex designs) remains exceptionally difficult. In addition, export controls only address dangerous materials that cross borders.
Secure Information-Sharing Networks	Establishing trusted communication channels and protocols for relevant stakeholders to securely share threat intelligence and biosecurity information.	Facilitating timely sharing of information about suspicious activities, novel threats (like specific AI misuse techniques observed "in the wild"), or identified vulnerabilities allows for more coordinated and rapid prevention or response efforts across different sectors.	Building and maintaining trusted communication channels between diverse organizations, especially across public and private sectors or internationally, is a significant hurdle. Information sharing must also carefully navigate legal frameworks, privacy concerns, and commercial sensitivities to be effective.

beyond what can be implemented by solely AI developers, and the importance of examining the deployment context of a model within an overall system. These factors underscore the importance of implementing a defense-in-depth approach to risk mitigation with multiple, complementary layers of mitigation, ensuring no single point of failure. This approach requires a coordinated stance across pre-deployment and post-deployment phases of model development, and across AI- and non-AI sectors of the ecosystem.

EMERGING PRACTICES FOR AI-BIO MISUSE SAFEGUARDS

Certain practices are emerging as foundational for managing AI bio-misuse risks:

- While the specific configuration of safeguards varies between frontier AI developers, refusal training and automated input and output monitoring are commonly used techniques to prevent misuse.
- Recognizing trade-offs around performance, effective risk reduction often requires implementing several safeguards in layered, holistic strategies.
- Safeguard strategies should be targeted to address specific, plausible threat scenarios relevant to AI-enabled biological misuse, within specific deployment contexts. The selection and prioritization of safeguards should be directly informed by risk management processes, including structured assessments that evaluate the plausibility and potential impact of different misuse pathways.
- Once safeguards have been implemented, it is critical to verify their efficacy in reducing risks to acceptable levels. Safeguards can be tested for effectiveness individually or in combination through a variety of methods, and the decision about whether to change safeguards should be based largely on this evidence. It is also important to take into account other tradeoffs in areas like cost, transparency, performance, research, and user privacy.[14]

Further research is needed to understand the trade-offs between various combinations and approaches. For example, monitoring and intervention techniques may be implemented as either a complement to, or a substitute for, output refusals, because refusing malicious requests outright may incentivize malicious actors to diversify their attack strategy and make misuse less visible. These strategies include supporting critical societal measures to prevent catastrophic bio-misuse. While AI developers do not control the physical inputs to creating a biological threat, they can help enhance ecosystem defences by establishing lines of communication for voluntary information-sharing with the organizations responsible for those safeguards, supporting the development of systems specifically designed to strengthen defensive capabilities, and establishing mechanisms for rapid threat detection and response.

CONCLUSION

Safeguards are a critical component of managing the misuse risks at the intersection of frontier AI and biology. This issue brief has presented a preliminary taxonomy of existing and exploratory potential technical safeguards, as well as downstream and societal mitigation measures that may contribute to a robust risk management strategy. While this brief provides an initial overview, the rapid rate of AI progress and the dynamic nature of biological threats requires continuous effort and forward-looking research.

Further work is needed to mature the landscape of AI-bio misuse safeguards. Key future directions include enhancing the robustness and adaptability of safeguards against evolving threats (such as the ability of agentic AI systems to use specialized biological tools), developing standardized evaluation methodologies to rigorously test safeguard effectiveness, and improving our understanding of potential trade-offs in model performance when implementing safeguards.

FOOTNOTES

1. Pannu, J., S. Gebauer, G. McKelvey Jr, A. Cicero, and T. Inglesby, "AI could pose pandemic-scale biosecurity risks. Here's how to make it safer," *Nature*, 2024; Sandbrink, J., "Artificial intelligence and biological misuse: Differentiating risks of language models and biological design tools," Arxiv, 2023.
2. METR, "Common Elements of Frontier AI Safety Policies," 2025, pg. 29.
3. Frontier Model Forum, "Risk Taxonomy and Thresholds for Frontier AI Frameworks," 2025.
4. For example, Anthropic's ASL-3 Deployment Standard calls for implementing a "defense in depth" approach to prevent misuse, such as through harm refusal techniques, automated detection, and harmlessness training. OpenAI's Preparedness Framework specifies techniques such as harmlessness training, unlearning or data filtering, adversarial training, and input and output monitoring.
5. Rose, S. and C. Nelson, "Understanding AI-Facilitated Biological Weapon Development," Center for Long-Term Resilience, 2023, pg. 2.
6. Carter, S., N. Wheeler, C. Isaac, and J. Yassif, "Developing Guardrails for AI Biodesign Tools," NTI, 2024.
7. Following a similar definition by the UK AISI, "Principles for Evaluating Misuse Safeguards of Frontier AI Systems," 2025.
8. See Frontier Model Forum, "Foundational Security Practices," 2024.
9. Safeguards may also be implemented at the organizational level. These include safety culture, risk governance, and organizational security. This brief does not cover this type of measures.
10. Scharre, P. "Future-Proofing Frontier AI Regulation: Projecting Future Compute for Frontier AI Models," CNAS, 2024, pg 27.
11. For further reading, see Toner, H., "Nonproliferation is the wrong approach to AI misuse," Rising Tide Substack, 2025.
12. Qi, X., et al., "On Evaluating the Durability of Safeguards for Open-Weight LLMs," ICLR 2025.
13. Carter et al., "Developing Guardrails for AI Biodesign Tools," 2024, pg. 8.
14. "Managing Misuse Risk for Dual-Use Foundation Models," NIST AI 800-1, 2024.