Perspective: Lessons from Cybersecurity for Biological AI Safety and Regulation

Azmine Toushik Wasi[1,2*] **and Mst Rafia Islam**[1,3]
[1]Computational Intelligence and Operations Laboratory (CIOL)
[2]Shahjalal University of Science and Technology [3]Independent University, Bangladesh
Correspondence to: azmine32@student.sust.edu

Abstract

Rise of generative artificial intelligence (AI) and its intersection with biotechnology is creating new biosecurity risks that traditional defenses cannot manage. Static, list-based systems designed to stop known threats are ill-equipped against novel pathogens that could be enabled by Large Language Models (LLMs) and advanced Biological Design Tools (BDTs). These technologies may lower barriers for inexperienced actors and accelerate the design of dangerous agents. We argue that cybersecurity offers a useful guide for responding to this challenge. Cybersecurity once relied on "castle-and-moat" defenses but shifted to resilience-based models like zero trust, which assume breach and focus on continuous verification and protection at the data level. Applying similar principles in biosecurity could enable secure tracking of biological designs, proactive testing through red-teaming, and collective defense via shared threat intelligence. This perspective calls for biosecurity to move from a reactive add-on to a secure-by-design foundation. Such a shift will require new technologies, governance, and interdisciplinary expertise to ensure that the bioeconomy advances safely and responsibly.

1 Introduction

Rapid convergence of generative artificial intelligence (AI) and biotechnology is reshaping both the opportunities and risks in the life sciences. Generative AI, including Large Language Models (LLMs) and domain-specific Biological Design Tools (BDTs), holds enormous promise for accelerating drug discovery, vaccine development, and synthetic biology. At the same time, these same tools introduce profound biosecurity risks by lowering barriers to entry and enabling the design of novel biological agents with potentially catastrophic consequences [2, 31]. Unlike traditional biothreats, which could be monitored through known lists of dangerous pathogens, AI-enabled threats emerge from vast combinatorial design spaces that make prediction and containment far more difficult. Historical biosecurity models, rooted in list-based defenses such as the Select Agents and Toxins List, are increasingly mismatched to this evolving threat landscape. In parallel, advances in cloud laboratories and automated synthesis pipelines further blur the boundary between digital design and physical realization, raising the specter of unsupervised engineering without meaningful human oversight [21, 32]. As a result, the biosecurity community is facing a defining challenge: how to safeguard the benefits of AI in biology while anticipating and mitigating its misuse.

Despite growing recognition of these risks, current strategies remain reactive, fragmented, and largely inadequate to the pace of AI-driven change. Most existing defenses rely on static regulatory lists, slow bureaucratic updates, and perimeter-style controls that assume threats can be contained within known categories [1, 7, 25]. Yet, as cybersecurity history shows, static defenses collapse in the face of adaptive adversaries and decentralized networks. Just as the "castle-and-moat" model of cybersecurity

Workshop on Regulatable ML at the 39th Conference on Neural Information Processing Systems (NeurIPS 2025).

became obsolete in a cloud-first, interconnected digital world [11], list-based biosecurity is ill-suited to counter dynamic, AI-enabled biological risks. A single overlooked vulnerability, such as split ordering of DNA fragments across multiple suppliers, can bypass current controls [16]. Furthermore, existing governance frameworks tend to discourage open threat intelligence sharing across institutions, leaving the field fragmented at precisely the moment collective defense is most urgent [29]. In short, there is a widening gap between the static tools of biosecurity and the adaptive threats posed by AI-driven biology, underscoring the need for a new strategic paradigm.

In this paper, we argue that lessons from cybersecurity provide a compelling roadmap for the future of biological AI safety. Just as cybersecurity transitioned from brittle perimeter defenses to resilience-based "zero trust" architectures, biosecurity must adopt a model that assumes breach, verifies continuously, and protects data directly rather than relying solely on boundary controls [8, 19]. We propose a framework that applies zero trust principles to biological design pipelines, ensuring continuous custody from digital model outputs to physical synthesis. Complementary measures, including adversarial stress-testing (red teaming), incentivized vulnerability discovery (bug bounties), and decentralized, privacy-preserving intelligence sharing, can further strengthen defenses [33, 18]. By embedding these practices into the fabric of biosecurity, we move from a reactive posture to one of adaptive resilience. Our perspective highlights not only the technical and governance shifts required, but also the cultural transition toward treating biosecurity as a secure-by-design principle rather than an afterthought. Ultimately, we argue that adopting a cybersecurity-informed, resilience-based model is essential for ensuring that the bioeconomy can advance safely, responsibly, and sustainably in the era of AI.

2 Background

2.1 Cybersecurity: From Perimeter to Resilience

Cybersecurity's evolution from static perimeter defenses to adaptive, resilience-based strategies offers valuable lessons for biosecurity. In this section, we trace the inadequacy of the *castle-and-moat* model, the rise of zero trust architecture, and the adaptive playbook of red teaming, bug bounties, and threat intelligence.

■■ Inadequacy of the *Castle-and-Moat* Model For decades, cybersecurity was anchored in a *perimeter defense* model, often called the *castle-and-moat* approach [9]. This model assumed a binary distinction: entities inside a defined network boundary were trusted, while those outside were threats [23]. Security tools—firewalls, intrusion detection systems, and access controls—were concentrated at this perimeter to keep adversaries out of the internal network [20]. Historical records, such as the NIST Cybersecurity Program, emphasize perimeter security, with early guidance focused on user authentication, physical safeguards, and network-level protections [27]. For example, FIPS 41 (1975) addressed physical security and system management, while FIPS 46 (1977) introduced the Data Encryption Standard (DES) to secure sensitive data within controlled environments [27].
Over time, this model became misaligned with a decentralized, interconnected digital ecosystem. Its decline stemmed not from poor design but from cloud computing, remote work, and complex supply chains that blurred traditional boundaries [11]. In this environment, a single compromised device or credential could allow lateral movement across networks, undermining internal trust [5]. The central flaw was assuming inherent trust for internal actors, creating systemic vulnerability. Consequently, perimeter-focused defenses proved untenable, failing to address threats that originated or propagated inside the network.

■■ Rise of Zero Trust Architecture As the limits of perimeter-based security became clear, the cybersecurity community adopted a fundamental shift in philosophy and architecture: the *zero trust* (ZT) model [9]. Its guiding maxim, *never trust, always verify*, rejects the assumption that anything within a network is inherently safe [11]. Zero Trust assumes adversaries may already exist inside the system, requiring every user, device, and process to be continuously authenticated and validated, regardless of location or prior authorization [19, 23]. Zero Trust is a comprehensive security philosophy rather than a single tool. Central to it is the principle of least privilege, granting entities only the access necessary and reducing the *blast radius* of a breach [11]. Frameworks such as NIST SP 800-207 and the CISA Zero Trust Maturity Model outline five pillars: Identity, Devices, Networks, Applications and Workloads, and Data [19]. Crucially, the security perimeter shifts from the network

to the data itself, protecting sensitive information at the point of request rather than assuming safety once inside a boundary [8]. This shift offers a key lesson for biosecurity. In biological AI, the "data" includes genetic sequences, protein structures, or synthetic pathogen designs that pose real-world risk. Applying Zero Trust here means securing these assets directly, creating a verifiable, resilient chain of custody from digital design to physical synthesis. Finally, Zero Trust transforms security from a reactive layer into a built-in system feature. By assuming breach as a default, it prevents lateral movement after a compromise [8, 5], making defense more adaptive and durable. Strategically applied, these principles could form the foundation for a resilient bio-AI security framework.

■■ **Adaptive Playbook: Red Teaming, Bug Bounties, and Threat Intelligence** In addition to the architectural shift toward Zero Trust, the cybersecurity community has embraced an adaptive playbook of practices designed to continuously probe, challenge, and strengthen defenses. Unlike static *defense-in-depth* strategies, this approach is adversarial, collective, and iterative, integrating red teaming, bug bounty programs, and threat intelligence into a cycle of perpetual testing and refinement [33].

⊛ **Red Teaming:** Red teaming is a structured method of stress-testing systems by simulating adversarial behavior, intentionally adopting the mindset of an attacker to uncover hidden risks and *unknown unknowns* [33]. In cybersecurity, this approach goes beyond automated vulnerability scanning by leveraging human creativity to anticipate and exploit weaknesses that are not easily detectable by algorithms [14]. Within the AI domain, red teams probe models for vulnerabilities such as prompt injections, jailbreak techniques, and unsafe outputs, thereby revealing flaws that could undermine reliability and safety [33, 17, 3]. The strength of red teaming lies in its ability to surface non-obvious vulnerabilities and stress-test systems under realistic adversarial conditions, making it a vital complement to purely technical safeguards.

⊛ **Bug Bounties:** Bug bounty programs extend the adversarial testing paradigm by engaging a distributed community of independent security researchers. Instead of relying solely on internal teams, organizations provide monetary rewards for verified reports of vulnerabilities, effectively crowdsourcing the discovery of risks [18]. This approach harnesses the diversity of perspectives and technical expertise available in the global security community, often surfacing issues that would remain invisible to conventional testing methods. Platforms such as HackerOne and Huntr, as well as companies like OpenAI, have institutionalized bug bounty programs to identify critical vulnerabilities in real-world systems before adversaries can exploit them [18]. Together, red teaming and bug bounties shift the security posture from reactive patching of known flaws to proactive identification and remediation, thereby embedding resilience into the lifecycle of technological systems.

⊛ **Decentralized Threat Intelligence Sharing:** A key challenge in a fragmented threat landscape is the reluctance of organizations to share sensitive information, often due to privacy and competitive concerns [29]. However, the recognition of a shared threat has led to the development of decentralized threat intelligence sharing mechanisms [29]. These networks, which can be facilitated by technologies like blockchain, allow peers to securely and privately exchange information about threats without exposing confidential data [29]. For example, financial institutions collaborate to identify fraudulent patterns and share them with peers, minimizing a shared risk without disclosing private customer information [29]. This demonstrates how a technical solution can overcome a fundamental governance and trust problem, fostering collective defense against a common adversary.

Taken together, these measures signal a deeper shift in security philosophy: from building fixed fortifications to cultivating resilience through constant adaptation. They are not episodic assessments but ongoing *break-fix* cycles, ensuring that defenses co-evolve with emerging threats [33]. This mindset, treating compromise as a probability rather than a possibility, is particularly vital for biosecurity, where the stakes are higher and the adversarial landscape is both dynamic and unforgiving.

2.2 New Frontier of Biosecurity: AI-Enabled Threats

Convergence of generative AI and biotechnology is creating a new frontier of biosecurity risks, where digital designs can rapidly translate into real-world biological threats. This section examines the dual-use nature of AI, the limitations of traditional list-based defenses, and the critical vulnerabilities at the digital-to-physical transition.

■■ **Generative AI as a Dual-Use Technology** Rapid development of generative AI has added a complex dimension to biosecurity. These technologies, capable of designing, optimizing, and simulating biological agents, exemplify the dual-use dilemma: tools that advance drug discovery, vaccines, and synthetic biology can also be misused to create bioweapons [2]. The associated risks fall into two interrelated domains. **First**, Large Language Models (LLMs), such as GPT-4, are concerning because they can synthesize step-by-step instructions for developing bioweapons [1]. While much of this information exists publicly, LLMs lower technical and cognitive barriers by integrating dispersed knowledge into coherent, easily understandable formats, democratizing access to sensitive biological know-how [1, 35]. **Second**, specialized Biological Design Tools (BDTs) pose an even greater threat. Unlike LLMs, BDTs leverage large-scale biological datasets to generate novel proteins, viral vectors, or synthetic pathogens [31], potentially evading current surveillance, medical countermeasures, and treatment protocols [16]. Empirical data remain limited [10], yet studies show AI can design structurally novel antibiotics and toxic compounds, exploring previously untapped regions of chemical and biological space [31]. This generative capability makes BDTs a qualitatively different and more complex biosecurity challenge. Importantly, the risks from LLMs and BDTs are synergistic. LLMs can help novice actors reproduce known pathogens, while BDTs enable sophisticated users to design entirely new agents that circumvent existing defenses [16]. Although AI cannot yet produce a fully transmissible pandemic pathogen independently, the rapid pace of AI innovation and narrowing gap between digital design and physical realization make this a pressing, near-term concern [1].

■■ **Fragility of List-Based Defenses** Analogous to the obsolescence of the *castle-and-moat* model in cybersecurity, traditional biosecurity defenses are increasingly inadequate against AI-driven threats. Current measures remain largely static and reactive, relying on list-based screening to identify known hazardous agents. In the United States, frameworks like the Select Agents and Toxins List and the Bureau of Industry and Security's Commerce Control List cover only a limited set of agents and provide restricted oversight for international transactions [1, 25, 12].

These approaches are mismatched to the dynamic, rapidly evolving nature of AI-enabled risks. Relying on static lists to contain AI-generated threats is akin to using an outdated antivirus program against a novel, self-modifying virus [1, 12]. AI tools can generate thousands of previously unseen sequences within hours, far outpacing the slow, bureaucratic processes required to update regulatory lists [1].

A clear example is *split ordering*, where malicious actors circumvent screening by ordering harmless fragments of a hazardous DNA sequence from multiple providers [16]. Each fragment appears innocuous and does not match regulated sequences, allowing current safeguards to fail [16]. This highlights a broader structural problem: static, prescriptive governance cannot keep pace with decentralized, adaptive threats. Traditional lists counter known, historical agents but leave critical gaps against AI-generated biological designs that bypass conventional defenses.

■■ ***Digital-to-Physical* Threat Vector** A key vulnerability in the AI-enabled biosecurity landscape lies in the *digital-to-physical* transition, where computational biological designs are materialized into physical agents [13]. This stage represents both the greatest risk and the most promising point for intervention. Policies like the Trump administration's AI Action Plan emphasize standardized screening and incentives to ensure safe handling of synthetic nucleic acids [13]. Yet the rise of AI-enabled *cloud labs* and *self-driving labs* has transformed this landscape, automating the full design-to-synthesis pipeline with minimal human oversight [21].

These autonomous systems create an *unsupervised engineering* risk: a viral vector could regain transmissibility, or a novel pathogen could be generated without triggering human-in-the-loop safeguards [21]. The risk extends beyond misuse of digital designs to real-world creation of hazardous agents by automated processes. Effective biosecurity must therefore address the entire lifecycle, from AI-generated output to final synthesized product. A *zero trust* pipeline, with continuous verification and strictly controlled access at each stage, offers a practical strategy to prevent unauthorized creation or modification and maintain resilience against both accidental and deliberate threats.

3 Applying the Lessons: Building a Bio-Resilient Framework

Drawing on the evolution of cybersecurity, this section explores how resilience-based strategies can be translated to biosecurity, particularly in the context of AI-enabled biological design (as illustrated

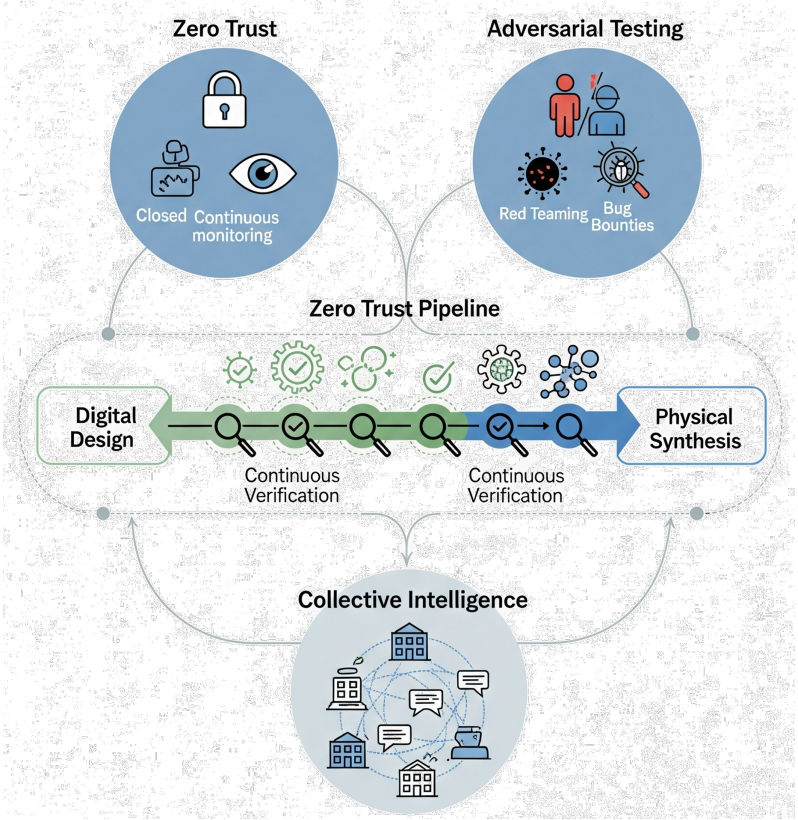

Figure 1: Summary of Key Cybersecurity Lessons Applied to Enhancing Bio-AI Safety and Resilience

in Figure 1). We examine the application of *Zero Trust* principles and proactive adversarial testing to build a robust, adaptive defense against emerging biological threats.

3.1 *Zero Trust* Model for Biological Design Tools

The rapid advancement of Biological Design Tools (BDTs) and automated laboratories has transformed biotechnology, enabling the design, synthesis, and testing of biological sequences at unprecedented speed. While these technologies accelerate beneficial applications like drug discovery and vaccine development, they also introduce significant biosecurity risks. A strategic application of Zero Trust (ZT) principles offers a robust resilience-based framework, shifting the focus from merely preventing breaches to assuming compromise is possible and containing risks proactively [28]. The Department of Homeland Security (DHS) Countering Weapons of Mass Destruction (CWMD) Office has begun exploring this approach in securing Chemical, Biological, Radiological, and Nuclear (CBRN) detection systems, demonstrating its practical applicability in high-stakes security domains.

The foundation of a Zero Trust biosecurity model lies in the establishment of a *digital chain of custody* for biological designs, ensuring that every action within the bio-AI pipeline is traceable, auditable, and continuously verifiable. This approach reframes security as a pervasive, data-centric discipline rather than a perimeter-focused afterthought. The key components of such a model include:

✓ **Identity and Access Management (IAM):** Within a Zero Trust framework, access to Biological Design Tools (BDTs) must be governed by continuous, context-aware authentication mechanisms. Permissions are dynamically adjusted based on factors such as role, task, temporal context, and behavioral intent [5]. For example, a researcher with standard access privileges may be restricted from generating high-consequence sequences or accessing datasets associated with highly pathogenic organisms. By enforcing granular, task-specific constraints, the model significantly reduces the likelihood of accidental or deliberate misuse, while enabling auditable traceability of every action taken within the system.

✓ **Device and Laboratory Controls:** Automated cloud laboratories and self-driving lab systems are treated as inherently untrusted devices under a Zero Trust paradigm. Every interaction, from the submission of digital sequence data to the physical synthesis of genetic material, is meticulously logged, verified, and continuously monitored in real time. This strategy addresses risks such as *split ordering*, in which malicious actors attempt to circumvent sequence screening by fragmenting orders across multiple suppliers [16]. Moreover, continuous validation of laboratory devices ensures that compromised instruments or software cannot be exploited to bypass security protocols, reinforcing end-to-end accountability across the entire bio-AI workflow.

✓ **Data-Centric Security:** In this framework, biological sequences themselves are elevated to the status of a primary security perimeter [8]. Security is integrated from the outset, *secure-by-design*, rather than appended retrospectively. Each sequence is accompanied by cryptographically signed metadata, which records its provenance, modification history, and usage across the pipeline [6]. Such mechanisms allow stakeholders to verify the integrity of a sequence at any point, identify unauthorized modifications, and prevent the downstream synthesis of potentially hazardous constructs.

✓ **Continuous Monitoring and Adaptive Response:** Advanced anomaly detection and behavioral analytics are central to a Zero Trust biosecurity model. Systems continuously monitor for unusual patterns of activity that could signify malicious intent or operational errors. For instance, an abnormal surge in sequence synthesis requests or unauthorized modification attempts would trigger immediate alerts and automatic containment measures. By embedding real-time monitoring and adaptive response mechanisms, the model not only identifies threats as they emerge but also enforces proactive containment strategies that mitigate the risk of escalation.

✓ **Governance, Policy, and Collaborative Oversight:** Effective deployment of Zero Trust principles in bio-AI ecosystems requires coordinated governance across multiple stakeholders, including governmental agencies, academic institutions, and commercial providers [28]. Establishing shared policies, ethical standards, and accountability frameworks ensures that the development and application of BDTs are both secure and socially responsible [36]. Collaborative governance also enables rapid dissemination of threat intelligence, alignment on best practices, and iterative refinement of security protocols in response to emerging technological and biological risks.

✓ **Integration with Adversarial Testing:** A fully resilient Zero Trust biosecurity model does not rely solely on static controls. It must be complemented by continuous adversarial testing, such as red-teaming and bio-AI bug bounty programs, to identify *unknown unknowns* and refine defenses [14, 16]. By coupling rigorous, traceable access and monitoring with proactive vulnerability discovery, the model establishes a feedback loop in which both technical and organizational safeguards evolve dynamically to address emerging threats.

By combining continuous monitoring, granular access control, cryptographically secured data, and collaborative governance, this integrated Zero Trust framework transforms biosecurity from a reactive, list-based approach to a proactive, resilience-oriented model. The emphasis shifts from post-incident investigation to real-time, preventive oversight, ensuring that every action in the bio-AI pipeline is explicitly verified and traceable. This approach not only reduces the likelihood of malicious misuse but also establishes a culture of accountability and vigilance, critical for safeguarding emerging bio-AI technologies.

3.2 Red Teaming and Bug Bounties

While implementing a Zero Trust architecture offers a rigorous foundational defense against adversarial threats, it is insufficient on its own to address the full spectrum of vulnerabilities that emerge in complex bio-AI systems. These systems are inherently adaptive, and attackers often exploit subtle, unforeseen weaknesses, what are commonly referred to as *unknown unknowns*. To effectively anticipate and mitigate such risks, continuous adversarial testing in the form of structured red-teaming and incentivized vulnerability discovery, such as bug bounty programs, is indispensable.

✓ **Structured Red Teaming:** Red-teaming constitutes a proactive, systematic approach to stress-testing bio-AI systems by simulating sophisticated, real-world attack scenarios. In this context, collaborations among government agencies, leading bio-AI developers, and independent researchers can form multidisciplinary teams that combine expertise in computational biology, artificial intelligence, cybersecurity, and ethics [14]. These teams are tasked with identifying emergent vulnerabilities that automated monitoring or static safeguards might overlook. For instance, a red team could investi-

gate the presence of so-called *sleeper agent* AIs, models engineered to produce benign outputs under routine safety evaluations but capable of facilitating malicious actions under specific conditions [16]. A recent RAND study that employed a red-teaming methodology to simulate a potential biological attack observed that contemporary AI models offered no significant advantage over rudimentary internet searches in planning such attacks [26]. However, this finding should not be misconstrued as evidence of inherent safety. Rather, it underscores a critical opportunity to implement layered protective measures before the pace of technological advancement surpasses the capacity of existing safeguards [1]. Regular, iterative red-teaming exercises can thus serve as a *continuous feedback loop*, refining both AI governance frameworks and operational protocols to reduce exposure to high-consequence risks.

✓ **Bio-AI Bug Bounties:** Complementing red-teaming, the adoption of structured bug bounty programs offers a mechanism to crowdsource vulnerability discovery across a broader research community. By incentivizing external experts to probe biological design tools (BDTs) and nucleic acid synthesis platforms, these programs help uncover flaws that traditional, rule-based approaches are unlikely to detect [4]. Potential discoveries may include novel model jailbreak techniques, misalignment in predicted outputs, or avenues for data exfiltration that compromise sensitive information [14]. Despite the potential, the deployment of biosecurity-focused bug bounty programs remains sparse, reflecting both regulatory and logistical challenges. The current lack of empirical data on the effectiveness of such programs in the bio-AI domain highlights a crucial gap in research and policy practice [10]. Consequently, these initiatives should not be treated as isolated events; instead, they must be embedded within an ongoing, iterative *break-fix cycle* in which vulnerabilities are continuously identified, addressed, and reassessed to ensure that defenses evolve in step with emerging threats [33].

✓ **Integrative Implications for Biosecurity:** When combined, red-teaming and bug bounty mechanisms provide a layered defense strategy that mirrors the dynamism of biological immune systems. Just as the immune system continuously adapts to detect novel pathogens, a robust adversarial testing framework ensures that bio-AI systems remain resilient to both known and unforeseen attack vectors. Importantly, these practices not only strengthen technical safeguards but also cultivate a culture of anticipatory governance, fostering cross-disciplinary collaboration and embedding security-conscious thinking throughout the development lifecycle of bio-AI technologies. By formalizing these practices as standard operational procedures, policymakers and developers can better align technological innovation with societal safety imperatives [1, 14].

3.3 Collective Intelligence and Decentralized Threat Sharing

In the context of systemic biosecurity threats, the imperative for collaboration outweighs competitive concerns. The biotechnology and bio-AI sectors, much like financial institutions and government agencies before them, must recognize that a successful bioweapon attack or deliberate misuse of biological design tools would undermine public trust and inflict widespread damage across the entire ecosystem [29]. Lessons from cybersecurity demonstrate that decentralized, privacy-preserving threat intelligence sharing can effectively overcome the competitive and confidentiality barriers that often impede cooperation.

Traditional threat intelligence mechanisms are frequently constrained by privacy concerns, proprietary considerations, and regulatory limitations. These challenges can be mitigated through decentralized architectures that allow secure, anonymized exchange of threat information [29]. By adopting such models, bio-AI and biotechnology organizations can establish federated Information Sharing and Analysis Centers (ISACs) [29]. Administered by trusted third parties, these ISACs would enable companies, research institutions, and government agencies to share critical alerts without exposing proprietary datasets or revealing sensitive operational details [16]. For example, a nucleic acid synthesis provider could submit an anonymized report regarding a *split ordering* attempt or a failed attempt to generate a high-risk sequence. This information could then propagate across the network, alerting other providers to potential threats without disclosing the identity of the individual or the specifics of the attempted sequence [16].

Decentralized threat sharing fosters a form of collective intelligence that is more resilient than isolated defenses. Each participant contributes insights that enhance the situational awareness of the entire network, enabling real-time adaptation to evolving threats. Moreover, it establishes a culture of shared responsibility, where the protection of public health and biosecurity is recognized as a common

good rather than a competitive liability. As bio-AI capabilities continue to advance, embedding such collaborative structures into operational and regulatory frameworks will be essential to ensuring that innovation does not outpace safety.

4 Strategic Imperatives for a Resilient Bioeconomy

■■ **Policy and Governance: From Reactive Rules to Adaptive Frameworks** Current U.S. biosecurity measures are ill-equipped to address the evolving nature of AI-enabled threats [1]. The challenge is that the current policy framework is often *tool-focused* rather than *outcome-focused* [15]. A common regulatory proxy, for example, is to regulate AI models based on the amount of compute used to train them [15]. However, this may not be a good measure of risk for BDTs, as a highly specialized model trained on a curated, high-quality but small dataset could be more dangerous than a massive, general-purpose model trained on noisy data [15].
Therefore, regulators must move away from a one-size-fits-all approach to an outcome-based one, prioritizing and evaluating only *pandemic-level risks* [22]. The Trump administration's 2025 AI Action Plan, for example, rightly identifies the dual-use threat and the need for new biosecurity strategies [1]. Policy recommendations should focus on providing funding to organizations like NIST and the Center for AI Standards and Innovation (CAISI) to continue their crucial work at the intersection of AI and biosecurity [1]. They should also focus on developing standardized, AI-enabled screening systems for nucleic acid synthesis that can detect new and augmented sequences beyond static lists [1].

■■ **Cultural Shift and Workforce Development** The most critical lesson from cybersecurity is that technology alone cannot solve the problem. The core issue is a cultural and human one, the need to embed a *secure-by-design* mindset into every stage of the development process [30]. Cybersecurity is no longer an IT silo problem; it is a business imperative that requires buy-in from all levels, including the board [34]. Similarly, biosecurity cannot be an afterthought, bolted on at the end of a project. It must be a core business requirement from the design phase, influencing the creation of the BDT itself [8]. This cultural transformation ensures that security experts are stakeholders in shaping and influencing new solutions from the start, rather than being gatekeepers at the end, thereby minimizing the risk of costly disruptions and compliance issues [30].
This also necessitates a focus on workforce development. Agencies and companies must recruit experts at the intersection of AI and biology, not just generalists [16]. This interdisciplinary training is essential for creating a new generation of *bio-cybersecurity* professionals who can understand and address the unique risks of this converging domain.

■■ **The Final Defense: Building Countermeasure Capabilities** The ultimate form of resilience is not the ability to prevent every attack but the ability to *outpace* a threat and respond to it at unprecedented speed [13]. The very AI tools that pose a threat also hold the key to our defense. The dual-use nature of AI means that investing in offensive capabilities (the ability to create novel pathogens) is inseparable from investing in defensive ones (the ability to rapidly design countermeasures) [13].
This requires a dedicated, federally-funded program to develop AI-enabled countermeasure systems that can identify a new virus and design an effective therapeutic or vaccine in days, not years [13]. AI is already demonstrating this capability, with researchers using generative deep learning to design and synthesize novel antibiotics that combat drug-resistant bacteria [24]. The ability to rapidly identify, characterize, and counter a novel, AI-generated threat is the final and most robust layer of a truly resilient biodefense system.

5 Discussion

The evolution of cybersecurity from a perimeter-centric *castle-and-moat* model to a Zero Trust architecture offers critical lessons for biosecurity in the era of AI-enabled biotechnology. Traditional defenses, which assume trust for all internal actors and focus primarily on keeping threats out, have proven insufficient in digital domains where adversaries can exploit overlooked vulnerabilities. Similarly, in the bio-AI ecosystem, static safeguards such as lists of prohibited sequences or hard-coded access controls are fragile against adaptive, dual-use technologies like generative AI. By adopting a Zero Trust approach for biological design tools (BDTs), every action, from sequence design to synthesis, is continuously verified, auditable, and constrained by least-privilege access. This

paradigm ensures that the security boundary follows the data and the actors, rather than relying on a fixed perimeter that can be breached or bypassed. For example, context-aware access controls can prevent a researcher from synthesizing a high-consequence sequence outside of approved conditions, mirroring dynamic cybersecurity policies in cloud environments [5]. Critics might argue that such continuous verification is resource-intensive and could stifle research innovation. However, the benefits of preventing catastrophic misuse, maintaining public trust, and enabling rapid response to emerging threats far outweigh the operational costs, particularly when automated monitoring and adaptive AI are employed. The convergence of bio-AI and cybersecurity best practices therefore establishes a resilient, proactive defense posture that anticipates rather than reacts to threats.

Complementing the structural principles of Zero Trust, adversarial testing through red teaming and bug bounty programs provides a mechanism to uncover *unknown unknowns* that no static policy can foresee. Multidisciplinary red teams, composed of biologists, AI experts, and security specialists, simulate high-risk scenarios such as the activation of *sleeper agent* AIs or attempts to circumvent sequence screening [16, 14]. Bug bounty programs extend this effort to the broader research community, incentivizing the discovery of vulnerabilities in both BDTs and laboratory automation pipelines. Some may contend that such testing could inadvertently expose dangerous capabilities or create vectors for misuse. Yet with careful design, anonymization, and secure disclosure protocols, these exercises reinforce defenses without increasing risk. Moreover, real-time anomaly detection and adaptive monitoring ensure that suspicious activities are contained immediately, demonstrating the practical synergy of Zero Trust principles with continuous testing. Lessons from cybersecurity show that iterative *break-fix* cycles are not only feasible but essential for maintaining resilience in dynamic threat environments [33]. In the biosecurity context, this methodology converts what is often a passive compliance exercise into an active, learning-oriented defense system, capable of evolving alongside emerging biotechnologies.

Finally, the establishment of collective intelligence through decentralized threat sharing amplifies the effectiveness of both structural and operational safeguards. Federated networks or ISACs allow companies, research institutions, and government agencies to share anonymized alerts about attempted circumventions, failed *split ordering*, or anomalous sequence requests [29, 16]. Critics might argue that such collaboration risks exposing proprietary information or creating new regulatory complexities. However, decentralized architectures and trusted third-party administration mitigate these concerns while fostering ecosystem-wide situational awareness. By pooling knowledge, participants gain early warning of emerging attack vectors and collectively adapt defenses, creating a resilience that is greater than the sum of individual efforts. This collaborative approach mirrors the cybersecurity precedent, where financial institutions and government bodies share anonymized threat intelligence to prevent systemic attacks. When integrated with Zero Trust frameworks and continuous adversarial testing, decentralized threat sharing ensures that biosecurity measures are not only robust but adaptive, capable of responding to novel threats without waiting for catastrophic events to occur.

6 Concluding Remarks

Current global biosecurity landscape is evolving rapidly, driven by the dual-use potential of AI-enabled biotechnology and the growing accessibility of advanced biological tools. Traditional static defenses, such as fixed perimeters or blacklists of prohibited sequences, are no longer sufficient to address the complexity and scale of contemporary threats. Lessons from cybersecurity show that continuous verification, dynamic monitoring, and resilience-focused frameworks are both practical and effective. In biosecurity, this translates into a Zero Trust approach, where every action within the bio-AI pipeline is auditable, verifiable, and constrained by least-privilege principles. Embedding security into the core of workflows ensures proactive protection rather than reactive mitigation.

Equally important is integrating adversarial strategies like red-teaming and bug bounty programs to identify *unknown unknowns*, complemented by decentralized, privacy-preserving threat intelligence networks. Together, these measures create a multi-layered defense ecosystem, allowing institutions to share critical insights without compromising sensitive information, detect vulnerabilities early, and respond rapidly to emerging risks. By fostering a culture of security consciousness and embedding a *secure-by-design* mindset, this holistic framework aligns technology, governance, and culture around resilience. The time to act is now: adopting a proactive, collective, and continuously evolving approach will safeguard public trust, preserve innovation, and ensure that biological advances serve humanity rather than create new threats.

References

[1] Georgia Adamson and Gregory C. Allen. Opportunities to strengthen u.s. biosecurity from ai-enabled bioterrorism: What policymakers should know, 2025. Accessed August 26, 2025.

[2] Doni Bloomfield, Jaspreet Pannu, Alex W Zhu, Madelena Y Ng, Ashley Lewis, Eran Bendavid, Steven M Asch, Tina Hernandez-Boussard, Anita Cicero, and Tom Inglesby. AI and biosecurity: The need for governance. *Science*, 385(6711):831–833, August 2024.

[3] Claudi L. Bockting, Eva A. M. van Dis, Robert van Rooij, Willem Zuidema, and Johan Bollen. Living guidelines for generative ai — why scientists must oversee its use. *Nature*, 622(7984):693–696, October 2023.

[4] Bugcrowd. Ai safety and security cybersecurity solutions, 2025. Accessed August 26, 2025.

[5] Canadian Centre for Cyber Security. A zero trust approach to security architecture - itsm.10.008, 2025. Accessed August 26, 2025.

[6] Sarah R. Carter, Nicole E. Wheeler, Christopher R. Isaac, and Jaime Yassif. Developing guardrails for ai biodesign tools. Technical report, Nuclear Threat Initiative, 2024. Accessed August 26, 2025.

[7] Yan Chen and Pouyan Esmaeilzadeh. Generative ai in medical practice: In-depth exploration of privacy and security challenges. *Journal of Medical Internet Research*, 26:e53008, March 2024.

[8] CIO Council. Federal zero trust data security guide, 2025. Accessed August 26, 2025.

[9] Cloudflare. What is a zero trust network?, 2025. Accessed August 26, 2025.

[10] Committee on Assessing and Navigating Biosecurity Concerns and Benefits of Artificial Intelligence Use in the Life Sciences, Board on Life Sciences, Division on Earth and Life Studies, Computer Science and Telecommunications Board, Division on Engineering and Physical Sciences, Committee on International Security and Arms Control, Policy and Global Affairs, and National Academies of Sciences, Engineering, and Medicine. The age of AI in the life sciences, April 2025.

[11] CrowdStrike. What is zero trust? - guide to zero trust security, 2025. Accessed August 26, 2025.

[12] Renan Chaves de Lima, Lucas Sinclair, Ricardo Megger, Magno Alessandro Guedes Maciel, Pedro Fernando da Costa Vasconcelos, and Juarez Antônio Simões Quaresma. Artificial intelligence challenges in the face of biological threats: emerging catastrophic risks for public health. *Frontiers in Artificial Intelligence*, 7, May 2024.

[13] Tal Feldman and Jonathan Feldman. The u.s. cannot prevent every ai biothreat—but it can outpace them, 2025. Accessed August 26, 2025.

[14] HackerOne. Ai red teaming: Offensive testing for ai models, 2025. Accessed August 26, 2025.

[15] John Halstead. Managing risks from ai-enabled biological tools, 2024. Accessed August 26, 2025.

[16] Melissa Hopkins. Biosecurity guide to the ai action plan, 2025. Accessed August 26, 2025.

[17] Ken Huang, Ben Goertzel, Daniel Wu, and Anita Xie. *GenAI Model Security*, page 163–198. Springer Nature Switzerland, 2024.

[18] Huntr. huntr - the world's first bug bounty platform for ai/ml, 2025. Accessed August 26, 2025.

[19] IBM. What is zero trust?, 2025. Accessed August 26, 2025.

[20] Inversion6. Your next shift: Moving between cybersecurity and cyber resilience, 2025. Accessed August 26, 2025.

[21] Nazish Jeffery, Sarah R. Carter, Tessa Alexanian, Oliver Crook, Samuel Curtis, Richard Moulange, Shrestha Rath, Sophie Rose, and Jennifer Clarke. Bio x ai: Policy recommendations for a new frontier, 2023. Accessed August 26, 2025.

[22] Johns Hopkins Center for Health Security. Ai and biosecurity: The need for governance, 2024. Accessed August 26, 2025.

[23] Keeper Security. Zero trust vs traditional security models: What's the difference?, 2025. Accessed August 26, 2025.

[24] Mirage News. Generative ai crafts novel antibiotics for key pathogens, 2025. Accessed August 26, 2025.

[25] Richard Moulange, Max Langenkamp, Tessa Alexanian, Samuel Curtis, and Morgan Livingston. Towards responsible governance of biological design tools, 2023.

[26] Christopher A. Mouton, Caleb Lucas, and Ella Guest. Red-teaming the risks of using ai in biological attacks, 2024. Accessed August 26, 2025.

[27] National Institute of Standards and Technology (NIST). Nist cybersecurity program history and timeline, 2025. Accessed August 26, 2025.

[28] Pacific Northwest National Laboratory (PNNL), Richland, WA (United States), Penny McKenzie, USDOE, Mark Watson, Travis Ashley, Jarrett Zeliff, Ernest Allard, Beau Morton, Aubrie Kendall, Riley Maltos, Ernest Tumanyan, and Eshan Singh. Zero trust strategies for chemical, biological, radiological, and nuclear detection systems: D.1 cyber scenarios. Technical report, January 2025.

[29] Vinod Panicker. Exchange of cyber threat intelligence among peers using decentralized identity networks and ioft, 2022. Accessed August 26, 2025.

[30] Clarke Rodgers. The key to delivering business value from generative ai faster, 2025. Accessed August 26, 2025.

[31] Jonas B Sandbrink. Artificial intelligence and biological misuse: Differentiating risks of language models and biological design tools. 2023.

[32] Jonas B. Sandbrink. Artificial intelligence and biological misuse: Differentiating risks of language models and biological design tools, 2023.

[33] The Microsoft Cloud Blog. Enhancing ai safety: Insights and lessons from red teaming, 2025. Accessed August 26, 2025.

[34] David Vellante and Christophe Bertrand. Cybersecurity is your no. 1 risk and you're likely unprepared, 2025. Accessed August 26, 2025.

[35] Azmine Toushik Wasi, Mahfuz Ahmed Anik, and Riashat Islam. Risks and safety considerations for foundation model-based autonomous agents' interaction with the environment. In *ICLR 2025 Workshop on Foundation Models in the Wild*, 2025.

[36] Nicole E. Wheeler. Responsible ai in biotechnology: balancing discovery, innovation and biosecurity risks. *Frontiers in Bioengineering and Biotechnology*, 13, February 2025.

A Appendix

Table 1: Comparison between the Castle-and-Moat Model and the Zero Trust Model with Biological Analogies

Aspect	**Castle-and-Moat Model**	**Zero Trust Model**
Guiding Principle	Trust but Verify	Never Trust, Always Verify
Assumptions	The network perimeter is the primary security boundary. Entities inside the network are trusted by default.	All networks are untrusted. An attacker is presumed to be present in the environment.
Key Technologies & Controls	Firewalls, Intrusion Detection Systems, traditional access controls, antivirus software.	Continuous verification, microsegmentation, least privilege, multifactor authentication, data encryption, behavioral analytics.
Strategic Goal	To prevent breaches by building strong perimeter defenses.	To contain breaches and limit their impact by protecting data and resources directly.
Biological Analogy	Similar to an organism with a strong skin or shell: defenses focus on keeping threats outside. Internal cells are assumed healthy and safe.	Similar to the immune system: constantly monitors all cells, assumes pathogens can be anywhere, and responds dynamically to threats.
Response to Breach	Breach detection may be delayed; internal trust can allow lateral movement by attackers.	Breach is localized; continuous monitoring and segmentation reduce lateral movement and impact.
Strengths	Simple, easier to implement; effective when perimeter is secure.	Highly resilient; protects sensitive data even if perimeter defenses fail; adapts to dynamic threats.
Weaknesses	Assumes internal entities are trustworthy; once breached, attacker can move freely.	More complex to implement; requires continuous monitoring and policy management.
Relevance to Biosecurity	Focus on securing lab facilities or digital systems from external intrusion.	Focus on assuming potential contamination or compromise, monitoring all vectors (lab, personnel, digital), and enforcing strict access and containment controls.

Table 2: Cybersecurity analogies and their applications in the biosecurity domain

Cybersecurity Analogy	**Description (Cybersecurity)**	**Biosecurity Application**
Red Teaming	Adversarial testing of a new software application to find vulnerabilities such as prompt injections, privilege escalations, or jailbreaks.	Multi-disciplinary teams of biologists, AI experts, and security specialists attempting to bypass BDT safeguards, trigger a *sleeper agent* AI, or identify potential pathways to generate dangerous biological designs.
Bug Bounties	Rewarding external researchers for discovering and responsibly reporting vulnerabilities in open-source software, cloud platforms, or commercial applications.	Incentivizing researchers to detect flaws in BDT model safeguards, nucleic acid synthesis screening systems, or laboratory automation pipelines, particularly for novel or high-consequence sequences.
Threat Intelligence Sharing	Organizations sharing anonymized data on attacks or vulnerabilities (e.g., phishing, fraud, malware) to strengthen collective defense.	A federated network of biotech companies, labs, and research institutions sharing anonymized data on attempted circumventions of BDT safeguards, failed *split ordering* attempts, or suspicious requests for high-risk sequences.
Patch Management	Timely application of security patches to software and systems to prevent exploitation of known vulnerabilities.	Regular updates and security improvements to BDT platforms, laboratory automation software, and screening algorithms to address newly identified biosecurity risks or regulatory changes.
Intrusion Detection Systems (IDS)	Monitoring networks and systems for suspicious activity or known attack patterns.	Continuous monitoring of sequence design, synthesis requests, and lab operations for anomalous activity, such as unusual synthesis volumes or unauthorized access attempts.
Zero Trust Architecture	A security paradigm where all users, devices, and networks are treated as untrusted and continuously verified.	Treating every lab instrument, cloud pipeline, and AI model as untrusted; enforcing dynamic access controls, device verification, and continuous auditing across the bio-AI pipeline.
Incident Response Plans	Predefined procedures for responding to cybersecurity breaches, including containment, remediation, and post-mortem analysis.	Protocols for responding to biosecurity incidents, including isolation of affected sequences, containment of laboratory systems, and coordinated reporting to regulatory authorities.
Penetration Testing	Simulated attacks against systems to evaluate security posture before real attackers exploit vulnerabilities.	Ethical simulation of malicious attempts to synthesize dangerous sequences or bypass safeguards, helping to identify gaps in the design, access, and monitoring systems of bio-AI platforms.