The Three Laws of AI Biosafety: A Constitutional Governance Framework for AI Biodesign Tools [1]

Yasmin Soltani
The George Washington University

With
Apart Research

Abstract

The proliferation of AI-enabled biological design tools presents an urgent governance challenge: existing frameworks diagnose pieces of the problem in isolation, and no unified architecture guides organizations from tool characterization through deployment and ongoing evolution. We introduce the Three Laws of AI Biosafety, a hierarchical constitutional governance framework operationalized through 21 questions across seven sub-pillars and three ordered modules: Characterize, Govern, and Evolve. Each law is a prerequisite for the next, enforcing the principle that you cannot govern what you have not characterized. Applied to eight AI biodesign tools spanning all eight established biological function categories, no tool achieved full governance readiness : five failed at Law 1 and three at Law 2. The dominant failure mode was not absent access control, but insufficient dual-use characterization. These findings suggest the biosecurity governance gap is more fundamental than previously understood, and the Three Laws framework offers a practical, auditable foundation for organizations seeking systematic governance readiness.

[1] Research conducted at the AIxBio Hackathon, April 2026

1. Introduction

AI-enabled biological design tools are advancing faster than the systems meant to regulate them. Technologies that were once confined to highly specialized labs are now accessible to a much wider group, including individuals who may misuse them. Yet the biosecurity community has largely responded with fragmented, single-purpose tools: frameworks that assess risk in isolation, propose technical guardrails without governance architecture, or raise organizational awareness without actionable remediation pathways.

This fragmentation is the core problem. No framework currently guides an organization through the full governance lifecycle of an AI biodesign tool, from understanding what it can do, to controlling who can access it, to ensuring that governance evolves as capabilities change.

To address this, we draw on two key ideas. The first is Isaac Asimov's hierarchical Laws of Robotics [2], which established that powerful tools require ordered, constitutional-level principles where higher-order rules constrain lower-order ones. The second is Anthropic's Constitutional AI approach [1], which treats governance not as an external compliance layer but as a foundational design principle embedded from the ground up. The assessment methodology builds on a Data Maturity Assessment developed by the lead author during a NASA internship, where the same challenge of translating abstract governance principles into operational, auditable tools was confronted at agency scale.

Main contributions are:

1. The Three Laws of AI Biosafety: the first hierarchical constitutional governance framework for AI biodesign tools, synthesizing existing biosecurity and AI governance literature into three ordered principles covering the full tool lifecycle.
2. A structured governance readiness assessment: 21 questions across seven sub-pillars scored on a five-level maturity scale, producing readiness scores that identify gaps and drive actionable improvement.
3. A proof-of-concept validation: application of the framework to eight publicly available AI biodesign tools spanning all established biological function categories, demonstrating differentiated, meaningful results across tools with varying governance maturity.

2. Related Work

The governance challenge addressed by this framework sits at the intersection of three active research streams: empirical risk assessment of AI biodesign tools, technical guardrail development, and organizational biosecurity vulnerability analysis.

The most directly relevant empirical foundation is the RAND Global Risk Index for AI-Enabled Biological Tools [3], which assessed state-of-the-art biodesign tools across eight biological

function categories and found no correlation between a tool's danger level and its accessibility. Our framework builds on RAND's eight categories as a classification scaffold for tool selection, and our proof-of-concept applies the Three Laws to one tool per category. Where RAND characterizes the risk landscape, our framework operationalizes the governance response.

The Nuclear Threat Initiative contributes one key piece: its technical guardrails report [4] proposed input screening, output filtering, and managed access as primary defenses for individual tools. Separately, the Netherlands Biosecurity Office developed the Biosecurity Vulnerability Scan [5], a web-based organizational self-assessment tool structured around eight biosecurity pillars, which shifts the unit of analysis from tools to the organizations deploying them. Our framework synthesizes both: Law 2 incorporates NTI's technical guardrail categories directly, while the maturity assessment structure draws on the organizational lens of the Vulnerability Scan. Neither product, however, provides a hierarchical architecture connecting characterization to safeguards to adaptive evolution. That integration is the core contribution of this work.

3. Methods

3.1 Framework Design

The Three Laws of AI Biosafety are structured as a hierarchical constitutional governance framework, where each law is a prerequisite for the next. Law 1 (Characterize) requires that any AI biodesign tool undergoes full capability and risk characterization before deployment. Law 2 (Govern) requires that characterized tools operate under built-in technical safeguards and structured access controls. Law 3 (Evolve) requires that governance frameworks are treated as living systems, revised as capabilities and scientific knowledge advance. The hierarchical dependency is intentional: an organization cannot credibly claim Law 2 compliance without first satisfying Law 1, and Law 3 is meaningless without an existing governance infrastructure to evolve. Figure 1 depicts this lifecycle as an ongoing, continuous cycle.

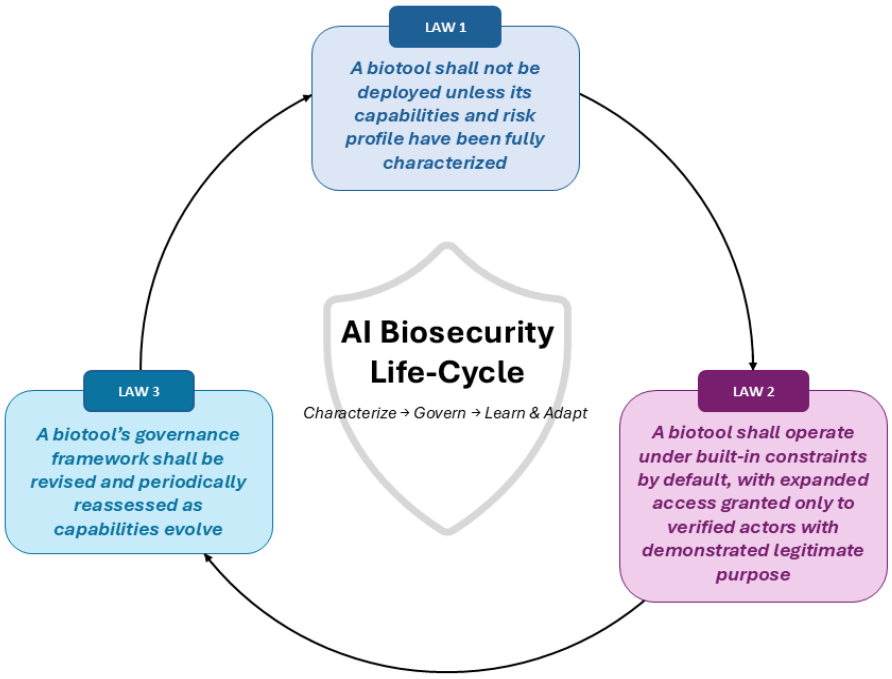

Figure 1: The AI Biosecurity Life-Cycle. The Three Laws of AI Biosafety govern the full lifecycle of an AI biodesign tool from characterization through governance to adaptive evolution. Each law is a prerequisite for the next, forming a continuous governance cycle rather than a one-time compliance checklist

3.2 Assessment Structure

The framework is operationalized through 21 questions organized across seven sub-pillars and three law modules. Law 1 comprises three sub-pillars: AI capability (use cases, dual-use potential, generalizability), technological maturity (scientific readiness, development stage, performance reliability), and availability (distribution format, deployment barriers, geographic reach). Law 2 comprises two sub-pillars: technical safeguards (input screening, output filtering, training constraints) and access control (credential verification, institutional legitimacy, purpose validation). Law 3 comprises two sub-pillars: adaptability (capability evolution, scientific advancement, organizational continuity) and iteration (systematic reflection, feedback loop, cross-sector collaboration). The complete question bank with five-level maturity descriptors for each domain is provided in [A2]. Readers can explore the full framework hierarchy interactively, including all sub-pillars and assessment questions with hover-based descriptions, at the GitHub repository linked in the Appendix [A1].

3.3 Scoring and Stop Logic

Each question is scored on a five-level maturity scale from Level 1 (Initial) through Level 5 (Optimized). Sub-pillar scores are averaged across their constituent questions, and law-level scores

across their sub-pillars. A passing threshold of 4.0 (Managed) was selected because it is the first level at which processes are formal and systematic rather than merely documented, representing the minimum bar for credible governance claims.

The framework produces four governance readiness outcomes. A Law 1 score below 4.0 yields Red, halting assessment entirely. If Law 1 passes, but Law 2 is below 4.0, that yields Amber. Both Laws 1 and 2 passing but Law 3 below 4.0 yields Yellow. All three Laws passing yields Green. While this borrows RAND's color taxonomy [3], it redefines what each color means: not a risk level assigned by an external evaluator, but a governance readiness stage an organization earns by satisfying ordered prerequisites.

The stop logic is the framework's core diagnostic contribution. It enforces the principle that you cannot govern what you have not characterized, and you cannot evolve a governance framework that does not yet exist.

3.4 Tool Selection and Evaluation

Eight AI biodesign tools were selected for proof-of-concept validation, one from each established biological function category: Addgene (viral vector design), AlphaFold2 (protein engineering), SyntheMol (small biomolecule design), Evo 2 (genome design), EVEscape (pathogen property prediction), AlphaFold-Multimer (host-pathogen interaction), NetMHCpan-4.1 (immune and vaccine design), and Coscientist (experimental automation).

All tools were scored using publicly available information only, including official documentation, peer-reviewed publications, GitHub repositories, and biosecurity literature. In production, organizations would self-assess using internal documentation; here we act as external evaluators, deliberately testing whether governance signals are legible from the outside. This mirrors the non-interactive assessment approach used by RAND [3]. Full source documentation for each tool is provided in [A3], enabling independent replication.

Sub-pillar averaging was preferred over equal weighting of all 21 questions, as it better preserves the diagnostic signal within each governance domain. Where RAND characterizes how dangerous and accessible each tool is, our framework asks whether the organizations behind those tools have the governance infrastructure to be trusted with them.

4. Results

Applying the Three Laws of AI Biosafety framework to eight publicly available AI biodesign tools produced differentiated outcomes across the governance readiness tiers. No tool achieved Green status. Five tools received Red outcomes, stopping at Law 1. Three tools reached Law 2 assessment and received Amber outcomes. No tool reached Law 3. Table 1 presents the full

scoring breakdown across all tools and sub-pillars. Detailed question-level scores and source documentation supporting each entry are provided in [A3].

	LAW 1: CHARACTERIZE			LAW 2: GOVERN		OUTCOME
Tool	**AI Capability**	**Tech Maturity**	**Availability**	**Tech Safeguards**	**Access Control**	**Tier**
Addgene	2.7	4.3	3.0	—	—	**Red**
AlphaFold2	4.0	5.0	3.0	1.3	1.0	**Amber**
SyntheMol	3.3	3.3	2.7	—	—	**Red**
Evo2	4.7	5.0	3.3	3.0	1.0	**Amber**
EVEscape	3.3	4.0	2.7	—	—	**Red**
AlphaFold-Multimer	3.7	4.7	3.3	—	—	**Red**
NetMHCpan-4.1	3.7	5.0	3.3	2.3	2.7	**Amber**
Coscientist	3.0	3.3	3.0	—	—	**Red**

Table 1: Governance readiness scores for eight AI biodesign tools across the Three Laws of AI Biosafety framework. Sub-pillar scores are averages of three constituent questions each, scored 1-5. Law-level scores are averages across sub-pillars. Dashes indicate assessment did not proceed due to the stop logic. Detailed scores breakdown are provided in [A3].

4.1 Key Findings

Five of eight tools failed at Law 1, before any governance or access control criteria were evaluated. The dominant failure mode was not technological immaturity but insufficient dual-use characterization. Across all tools, technological maturity scores were consistently high, reflecting strong scientific validation and production-ready development, while dual-use potential scores were systematically low. Tools that have been rigorously peer-reviewed and widely deployed have not undergone equivalent rigor in assessing how their capabilities could be exploited by malicious actors. AlphaFold-Multimer illustrates this precisely: a Law 1 score of 3.9, just below the threshold, driven entirely by weak dual-use characterization despite a technological maturity score of 4.7. Among the three tools that passed Law 1, none achieved sufficient Law 2 scores. Evo 2 presented the most nuanced result, scoring 3.0 on technical safeguards, reflecting genuine biosecurity investment including deliberate exclusion of human-infecting pathogen sequences from

training data and documented red-teaming, but scoring 1.0 on access control due to fully open-source distribution with no credential verification or purpose validation. We term this pattern "effort on the inside, nothing on the outside." NetMHCpan-4.1 showed the most balanced Law 2 profile with differentiated academic and commercial licensing, yet still fell short at 2.5.

4.2 Robustness and Cross-Cutting Pattern

The consistent finding across all eight tools is that organizations know what their tools do, but have not formally assessed what their tools could be made to do. This dual-use diagnosis gap appears across tools spanning all biological function categories, risk levels, and developer contexts, suggesting it is not an artifact of tool selection or scorer bias. The framework's stop logic further strengthens robustness: small variations in individual question scores would need to be both substantial and systematic to change a tool's tier outcome. Law 1 is not merely a prerequisite for the framework. It is doing the heaviest diagnostic work. The field has not yet reached the starting line of governance.

5. Discussion and Limitations

5.1 Broader Implications

Our findings suggest the AI biosecurity governance gap is more fundamental than existing frameworks have diagnosed. Prior work has focused on characterizing tool risk and proposing technical guardrails [3, 4], but these interventions are premature without a prior step: organizations must formally characterize what they are governing before guardrails can be meaningfully designed. The Three Laws framework reframes the governance challenge as sequential rather than parallel.

The "effort on the inside, nothing on the outside" pattern observed in Evo 2 points to a structural incentive problem. Developers face pressure to demonstrate responsible AI through training constraints and red-teaming, while access control creates user friction and may reduce adoption. Until external accountability frameworks address access decisions specifically, this asymmetry is likely to persist.

5.2 Limitations

All tools were scored by a single evaluator using publicly available information, meaning scores reflect observable governance rather than actual internal practice. The threshold of 4.0 and equal sub-pillar weighting are design choices made without empirical validation. Law 3 remains empirically unvalidated as no tool passed both prior laws, and the scoring rubric has not been tested for inter-rater reliability within this hackathon timeframe. The framework also does not yet address multi-tool threat models, where individually low-risk tools become dangerous in sequence.

5.3 Future Work

Immediate next steps include inter-rater reliability testing, weighted sub-pillar scoring validated through expert elicitation, extension to multi-tool threat models, and development of a self-assessment interface to shift the framework from external audit to internal governance tool.

6. Conclusion

This paper introduced the Three Laws of AI Biosafety, a hierarchical constitutional governance framework for organizations developing or deploying AI biodesign tools, operationalized through a 21-question maturity assessment across seven sub-pillars. Applied to eight tools spanning all established biological function categories, the framework found that no tool achieved full governance readiness, and that the dominant failure mode is insufficient dual-use characterization at Law 1, before access controls or evolution mechanisms are even evaluated. The findings suggest that governance frameworks for AI biodesign tools should follow a phased, sequential approach rather than being implemented all at once: characterization is not a step that can be bypassed in favor of more visible interventions, but the foundational layer upon which all other governance depends.

References

1. *Anthropic. (2025). Claude's Constitution. Anthropic.* *https://www.anthropic.com/constitution*
2. *Asimov, I. (1942). Runaround. Astounding Science Fiction.*
3. *RAND Europe and Centre for Long-Term Resilience. (2025). Global Risk Index for AI-Enabled Biological Tools. RAND Corporation.* *https://www.rand.org/pubs/external_publications/EP71093.html*
4. *Nuclear Threat Initiative (NTI). (2025). Developing Guardrails for AI Biodesign Tools. NTI Bio.* *https://www.nti.org/analysis/articles/developing-guardrails-for-ai-biodesign-tools/*
5. *Meulenbelt, S.E. et al. (2019). The Vulnerability Scan, a Web Tool to Increase Institutional Biosecurity Resilience. Frontiers in Public Health, 7:47.* *https://doi.org/10.3389/fpubh.2019.00047*

Appendix

[A1] Interactive Framework Visualization An interactive visualization of the Three Laws of AI Biosafety governance hierarchy, including all seven sub-pillars and 21 assessment questions with hover-based descriptions, is available at: https://yasminsoltani.github.io/AIxBIO-Hackathon

[A2] Supporting Document 1: Assessment Framework The complete 21-question maturity assessment bank, organized by Law and sub-pillar, with five-level maturity descriptors (Initial through Optimized) for each question domain: https://docs.google.com/document/d/1teYGi1PXgcCmoxNM3JaFnREmM84ODZea/edit?usp=sharing

[A3] Supporting Document 2: Assessment Results Full scored assessments for all eight AI biodesign tools, including sub-pillar scores, overall Law scores, governance readiness outcomes, and documented public sources used for each evaluation: https://docs.google.com/document/d/1cyKdthjNms3wE-PY9rswup11z4LlduPo/edit?usp=sharing

LLM Usage Statement

We used Claude (Anthropic) to support drafting and editing sections of this report. The framework design, assessment questions, tool selection, and all scores were developed and verified independently by the author using publicly available sources documented in the Appendix. All claims and results were independently verified.