IntuitionLabs

INTUITIONLABS / RESEARCH & PRACTICAL GUIDANCE

Claude Biosecurity Controls for Pharma: 2026 Analysis

> Build practical AI for pharma and biotech with IntuitionLabs. We help life-science teams turn complex information and workflows into useful software, governed knowledge systems and AI tools.

Published by IntuitionLabs · 2026-09-24 · pharma ai governance · dual-use ai biology · biotech security · biosecurity governance · life sciences ai controls · ai risk management in biotech · tabletop exercises · durc

Original article: https://intuitionlabs.ai/articles/claude-biosecurity-controls-pharma

In this article

Executive Summary

Anthropic's September 2026 threat report changes the governance question for pharmaceutical and biotechnology organizations. The relevant question is no longer whether a model will always refuse a hazardous prompt. It is whether the institution can bind every sensitive interaction to a verified person, an approved research purpose, a known access path, continuous review, and an accountable response process.

The defensive architecture recommended here has four linked layers. First, verify identity, employer or institutional affiliation, geography, and the true route to the model, including resellers. Second, authorize a **project**, not merely an account, with the scientific aim, model and tool access, data classes, funding and affiliation, Institutional Biosafety Committee (IBC) or equivalent review, and expiry recorded. Third, monitor the whole project context for identity churn, routing changes, repeated refusals, obfuscated intent, tool escalation, and long-running accumulation of capabilities. Fourth, preserve proportionate evidence and run a practiced response process. Anthropic's new Life Sciences Verification Program (LSVP) reviews research credentials, security standards, and ethical oversight (www.anthropic.com), but provider verification does not replace institutional scientific review.

The global governance frame is the World Health Organization's Biorisk Implementation and Evaluation Tool (BRIET), launched **August 28, 2026**. WHO describes it as a voluntary, structured self-assessment and capacity-strengthening tool (www.who.int). It supports identifying gaps, prioritizing actions, and monitoring progress (www.who.int). BRIET data remain on the user's systems (www.who.int). That property helps an institution assess itself without exporting the assessment, but it does not decide what model telemetry the institution or vendor must retain.

Quantitative governance should use local denominators rather than invented universal thresholds. Track verified-user coverage, approved-project coverage, monitored-session coverage, median and tail human-review latency, repeat-refusal rate, reseller traffic share, and time to containment. Capability tests are not safety certificates: one AISI evaluation used **five models and more than 600 private expert-written**

questions (www.aisi.gov.uk), yet AISI calls such results a snapshot (www.aisi.gov.uk). The decision standard should therefore be evidence that controls work together, with ownership, records, exercises, and remediation, rather than confidence in model refusals alone.

Introduction and Background

Claude and other frontier models can assist literature analysis, coding, data interpretation, protein research, toxicology, and experimental planning. The same versatility complicates governance. WHO defines dual-use research as beneficial work that may generate knowledge, methods, products, or technologies that could intentionally be misused (www.who.int). Its narrower category, dual-use research of concern (DURC), covers work readily misapplied with little or no modification (www.who.int).

This distinction matters because a pharmaceutical project can be institutionally legitimate and still produce hazardous capabilities. A clean corporate email address does not establish that a specific project has the right approvals. Conversely, technical terms alone do not establish harmful intent. The proper unit of governance is the combination of user, affiliation, project, data, model, tools, route, and time.

The control vocabulary also needs precision. WHO describes **biosafety** as preventing unintended exposure or accidental release (ihrbenchmark.who.int), while **biosecurity** addresses unauthorized access, loss, theft, misuse, diversion, or intentional release (ihrbenchmark.who.int). AI governance can affect both, but this report concentrates on deliberate misuse prevention and the institutional controls that also reduce accidental pathways.

Independent evidence also cautions against static assumptions. The United Kingdom AI Security Institute (AISI) reports that tool access advances automation of complex tasks that precede wet-lab work (www.aisi.gov.uk), while current models still struggle with some end-to-end biological-design workflows (www.aisi.gov.uk). RAND's 2024 red-team study found no statistically significant improvement in plan viability from model assistance relative to internet-only access (www.rand.org), but warned that the study did not measure the distance to a concerning capability threshold (www.rand.org). These findings are compatible: capability is uneven, changing, and sensitive to scaffolding.

For IntuitionLabs, an adjacent life-sciences AI consultancy rather than a model provider, the practical perspective is operational. Its first-party description emphasizes governed information, specialist implementation, role-based adoption, and measured results (intuitionlabs.ai). That posture belongs in implementation guidance, not in a vendor-comparison row.

Key Changes

Stronger Model and Access Safeguards

As of September 23, 2026, Anthropic says Fable 5.1 is generally available, while Mythos 5.1 is available only through its trusted access programs. In Anthropic's August 7, 2026 description of the earlier Fable 5 model, a biological classifier can route a flagged request to the less biologically capable Opus 5 (www.anthropic.com).

Its LSVP gives verified life-sciences professionals access to Mythos, Opus, and Sonnet under refined safeguards (www.anthropic.com). Anthropic says Opus 5.5 launches with a similar class of safeguards to Fable 5.1 in biology.

These measures are relevant, but their governance meaning must be precise. **Classifier action** changes the model path for a request but does not approve an institutional study. **Professional verification** establishes evidence about a user and organization but does not authorize every project. **Trusted access** narrows who receives a capability but does not remove local supervision and tool controls. **Provider monitoring** observes the provider's service but may not see local data preparation, external tools, or downstream laboratory actions. **Institutional review** evaluates the research purpose, people, facilities, funding, materials, and downstream consequences within local authority.

Independent research reinforces the need for multiple layers. RAND describes access controls as requiring complementary detection that identifies and disrupts misuse patterns (www.rand.org). That supports an architecture in which verification, authorization, monitoring, and response fail independently and create opportunities for human intervention.

The core change is therefore not a single refusal rule. It is a shift from open, prompt-level access toward trusted-user programs, account signals, and layered monitoring. Anthropic itself says sensitive access requires account and institutional signals (www.anthropic.com).

Why Beneficial and Harmful Intent Are Hard to Separate

Intent is not directly observable, and scientific language is not a reliable proxy. WHO's definitions focus on potential use of outputs, not only stated motive. Anthropic likewise says biological capabilities can serve beneficial or harmful purposes (www.anthropic.com).

Three implications follow. **Legitimacy is contextual:** affiliation, funding, facilities, approvals, and expected outputs must cohere. **Risk is cumulative:** a sequence of individually permissible requests can assemble a higher-risk workflow. **Controls must be reversible:** access should expire, narrow, pause, or step down as evidence changes.

This is also why capability evaluations cannot designate a system safe. UK government guidance says evaluation is not meant to confer that label (www.gov.uk). A Nature study notes that saturation of public benchmarks limits precise measurement of frontier capabilities (www.nature.com). Governance must remain responsive to new models, tools, and access patterns.

Provider and Institution Responsibilities

Model providers and research institutions observe different parts of the system. Neither can substitute for the other. The provider controls model deployment, service policy, classifiers, and much of the platform telemetry. The institution controls employment, project approval, local data, connected tools, research facilities, and its own response obligations.

Table 2 separates the primary duties while identifying shared interfaces.

Control area	Model-provider responsibility	Pharma or biotech responsibility	Shared evidence
Identity and geography	Verify accounts, supported-region eligibility, and abnormal platform access. Anthropic considers majority ownership in supported-region eligibility (www.anthropic.com).	Verify worker, contractor, guest, institution, role, and permitted work location.	Stable user ID, organization ID, country and access-route record
Project authorization	Offer program eligibility and capability tiers.	Approve a named project, scope, models, tools, datasets, duration, and scientific accountable owner.	Project ID carried with each request
Reseller control	Enforce contractual routing and service terms. Anthropic's commercial terms restrict unapproved resale (www.anthropic.com).	Prohibit hidden relays, inventory every intermediary, and reconcile bills with gateway logs.	End-user attribution and route lineage
Content safeguards	Run classifiers, refusals, tiering, and abuse detection.	Set stricter local rules where project, facility, or regulatory context requires them.	Reason code, model selected, reviewer outcome
Tool and data access	Secure hosted tools and document their scope.	Apply least privilege to repositories, code execution, synthesis ordering, laboratory systems, and export paths.	Authorization decision and tool-use event
Evidence retention	Publish retention behavior and preserve agreed safety signals. Human access to covered-model conversations is recorded in a tamper-proof log (support.claude.com).	Define legal, scientific, privacy, and security retention for local records.	Retention schedule, access record, deletion evidence
Response and learning	Restrict service, investigate platform patterns, and update safeguards.	Pause projects, preserve records, protect facilities and data, notify internal authorities, and remediate.	Joint contact path, timestamps, after-action record

The important boundary is decision rights. A provider may decide whether its service will answer. An institution must decide whether the research may proceed and under what conditions. The controls should exchange only the minimum evidence required for those decisions.

Identity, Affiliation, Geography, and Reseller Controls

Identity proofing should establish that the claimed identity exists and belongs to the applicant, consistent with NIST's objective (pages.nist.gov). For sensitive access, identity alone is insufficient. IBBIS recommends confirming life-sciences affiliation and a legitimate reason at onboarding (ibbis.bio). Its project-level approach considers intended use, institutional approvals, and biorisk-management practices (ibbis.bio).

For stronger assurance, NIST specifies collection of core identity attributes, including at least one government identifier (pages.nist.gov). In a neighboring high-consequence supply chain, UK synthesis-screening guidance also tells providers and third-party vendors to verify customer identity (www.gov.uk). These sources support stronger identity assurance, not indiscriminate collection of credentials.

Minimum controls should include:

- **One person, one durable identity:** map workforce, collaborator, and service identities to one subject record.
- **Affiliation evidence:** validate employer, laboratory, role, principal investigator, and sponsor independently.
- **Project binding:** require an active project ID on sensitive model, data, and tool calls.
- **Geography evaluation:** compare approved location with actual device, network, and account signals.
- **Ownership review:** assess institutional and reseller ownership where provider regional policy depends on it.
- **Route inventory:** record direct API, cloud marketplace, managed platform, broker, and downstream model hops.
- **Anti-churn controls:** link replacement accounts, credentials, devices, and payment or contract identifiers.
- **Periodic attestation:** require owner confirmation when staff, scope, funding, model, tools, or location changes.

The UK National Cyber Security Centre recommends combining user, device, status, and location signals for access decisions (www.ncsc.gov.uk). It also treats access from a different geographic region as potentially unexpected behavior (www.ncsc.gov.uk). CISA's zero-trust goal is granular enforcement (www.cisa.gov). For a research platform, that means per-request authorization based on subject, project, model, tool, data class, location, and current risk state.

Implementation Considerations and Process Changes

Project-Level Context and Risk Tiering

An effective intake record should be compact enough to maintain but rich enough to support review. It should capture:

- **Scientific purpose:** beneficial objective and decision the model will support.
- **Biological scope:** organism, material, function, and hazard class at a governance level, without copying sensitive procedures into the approval summary.
- **People and affiliations:** principal investigator, users, collaborators, funder, and facilities.
- **AI scope:** permitted models, capability tier, system prompts, retrieval sources, code or search tools, and agent autonomy.
- **Downstream scope:** whether outputs can reach design software, ordering, automation, laboratory control, or external partners.
- **Controls and approvals:** IBC or equivalent decision, biosecurity review, security owner, data owner, conditions, and expiry.
- **Monitoring plan:** events collected, reviewer, alert route, privacy basis, and retention schedule.
- **Change triggers:** new model, new tool, new location, new collaborator, expanded biological aim, refusal pattern, or route change.

Risk tiering should be locally defined. A practical model uses consequence, plausibility, model capability, tool access, user assurance, project assurance, and monitoring strength. It should never infer a universal numerical threshold from a vendor case report. NIST's generative AI profile allows organizations to tailor measurement to system characteristics (doi.org) and calls for reviewed approval thresholds tied to measurement (doi.org).

IBC and Biosecurity Review

An Institutional Biosafety Committee can be part of the workflow, but the organization should define when the IBC acts, when a dedicated dual-use or biosecurity body acts, and when both are required. U.S. NIH guidance expects procedures for initial and continuing IBC review and approval (osp.od.nih.gov). It also describes robust risk assessment for appropriate biocontainment (osp.od.nih.gov). Canadian guidance likewise identifies assessment of dual-use potential as a possible IBC function (www.canada.ca).

The review system also needs trained participants. NIH guidance calls for appropriate training for the IBC chair, members, biosafety officer, investigators, and staff (osp.od.nih.gov). Training should include how model context, tool use, and reseller routes change the evidence available to a biological reviewer.

As of September 19, 2026, U.S. policy requires careful date labeling. The policy approved on **July 20, 2026** replaced the 2024 DURC and pathogen policy (www.aspr.gov). NIH said its identified potential high-risk activities would remain paused until NIH-specific implementation requirements were established (www.grants.nih.gov). Institutions should therefore verify current agency implementation rather than operationalizing the superseded 2024 policy.

The transition timetable is itself material. NIH reported **120 days** for departments and agencies to publish implementation guidance (www.grants.nih.gov) and **90 days** to establish a single independent third-party review body (www.grants.nih.gov). ASPR's summary says researchers, institutions, and funders must review and attest that research is outside prohibited categories before work or funding proceeds (www.aspr.gov). Legal and research-policy owners should verify the operative agency requirements on the date of each decision.

A decision tree can be implemented as six questions. **User:** is the person verified and affiliated? If no, deny sensitive access and route to onboarding. **Project:** is there an active approval? If no, allow only low-risk general use or require intake. **Scope:** does the request match that approval? If no, pause for human disposition. **Change:** has the model, tool, route, geography, or affiliation changed? If yes, re-evaluate authorization. **Accumulation:** does cumulative context raise the risk tier? If yes, step down capability or suspend pending review. **Readiness:** are approvals, monitoring, and evidence retention current? If no, do not proceed.

Online Monitoring, Refusal, and Escalation

Monitoring should join platform events with institutional context. A single prompt detector cannot see a weeks-long sequence unless conversations are linked to the same user and project. Nor can it distinguish an authorized program from an unauthorized one without approval data.

Monitoring also needs a review cycle. NIST's generative AI profile calls for ongoing monitoring and periodic risk-management review (doi.org). AISI recommends testing across the lifecycle when risk is expected to increase, rather than relying on a universal calendar cadence (www.aisi.gov.uk). Model replacement, new tools, major prompt architecture changes, and expanded downstream access are sensible local triggers.

Monitor at least these signals:

- **Repeated refusals:** subsequent rephrasing, fragmentation, or migration to another model.
- **Route change:** movement from direct access to a broker, relay, personal account, or unapproved endpoint.
- **Identity churn:** new accounts or credentials linked to the same project, device, organization, or payment path.
- **Geographic inconsistency:** access that does not match approved sites or provider regions.
- **Obfuscated context:** deliberate low-detail labels, unexplained aliases, or unstable project descriptions.
- **Tool escalation:** addition of code execution, external search, design tools, ordering, robotics, or laboratory interfaces.
- **Scope accumulation:** individually low-risk requests forming a higher-risk workflow over time.
- **Evidence gaps:** missing project ID, expired approval, unknown data source, or unlogged downstream action.

Escalation should be proportional and reversible. Level 1 can request context. Level 2 can narrow tools or route to a weaker capability. Level 3 can freeze the session and require biosecurity review. Level 4 can suspend project access, preserve relevant evidence, and invoke the institution's incident process. The labels and triggers should be approved locally, tested against representative workflows, and reviewed when models change.

Privacy, Zero Data Retention, and Evidence

Privacy and safety monitoring create a real design trade-off. Under standard API retention, Anthropic says it deletes inputs and outputs within **30 days**, subject to stated exceptions (privacy.claude.com). For LSVP traffic, Anthropic requires **30-day retention** to support offline monitoring (www.anthropic.com). Under approved zero-data-retention arrangements, it still retains user-safety classifier results for policy enforcement (privacy.claude.com).

The institution should document which evidence exists in each layer:

- **Provider content:** prompts and outputs, if retained under the selected program.
- **Provider safety metadata:** classifier outcomes, refusal reasons, model routing, and enforcement signals.
- **Institutional gateway:** user, project, model, timestamp, route, tool, decision, and correlation identifiers.
- **Research systems:** source data, code, output destination, approval version, and downstream action.
- **Human review:** reviewer, evidence examined, disposition, rationale, conditions, and closure time.

Evidence access must also be auditable. ENISA recommends logging privileged operations delegated to a healthcare cloud provider (www.enisa.europa.eu). UK synthesis-screening guidance recommends retaining user information for at least **three years** (www.gov.uk). Neither rule should be copied automatically into model telemetry policy, but both show why retention, access, and audit evidence should be explicit.

Collect the minimum content necessary, separate highly sensitive scientific content from metadata, restrict reviewer access, and set purpose-specific deletion rules. FDA guidance provides a useful record principle for regulated computerized systems: documentation should show who made a change, when, and why (www.fda.gov). It also says audit trails should be retained at least as long as associated records (www.fda.gov). Those statements do not create a universal retention period for AI biosecurity telemetry. They illustrate why retention must be tied to the governing record and use case.

WHO BRIET to AI Control Crosswalk

BRIET is an institutional self-assessment, not a model certification. WHO says it is intended for authorities, research institutions, and laboratories (www.who.int). Its value is the cycle from assessment to prioritized action and later monitoring.

Table 3 translates that cycle into evidence for an AI-enabled research environment.

BRIET-aligned function	AI control question	Evidence to retain	Accountable owner
Governance and scope	Which AI-supported biological work is in scope, and who can stop it?	Policy, system inventory, committee charters, decision rights	Executive sponsor, biosecurity officer
Roles and competence	Are reviewers trained in both biological risk and AI workflow behavior?	Role matrix, training record, reviewer coverage, conflict process	IBC chair, AI governance lead
Risk assessment	Is risk assessed at project level across user, model, data, tools, route, and downstream action?	Intake, tier rationale, approvals, expiry, change history	Scientific review body
Controls and operation	Are identity, least privilege, monitoring, refusal, escalation, and retention functioning together?	Access decisions, gateway logs, control tests, exceptions	Platform and security owners
Incident readiness	Can the institution contain access, preserve evidence, protect research systems, and coordinate with the provider?	Playbook, contacts, exercise record, containment timestamps	Incident commander
Improvement	Are gaps prioritized, assigned, funded, tested, and closed?	Action plan, due dates, validation result, residual-risk acceptance	Governance committee

This crosswalk deliberately leaves scoring and thresholds local. WHO describes BRIET as voluntary (www.who.int). It is not a law, an approval to conduct research, or evidence that a specific Claude deployment is safe. WHO's broader framework addresses stakeholders across the research lifecycle (www.who.int), while NIST organizes AI risk management into Govern, Map, Measure, and Manage (airc.nist.gov). Together, they support a continuous management system rather than a one-time checklist.

The evidence base should remain comparable across model changes. NIST recommends a document-retention policy that preserves the history of testing and evaluation (doi.org). That history helps explain whether an apparent improvement reflects a model, a control, a changed test, or a different user population.

Data Analysis and Evidence

No public evidence establishes a universal safe refusal rate, reviewer response time, or monitoring threshold for pharmaceutical use of frontier models. The correct quantitative approach is to define a population, eligibility rule, unit of analysis, time window, and data owner for each metric. Percentages should use a qualifying numerator divided by the eligible denominator, multiplied by 100. Older NIST measurement examples express that basic percentage construction (nvlpubs.nist.gov), but local metric definitions should follow the organization's current measurement policy.

Recommended operating metrics are:

- **Verified-user coverage:** verified active sensitive users divided by all active users eligible for sensitive access, times 100.
- **Approved-project coverage:** sensitive interactions carrying a current approved project ID divided by all sensitive interactions, times 100.
- **Monitor coverage:** sensitive interactions successfully evaluated by required online and offline controls divided by all sensitive interactions, times 100.
- **Human-review latency:** time from queue entry to recorded disposition; report median and upper-tail values, not only an average.
- **Repeat-refusal rate:** users or sessions with a new related attempt after refusal divided by the explicitly defined prior-refusal cohort, times 100.
- **Reseller traffic share:** sensitive interactions reaching the model through approved intermediaries divided by all sensitive interactions, times 100.
- **Time to containment:** effective-containment timestamp minus detection timestamp, consistent with ENISA's detection-to-containment formulation (www.enisa.europa.eu).

Each result needs a denominator health check. A high verified-user percentage can conceal uncounted personal accounts. High monitor coverage can conceal dropped events if total traffic is derived from the monitoring system itself. Low refusal recurrence can mean strong deterrence, or it can mean users moved to an invisible channel. Reconcile independent sources such as identity systems, procurement, network egress, provider billing, research inventory, and gateway logs.

Capability evidence should also be interpreted with uncertainty. AISI evaluated **five language models** on **more than 600** private expert-written chemistry and biology questions (www.aisi.gov.uk). Human experts had web access and up to **one hour per question** (www.aisi.gov.uk). Its reported agreement was **0.52** between an automated grader and humans, compared with **0.8** between humans (www.aisi.gov.uk). These details show why a single score should not become an access decision by itself.

Benchmark drift matters as well. AISI reported a **+0.6 relative result** against its biology expert baseline for tested 2025 models (www.aisi.gov.uk). RAND later found that many dual-use benchmark tasks were saturated while harder discriminating tasks remained (www.rand.org). The institution should connect evaluation changes

to review triggers, not assume that a model name denotes a stable risk tier forever.

Tabletop Exercise and Audit Checklist

A tabletop exercise should test decisions and evidence flows without reproducing hazardous biological detail. A useful hypothetical scenario is a legitimate, approved virology project whose user begins accessing through an unrecorded reseller after several refusals, from a new region, while adding code execution and an external design tool. The exercise is explicitly a **Hypothetical Example** and should use fictitious organizations, users, and data.

Exercise injects can proceed as follows:

1. **Onboarding discrepancy:** affiliation is valid, but the model account is personal and not linked to the approved project.
2. **Route change:** traffic begins arriving through an intermediary not listed in procurement records.
3. **Refusal recurrence:** similar objectives appear in fragmented requests across two model tiers.
4. **Location signal:** the access location conflicts with the project's approved sites.
5. **Tool expansion:** the user connects code execution and an external biological design service.
6. **Privacy constraint:** content is not centrally retained, but safety and gateway metadata remain available.
7. **Decision pressure:** the scientific owner argues that a deadline justifies continued access.
8. **Containment:** the team must narrow or suspend access without disrupting unrelated approved work.
9. **Recovery:** reviewers decide what evidence and approval changes are required to resume.
10. **Improvement:** owners assign corrective actions and validate closure.

Australian guidance offers discussion-based tabletop formats lasting **30 to 120 minutes** (www.cyber.gov.au), but duration should follow the exercise objective. ENISA says metrics may be qualitative or quantitative and should be framed around exercise objectives (www.enisa.europa.eu). FEMA's improvement guidance emphasizes continuously monitoring corrective actions (preptoolkit.fema.gov).

After the exercise, conduct an after-action assessment. NIST's generative AI profile recommends this for incidents to verify the response process (doi.org). If biological records fall under other institutional regimes, their retention must be reconciled separately. For example, Canada's biosafety standard requires some biosafety and biosecurity incident records to remain on file for at least **10 years** (www.canada.ca).

The audit checklist should confirm:

- **Inventory:** every model, endpoint, intermediary, plug-in, tool, and service account is recorded.
- **Identity:** every sensitive user maps to a verified person, affiliation, manager, and research role.
- **Project:** every sensitive interaction carries a current approved project identifier.
- **Approval:** scientific, biosafety, biosecurity, security, and data decisions are traceable and current.
- **Least privilege:** model, tool, dataset, export, and laboratory permissions match project scope.
- **Routing:** provider, cloud, broker, and reseller paths preserve end-user and project attribution.
- **Monitoring:** events arrive completely, controls execute, alerts route, and failures are visible.
- **Human review:** queues have qualified coverage, recorded disposition, and conflict handling.

- **Retention:** content and metadata schedules are explicit, lawful, access-controlled, and testable.
- **Response:** contacts, containment mechanisms, evidence preservation, and restoration steps work.
- **Metrics:** numerators, denominators, timestamps, exclusions, and independent reconciliations are documented.
- **Change management:** new models, tools, regions, affiliations, and purposes trigger reassessment.

Implications and Future Directions

Three conclusions should shape 2026 roadmaps. First, model capability and safeguard performance will move at different rates. AISI reports that stronger capability does not reliably imply stronger safeguards (www.aisi.gov.uk). Risk classification must therefore consider both capability and the complete access architecture.

Second, biological AI governance will increasingly resemble controlled research infrastructure rather than an ordinary productivity-software rollout. Project identifiers, verified affiliation, least privilege, tool lineage, continuing review, and evidence-quality requirements will become more important as models gain tool access. ISO/IEC 42001 frames an AI management system as something established, maintained, and continually improved (www.iso.org). OECD's accountability principle similarly calls for traceability of datasets, processes, and decisions (www.oecd.org).

Third, privacy architecture must become more deliberate. Local BRIET data storage supports institutional control (www.who.int), while model-provider monitoring may require specific retained evidence. The design goal is not maximal collection. It is sufficient, access-controlled, purpose-bound evidence to make and defend decisions.

An adjacent advisor can help an institution define operating processes, integration, measurement, and assurance, but cannot confer model safety or research approval. IntuitionLabs describes measurement across workflow penetration, time recovered, quality, risk signals, reliability, and support burden (intuitionlabs.ai). In this context, that evidence-first approach should be extended to biosecurity controls, with the institution retaining decision authority.

Frequently Asked Questions (FAQs)

What Claude dual-use AI controls are needed for pharmaceutical research?

At minimum: verified identity and affiliation, approved project scope, supported geography and route, least-privilege model and tool access, project-linked monitoring, a human escalation path, retention rules, periodic review, and a tested containment process. More capable or more connected workflows need stronger assurance. This is dual-use AI governance for life sciences in operational form.

Do Claude's biological safeguards replace institutional DURC review?

No. Provider safeguards decide how a service responds and who receives certain access. Institutional oversight decides whether the research is authorized, properly supervised, and supported by suitable facilities, controls, and approvals. NIH guidance expects continuing review for work in its scope (osp.od.nih.gov).

How should AI biosecurity risk in pharma be tiered?

Use locally approved factors: consequence, plausibility, model capability, data sensitivity, tool and laboratory connectivity, user and project assurance, route, monitoring, and reversibility. Record the rationale and define triggers for re-review. Do not import a universal score from a vendor report or benchmark.

How can pharma and biotech prevent AI misuse in drug discovery?

Responsible AI controls for biotech start with separating routine productivity use from sensitive scientific workflows. For the latter, verify the user and affiliation, approve the project, restrict model and tool access, preserve end-user attribution through every intermediary, monitor cumulative context, and require human review when scope or risk changes. Drug-discovery governance should cover external design tools, code execution, ordering interfaces, and laboratory automation as well as the conversational model.

Can zero data retention coexist with misuse prevention?

Sometimes, but only if the remaining safety metadata and institutional records are sufficient for detection, review, and response. The organization should map exactly which content and metadata exist at the provider, gateway, research system, and human-review layers.

What is the role of WHO BRIET?

BRIET is a voluntary institutional self-assessment that helps identify gaps, prioritize actions, and monitor progress. It offers a global governance frame. It is not law, a product certification, or a substitute for applicable national requirements and local research approval.

How should model-shopping after a refusal be handled?

Treat it as a contextual signal, not automatic proof of intent. Link attempts across users, projects, accounts, routes, and model tiers. A qualified reviewer should determine whether the activity is in scope, needs clarification, requires narrower capability, or should be paused.

Conclusion

Claude biosecurity controls for pharma should be designed as a research-governance system, not a prompt filter.

The defensible architecture binds a verified person to a verified institution and an approved, expiring project. It authorizes model, data, tools, route, geography, and downstream actions at the narrowest practical level. It monitors the project over time, escalates uncertain cases to qualified humans, preserves proportionate evidence, and can contain access without stopping unrelated work.

WHO BRIET supplies a useful global cycle of self-assessment, gap identification, prioritized action, and progress monitoring. Current national rules and institutional biosafety duties still require separate verification. Vendor enforcement and institutional oversight remain complementary, with different evidence and decision rights.

The final readiness test is operational: the organization can show coverage metrics with valid denominators, explain every exception, trace a sensitive request to its user and project, reproduce an approval decision, contain a simulated event, and verify that corrective actions were completed. Until those capabilities exist, model refusals alone are not an adequate control architecture for dual-use biological research.

Source links

Links cited in this article, in order of first appearance. Full addresses are provided for readers using extracted text.

1. https://www.anthropic.com/news/life-sciences-verification-program
2. https://www.who.int/news/item/28-08-2026-who-launches-the-biorisk-implementation-and-evaluation-tool-for-biorisk-management
3. https://www.aisi.gov.uk/blog/advanced-ai-evaluations-may-update
4. https://ihrbenchmark.who.int/document/glossary
5. https://www.aisi.gov.uk/frontier-ai-trends-report
6. https://www.rand.org/pubs/research_reports/RRA2977-2.html
7. https://intuitionlabs.ai
8. https://intuitionlabs.ai/articles/claude-fable-5-1-mythos-5-1-biology-access
9. https://www.anthropic.com/news/improving-fable-5-s-biology-safeguards
10. https://www.rand.org/pubs/research_reports/RRA4999-1.html
11. https://www.anthropic.com/threat-intelligence-report-september-2026
12. https://www.gov.uk/government/publications/ai-safety-institute-approach-to-evaluations/ai-safety-institute-approach-to-evaluations
13. https://www.nature.com/articles/s41586-025-09962-4
14. https://www.anthropic.com/supported-countries
15. https://www.anthropic.com/legal/commercial-terms
16. https://support.claude.com/en/articles/15425996-data-retention-practices-for-covered-models
17. https://intuitionlabs.ai/articles/ai-governance-in-biotech-charter-decision-rights
18. https://pages.nist.gov/800-63-4/sp800-63a.html
19. https://ibbis.bio/our-work/customer-screening/
20. https://www.gov.uk/government/publications/uk-screening-guidance-on-synthetic-nucleic-acids/uk-screening-guidance-on-synthetic-nucleic-acids-for-users-and-providers
21. https://www.ncsc.gov.uk/collection/zero-trust/architecture-design-principles/authenticate-and-authorise
22. https://www.ncsc.gov.uk/collection/zero-trust/architecture-design-principles/assess-user-behaviour-service-and-device-health
23. https://www.cisa.gov/resources-tools/resources/zero-trust-maturity-model
24. https://doi.org/10.6028/NIST.AI.600-1
25. https://osp.od.nih.gov/policies/biosafety-and-biosecurity-policy/faqs-on-institutional-biosafety-committee-ibc-administration-april-2024/
26. https://osp.od.nih.gov/policies/biosafety-and-biosecurity-policy/faqs-about-ibc-meetings-and-minutes/

27. https://www.canada.ca/en/public-health/services/canadian-biosafety-standards-guidelines/guidance/biosafety-program-management/guideline.html
28. https://www.aspr.gov/readiness-response/medical-countermeasures-biodefense/s3/high-consequence-research-oversight/USG-Policy-For-Stopping-High-Risk-Life-Sciences-Research
29. https://www.grants.nih.gov/grants/guide/notice-files/NOT-OD-26-101.html
30. https://www.aisi.gov.uk/blog/early-lessons-from-evaluating-frontier-ai-systems
31. https://intuitionlabs.ai/articles/responsible-enterprise-ai-privacy-governance
32. https://privacy.claude.com/en/articles/7996866-how-long-do-you-store-my-organization-s-data
33. https://privacy.claude.com/en/articles/8956058-i-have-a-zero-data-retention-agreement-with-anthropic-what-products-does-it-apply-to
34. https://www.enisa.europa.eu/sites/default/files/publications/ENISA%20Report%20-%20Cloud%20Security%20for%20Healthcare%20Services.pdf
35. https://www.fda.gov/inspections-compliance-enforcement-and-criminal-investigations/fda-bioresearch-monitoring-information/guidance-industry-computerized-systems-used-clinical-trials
36. https://www.who.int/publications/i/item/9789240056107
37. https://airc.nist.gov/airmf-resources/airmf/5-sec-core/
38. https://nvlpubs.nist.gov/nistpubs/legacy/sp/nistspecialpublication800-55.pdf
39. https://www.enisa.europa.eu/news/enisa-news/enisa-publishes-training-course-material-on-network-forensics-for-cybersecurity-specialists
40. https://www.rand.org/pubs/research_reports/RRA5055-1.html
41. https://www.cyber.gov.au/business-government/exercise-in-a-box
42. https://www.enisa.europa.eu/sites/default/files/2025-05/2025.04311_01_ms_v2.0_Handbook%20for%20Cyber%20Stress%20Tests_en.pdf
43. https://preptoolkit.fema.gov/web/hseep-resources/improvement-planning
44. https://www.canada.ca/en/public-health/services/canadian-biosafety-standards-guidelines/third-edition.html
45. https://www.iso.org/standard/42001
46. https://www.oecd.org/en/topics/ai-principles.html

THE PUBLISHER

About IntuitionLabs

IntuitionLabs is an AI consulting, custom software development and data engineering firm serving pharmaceutical, biotechnology, medical-device and other life-science organizations. We work with clinical, regulatory, medical-affairs, commercial, quality and IT teams to connect technology decisions with the work people need to accomplish.

AI consulting and adoption

Our AI enablement services cover readiness assessments, use-case selection, governance and policies, team workshops, adoption measurement and ongoing advisory support. We help organizations structure the information layer behind AI: source material, context, permissions and maintained knowledge that make generated answers useful and reviewable. Private LLM inference and hosted AI options support teams evaluating how to operate AI with appropriate control over their data and infrastructure.

Software, data and life-science workflows

IntuitionLabs develops custom software for pharma and biotech, integrates enterprise systems, and builds data engineering and business intelligence solutions. Areas of focus include AI agents, regulatory research, medical writing, medical affairs, CMC information, competitive intelligence and clinical-document workflows. Our eTMF intelligence work includes cross-system reconciliation and inspection-readiness support.

Enterprise platforms and regulated delivery

We provide Veeva services, application support, managed services, integrations and custom applications, alongside enterprise content work involving platforms such as Egnyte. For regulated workflows, our services include GxP enablement, computer-system validation and software development addressing 21 CFR Part 11 requirements. The applicable controls, validation responsibilities and acceptance criteria are defined for each engagement.

Work with IntuitionLabs

Explore AI enablement, pharma and biotech software development, data engineering and BI, and Veeva services. Contact IntuitionLabs to discuss your workflow, information sources and implementation needs.

IntuitionLabs publishes educational research to help life-science teams make informed technology decisions. Coverage of a product or organization does not imply a client relationship, endorsement or partnership.

Website: https://intuitionlabs.ai