Governing AI in Pharmaceutical Operations

By Brian Drapeau

How to decide what AI may do and whether people can use it reliably

What must be true before the work proceeds

Consider a deviation summary prepared by AI. The chronology looks sound, the documents are linked, and a reviewer approves it. Months later, someone finds an equipment record that contradicts the conclusion. The summary left it out. To understand this illustrative failure, we need to know which records the system could access, how it selected them, what the reviewer could see, and why the summary was approved.

AI should not support pharmaceutical quality work until the organization has defined the task, established what the system may do, and tested whether the complete human and AI workflow can perform representative work reliably. The resulting evidence should determine whether the work may proceed, what limits apply, and what findings would require correction, restriction, or withdrawal.

This approach connects four questions that organizations often address separately: whether the evidence is adequate, whether the system has appropriate authority, whether people can perform their assigned role, and whether the organization can find and correct failures. Passing a model test or completing a course does not establish that the combined workflow will produce acceptable work. A qualified person can still miss a record that changes the conclusion, especially when the system controls which evidence reaches the reviewer.

I developed this operating method through earlier work on manufacturing readiness, governance by design, AI qualification, and agent authority. The method treats the relationship among the person, system, version, use case, source environment, and granted authority as the operating unit.

The focus here is pharmaceutical manufacturing and quality operations. Other GxP settings require controls suited to their records, decisions, and governing requirements.

What industry experience tells us

Industry experience supports the need to evaluate the work around the technology. Roche describes trusted data, employee development, process redesign, and governance as connected conditions for useful AI outcomes. McKinsey's pharmaceutical analyses similarly place talent and change management alongside strategy, technology, analytics, and data, and call for operators and technical staff who can recognize errors and make informed decisions. [1-3]

Lilly's responsible-use disclosure assigns employees accountability for AI outputs and describes evaluation for specific uses. GSK reports generative AI across manufacturing sites for examining historical investigations, alongside further inspection-readiness work. These accounts show organizations confronting evidence access, human judgment, and workflow design together. They do not establish that any one governance method has solved the problem. [4-5]

The useful question is whether an organization can use available evidence to make a defensible decision about a particular AI-supported job.

The AI-Supported Work Authorization Cycle

The following cycle turns the argument into a set of gate decisions. Each stage produces evidence for proceeding, narrowing the use, correcting it, or stopping it.

Stage	Question	Required evidence	Accountable owner	Gate decision
0. Inventory and ownership	What AI is in use, including unapproved use, and who owns each use?	Use inventory, systems and data accessed, users, suppliers, named process, system, and quality owners	AI governance lead and quality unit	Register, contain, prohibit, or assign ownership
1. Define the work and risk	What task and decision will AI support, and what could failure affect?	Intended use, source boundary, applicable requirements, baseline process, risk assessment	Process owner and quality owner	Approve the scope, narrow it, or reject it
2. Set authority and records	What may the system read, recommend, change, or approve, and which records must be retained?	Permissions, prohibited actions, approval design, record and signature requirements, retention plan	System owner, quality owner, IT, and data owner	Authorize bounded functions or prohibit them
3. Test the complete workflow	Can the system and people perform representative work under realistic conditions?	Challenge cases, system results, reviewer performance, omissions, false alarms, time, and recovery tests	Evaluation owner and quality unit	Authorize a pilot, correct, narrow, or stop
4. Authorize routine use and monitor	Does pilot evidence justify routine use, and does performance remain acceptable?	Pilot results, unresolved risks, incidents, near misses, overrides, workload, outcome measures, drift indicators, and user feedback	Process owner and quality unit	Authorize routine use or expansion, continue, restrict, investigate, or suspend
5. Change, correct, and retire	What changed, what was affected, and did the response work?	Change impact, regression results, requalification, corrective-action evidence, records, dependencies, and continuity plan	Change owner, system owner, and quality unit	Reinstate, narrow, stop, or retire

Stage 0: Inventory and ownership

Stage 0 matters because an organization cannot govern a use it has not identified. The inventory should describe actual work, not merely list vendors. "AI used in quality" is too broad. The entry should identify the task, source systems, output, users, downstream decision, and whether the system can act. Unapproved consumer-tool use belongs in the same inventory process because it may expose regulated information or create outputs that enter the quality system without an accountable owner.

Ownership should also be divided clearly. The process owner defines the work and acceptable outcome. The system owner maintains the configured service, integrations, access, and technical evidence. The data owner governs the sources and permitted uses of information. The quality unit determines whether the proposed controls and evidence support the regulated use and retains its existing approval responsibilities. An AI governance body can set common policy, coordinate expertise, and resolve cross-functional questions, but it should not become a substitute owner for every local process. Suppliers provide evidence and change information; they do not accept the manufacturer's regulated decision on its behalf.

The inventory should remain connected to procurement, access management, incident reporting, and retirement. Otherwise, it becomes a periodic spreadsheet that misses the work it was created to govern. New integrations, expanded permissions, unexpected outputs in controlled records, and reports of consumer-tool use should trigger review of the inventory rather than wait for an annual reconciliation.

The cycle does not require the same ceremony for every use. A bounded formatting tool and an agent that can alter a regulated record present different questions. Evidence, documentation, and review should follow the significance, uncertainty, and complexity of the work.

Each gate needs an actual decision maker and a recorded outcome. Authorization for a pilot does not authorize routine use, and routine use does not authorize expansion to a new task, source, site, or authority level. A collection of completed activities is not itself authorization. The organization should be able to show who reviewed the evidence, which conditions were accepted, what remained unresolved, and which decision followed. The same record should identify the trigger for reconsidering that decision. Without that connection, validation, training, risk assessment, and monitoring can all exist while nobody can explain why the workflow was allowed to operate.

Establish accountability and the applicable control basis

AI does not displace existing quality responsibilities. FDA's April 2026 warning letter to Purolea Cosmetics Lab described AI use to prepare specifications, procedures, and master production or control records within a wider pattern of manufacturing deficiencies. FDA cited inadequate quality-unit review under 21 CFR 211.22(c) and directed the firm, if it resumed relevant activities, to have AI outputs or recommendations reviewed and cleared by an authorized quality-unit representative. [6] This was a specific enforcement finding involving products regulated as drugs, not a universal AI-certification requirement. My earlier article and a separate Pharmaceutical Technology analysis interpret the letter as a reminder that accountable review cannot be outsourced to the system. [7-8]

Personnel requirements provide a related foundation. Under 21 CFR 211.25, education, training, experience, or their combination must enable assigned functions, and relevant training must continue with sufficient frequency. The provision does not prescribe the assessment method proposed here. It does require the organization to have enough qualified people for the work it assigns. [9]

ICH Q9(R1) supplies the quality-risk basis for proportionality. It calls for effort, formality, and documentation commensurate with risk, uncertainty, importance, and complexity. It also treats risk management as a lifecycle process of assessment, control, communication, and review. This supports stronger evidence for consequential uses and simpler controls for bounded ones. It does not permit risk language to excuse an otherwise unacceptable practice. [10]

Data integrity is part of the operating design. The commonly used ALCOA+ shorthand is useful only when translated into the records and controls needed for the job. The organization should be able to identify who or what created or changed a record, when it happened, which source and version were used, and whether the record remained complete, consistent, accurate, available, and protected through its lifecycle. FDA's drug-CGMP data-integrity guidance emphasizes reliable and accurate data and risk-based strategies grounded in process understanding and knowledge of the technology and business model. [11]

European requirements and proposals must also be distinguished. The effective EudraLex Volume 4 listing and the consultation materials for revised Chapter 4, Annex 11, and new Annex 22 are not the same thing. For critical GMP applications, draft Annex 22 addresses static machine-learning models with deterministic outputs and says dynamic models, probabilistic-output models, generative AI, and large language models are outside its scope and should not be used. For non-critical uses of generative AI and large language models, it calls for qualified personnel to remain responsible for output suitability. The text remains a draft unless and until adopted. Adding a human reviewer does not make a critical function non-critical; criticality follows the function and consequences. [12-14]

The joint FDA and EMA principles for good AI practice in drug development support examining the complete system, including human and AI interaction, data provenance, multidisciplinary expertise, and continuing performance. They are principles, not a statutory qualification scheme or an endorsement of this method. [15]

FDA's February 2026 Computer Software Assurance guidance applies to software used in medical-device production and quality-management systems under Part 820; its risk-based intended-use logic is only an analogy here. ISPE's July 2025 GAMP Guide: Artificial Intelligence, read alongside GAMP 5 Second Edition, offers lifecycle practitioner guidance for GxP AI rather than product certification. [16-18]

Stage 1: Define the work and risk

I use the shorthand person × system × version × use case. For an agent, add the source environment and authority envelope: the systems and records it may access, the actions it may take, the approvals it needs, and the conditions that stop it. Readiness depends on this complete relationship and can change when any part changes. [19]

"Prepare a draft chronology from a defined set of approved deviation records for an investigator who verifies each consequential statement" can be evaluated. "AI for Quality" cannot. A useful intended use identifies the task, source boundary, expected output, user, downstream decision, and prohibited actions. It also identifies how failure could affect product quality, patient safety, data integrity, compliance, or availability.

The system description should include the interface, connected tools, retrieval sources, configurations, permissions, and points where a person must intervene. A capable reviewer may face a materially different job after a retrieval change, model update, new product, altered interface, or grant of write access. The name of the product may remain unchanged while the qualified state has changed.

Define the baseline process at the same time. If the current human process already misses records, produces inconsistent classifications, or takes an unacceptable amount of time, the comparison should expose that weakness. The objective is not to preserve a flawed baseline. It is to determine whether the proposed workflow produces acceptable work and whether its benefit justifies the new risks and controls. A comparison that measures only the AI condition can misattribute pre-existing failures to the tool or conceal failures that both methods share.

Different technologies require different evidence. A classifier used in visual inspection, a predictive model used in process analysis, and a language model drafting an investigation create different failure modes. For a classifier, evaluation should address independent test data, important defect classes, subgroup performance, calibration, and the consequences of false positives and false negatives. A strong average can conceal poor performance on a rare but consequential defect.

For generative systems, verification must address meaning. A fluent sentence can reverse a negation, remove a qualification, or imply causation where the source shows only sequence. Agreement between repeated answers gives limited reassurance when the same missing source or unsupported inference drives each answer. Tested calculation tools should perform controlled calculations, deterministic rules should enforce permissions, retrieval should identify candidate evidence, and generative AI may organize or draft material within those boundaries. Test each component for its assigned job and then test the complete workflow.

The product and task also shape the challenge cases. In QC, a reviewer may need to find omitted raw data, a dilution error, a method-version mismatch, or an unsupported explanation for an out-of-specification result. In technology transfer, the risk may be extrapolation beyond studied conditions. In patient-specific manufacturing, identity and custody require explicit testing because a plausible summary attached to the wrong patient or material is a serious failure. For external partners, the system must preserve site identity, supplier status, quality-agreement responsibility, and provenance across organizational boundaries.

Stage 2: Set authority and preserve decision evidence

As AI moves from producing content to taking action, the organization needs a precise account of what each system may observe, interpret, recommend, decide, execute, and approve. Capability describes what the technology can do. Permission describes what the organization has authorized it to do. The authority envelope should identify accessible systems and records, allowed operations, prohibited actions, approval points, stop conditions, and the person accountable for the result. [20]

Part 11 should be applied with equal precision. It does not make every AI prompt or output a Part 11 record. Under FDA's scope guidance, Part 11 applies when records required by predicate rules are maintained electronically in place of paper or relied upon for regulated activities, when designated records are submitted electronically to FDA, and when electronic signatures are intended to serve as required signatures. The organization should determine which records and signatures fall within that scope. For applicable systems, authority checks, operational checks, access controls, accountability for electronic signatures, and reliable records are directly relevant to AI-supported approvals and actions. [21]

Make that determination here, before a system has generated months of records. The team should identify which artifact is the regulated record, which supporting evidence must remain linked to it, whether an electronic approval is intended to have the force of a required signature, and what information must be available during an investigation or inspection. An AI interaction may be transient in one workflow and consequential evidence in another. The applicable predicate rule and actual business reliance should decide the treatment.

These limits require technical enforcement. An instruction telling an agent not to approve a record provides weak protection if its credentials permit approval. The agent should have its own attributable identity and only the access required for its assigned function. It should not inherit all the privileges of the employee who launches it.

Approval must remain tied to the state reviewed. If someone approves an update and the underlying record changes before the agent acts, the system should determine whether the approval still applies. Where the change undermines the original basis, the action requires another review. Permission to retrieve records does not confer permission to determine root cause, open a corrective action, or approve closure.

The record should let another person reconstruct what happened: the task, relevant inputs, source versions, retrieval results, configuration, tool actions, output, changes, reviewer response, approval, and resulting system state. A model's narrative explanation of its answer is not proof of its internal path. Keep that narrative distinct from verifiable inputs, outputs, actions, and human reasons.

This is decision provenance rather than logging alone. The audit trail may show that a field changed at a particular time. Decision provenance connects that event to the evidence selected, the intermediate recommendation, the approval state, and the authority under which the change occurred. The organization does not need to preserve every internal computation a supplier cannot expose. It does need enough reliable evidence to explain its own regulated action and investigate a failure.

Segregation of duties also requires attention. If one agent retrieves the evidence, selects what is relevant, interprets it, drafts the conclusion, and evaluates its own result, several responsibilities that were separate in the human process may have collapsed into one technical pathway. A second agent is not independent merely because it has a different label. Independence depends on the data, methods, failure modes, and authority used for the check.

A complete audit trail can record a bad decision perfectly. The organization needs both a sound decision and reliable evidence of how it was made. Retention should follow the task, applicable predicate rules, risk,

investigation needs, privacy, confidentiality, and contractual obligations. Indiscriminate logging can create uncontrolled copies without improving assurance. Exact reproducibility may be impossible for a probabilistic service, but the organization should still preserve enough historical evidence to explain the regulated decision. If exact reproduction is required, that requirement should shape the system choice before approval.

Stage 3: Test the complete workflow

Stage 3 determines whether the configured system and assigned people can perform representative work under realistic conditions, first well enough for a bounded pilot and then well enough to support a separate routine-use decision.

Test the evidence and the review conditions

A citation is a starting point for verification. The source must exist, support the statement, and apply to the case. A true statement about another product, site, jurisdiction, process stage, or document revision can produce a wrong conclusion in the present workflow.

A summary can accurately describe five retrieved deviations and still omit a sixth that changes the conclusion. Citations to the five will not reveal the missing record. The team needs a justified method for identifying which records belong in the review and detecting consequential exclusions. Two agents can agree because both missed the same source. Calling one a writer and the other a reviewer does not make their checks independent.

Make clear whether an output reports a fact, presents a calculation, offers an explanation, or proposes an action. Compare a factual statement with its source, repeat a calculation, and test an explanation against alternatives and contradictory evidence. Preserve the distinction between chronology and causation. If an excursion follows an intervention, the sequence may justify investigating a connection; it does not establish that the intervention caused the excursion.

The phrase "human in the loop" identifies a person's position, not the effectiveness of the review. Review depends on competence, evidence access, available time, interface design, and practical authority to reject or escalate. Showing a persuasive AI conclusion before the evidence may bias the reviewer's search. For consequential recommendations, the workflow can require an initial evidence assessment before displaying the AI conclusion, or another method that reduces anchoring. Lower-risk transformations may need only a direct source comparison.

Workload is part of the control. Increasing output while reducing the time available for each review can weaken the approval on which the process depends. A reviewer who may reject an output but is penalized for delay receives conflicting instructions. The evaluation and later monitoring should examine review time, queue size, escalation patterns, and whether staff can report uncertainty and near misses without pressure to approve.

The evaluation should preserve evidence identity. Record the test collection, document versions, inclusion rules, expected findings, tool configuration, reviewer instructions, scoring rubric, and any changes made after results were examined. Reusing the same visible cases for repeated tuning can turn a test into practice. Fresh cases or a protected holdout set are needed when the result will support authorization.

The comparison should count more than average accuracy. Identify critical misses that cannot be offset by easier successes, false connections that could misdirect an investigation, correct findings rejected without basis, and unnecessary escalations that add burden. Measure the time required to reach an acceptable evidence packet, not merely the time to generate the first draft. Record withdrawals and incomplete work

rather than silently excluding them. If the sample is small, present the uncertainty instead of converting a few cases into a precise-looking percentage.

Before evaluation, define the performance needed for the bounded use and the improvement that would justify the added intervention. A qualification method that raises examination scores without improving independently assessed work has not demonstrated its intended value. If an interface change produces the same benefit with less burden, that alternative deserves consideration. The added controls must earn their place through evidence.

Empirical research supports testing the combined arrangement. Vaccaro and colleagues' meta-analysis covered 106 experiments and 370 effect sizes. Human and AI combinations performed better than humans alone on average, but worse than whichever component performed best alone. The studies span varied settings and largely predate current generative systems, so they do not establish GMP performance. They do show why adding a reviewer cannot be assumed to improve an outcome. [22]

Qualify people for the role they actually perform

Awareness that AI can make mistakes is a beginning, not evidence of readiness. The test is whether someone can identify a consequential defect in representative work, explain why it matters, and accept, reject, correct, or escalate the result appropriately.

Challenge cases should include correct outputs, awkwardly written but sound outputs, fluent dangerous outputs, incomplete evidence, wrong-product sources, and conditions where rejecting everything is itself an error. Count both missed defects and unsupported accusations of error. Decide in advance which serious failures cannot be offset by success on easier items. Assessors should agree on what acceptable work looks like and why; unresolved scoring disagreement weakens the permission decision.

Dell'Acqua and colleagues' randomized study of 758 knowledge workers illustrates the importance of task boundaries. AI improved speed and completion within the tested capability range. On a deliberately selected task outside that range, the AI groups' correctness was about 19 percentage points lower than the control group's. The study concerns consulting work with GPT-4, not pharmaceutical qualification. Its methodological lesson is still useful: evaluation should include conditions where assistance can mislead as well as conditions where it helps. [23]

Qualification should match responsibility. A person using a bounded tool to reformat approved text does not need the same demonstration as a quality approver evaluating an investigation or a system owner granting an agent write access. Training hours and course completion are inputs, not evidence that someone can perform a particular employer's work. This follows the readiness argument I developed for advanced-therapy manufacturing. [24]

The same evidentiary standard applies to my Life Sciences AI Academy blueprint. Its exercises and qualification pathways are proposed designs; they still need to demonstrate improvement in independently assessed workplace performance. [25]

A qualification program can also fail by blaming the operator for impossible conditions. If the interface hides source status, the evidence collection is incomplete, or volume makes careful review unrealistic, repeated retraining will not repair the workflow. Training cannot recover evidence the system never makes available.

A passing standard needs a reason. Define which outcomes matter, how cases represent the assigned work, and how assessors will distinguish an acceptable decision from a plausible but unsupported one. Use enough varied cases to avoid granting authorization on the strength of a rehearsed example. Where judgment is involved, calibrate assessors against shared examples and investigate material disagreement. If competent

assessors cannot apply the standard consistently, the organization does not yet have a dependable basis for using the result to grant permission.

Qualification should also decay when its supporting conditions change. A reviewer qualified on one evidence repository, interface, language, and authority level should not automatically remain qualified after the system begins searching additional repositories or initiating workflows. Requalification can be targeted to the changed function. It need not repeat every prior activity, but the retained evidence should have an explicit rationale.

Challenge agents before routine use

An agent can turn one mistaken interpretation into changes across several systems. Problems can also arise when each tool works but the sequence fails. Consider an agent authorized to open a follow-up task and update a linked record. The first operation succeeds, the second times out, and the agent retries the entire sequence. It creates a duplicate task. The request and generated text were appropriate; the agent lost track of what had already happened.

The system should recognize repeated requests, check the state of prior operations, and expose incomplete work. Verify the actual records rather than treating the agent's completion message as proof. Test duplicate requests, expired approvals, unavailable sources, changed records, interrupted execution, concurrent edits, and partial success.

Recovery should distinguish reversible and irreversible actions. A record change may be reversed; an external notification cannot be unsent. Restoring one database may leave another system inconsistent. The recovery plan should specify reconciliation, compensating action, communication, and the authority to suspend the workflow.

Operating hours also matter. If an agent can take consequential action outside staffed periods, the organization needs a response arrangement for those periods. Otherwise, operation may need to be limited to times when responsible people can intervene. Training users to watch carefully cannot replace least privilege, state checks, observable execution, and tested recovery.

Stage 4: Authorize routine use and monitor

A successful evaluation establishes evidence for the conditions tested. It does not by itself authorize routine use. The process and quality owners should review pilot performance, critical failures, unresolved risks, review burden, recovery results, and operating controls before granting routine authorization. Expansion to another task, site, source collection, user group, or authority level requires another explicit decision.

Continued use requires determining whether the authorized conditions still apply. Model updates, retrieval changes, revised procedures, new products, staffing changes, and added permissions can alter performance or risk. McKinsey's learning-loop model in biopharma R&D offers a useful parallel: later evidence can inform earlier decisions while authority for protocol and regulatory changes remains human. It is a strategic model, not a GMP control standard. [26]

Monitoring should focus on consequential performance. Average quality can remain stable while a rare defect family worsens. Examine relevant products, sites, roles, document types, languages, and failure classes where the data support those distinctions. Keep small samples and uncertainty visible.

Overrides and near misses can provide useful signals, but their meaning requires context. A sudden absence of overrides may reflect improvement, rubber-stamping, poor capture, or migration to an unapproved tool. Stage 0 inventory and ordinary user feedback help identify shadow use that formal metrics miss.

Stage 5: Change, correct, reauthorize, and retire

Assess what a change could affect before selecting the tests. A wording edit that leaves behavior unchanged may need little additional assurance. A change in source selection may require targeted regression cases. Granting write access creates a different authority envelope and a different qualified state. The organization should document which prior evidence still applies and which conditions require new testing or requalification.

Approved changes must also reach the right people and systems. A revised procedure may affect particular sites, roles, work instructions, assessments, and permissions. Assigning the document to everyone is easy to count but does not show that the correct intervention occurred. Some changes require awareness; others require practice, observed demonstration, requalification, or a change in authorization. A governed applicability process should identify affected relationships from approved sources, show the basis for each proposed connection, and leave uncertain cases for responsible review. [27]

When a failure occurs, determine its scope, preserve the evidence, restrict affected use when warranted, and investigate human and system causes. Retraining is appropriate when a person lacked a required skill. It will not repair missing evidence, an interface that hides limitations, a defective permission design, or unreliable equipment. Corrective action should face fresh cases that reproduce the failure mechanism, along with unaffected cases that check for new errors.

Correctability means being able to find a problem, determine what caused it and what it affected, correct it under change control, and test whether the correction worked. Completing retraining or changing a prompt establishes that an action occurred. It does not establish that the action was effective. The final decision may be to reinstate the workflow, narrow it, stop it, or retire it. Retirement should address remaining records, access, supplier dependencies, and continuity obligations.

Count the time and cost of the whole job

AI saves little if people cannot check its work well enough to use it. Measure the time and effort required to finish an acceptable piece of work, including review, correction, escalation, maintenance, and downstream rework. Drafting speed alone does not show that the process improved. [28]

Sanofi reports reductions of up to 80% in time and effort for generative-AI-assisted manufacturing reports, beginning with annual Product Quality Reports (PQR) and extending to other manufacturing documentation. The public account does not provide enough detail to calculate the cost of an approved report or the burden of all controls. It nevertheless provides a relevant operating example of realized drafting benefit. [29]

A published case study of AstraZeneca's AI-governance implementation describes approximately four full-time staff coordinating the effort during 2020 and 2021 and a 14-week audit requiring about 2,000 person-hours, without technical testing of individual models. These estimates belong to that program and do not measure reduced AI error. They show why governance labor must be included in the business case. [30]

Evidence outside pharma also shows why benefits should be measured by task and user. Brynjolfsson, Li, and Raymond's published study of 5,172 customer-support agents found a 15% average increase in issues resolved per hour, with larger gains among less-experienced workers. It was a staggered deployment in one firm, not a GMP trial. It cautions against assuming that the most experienced workers always benefit most or that AI necessarily prevents learning. [31]

Before a pilot starts, define what should improve and what cannot worsen. An SOP gap-check tool might be expected to reduce search time without increasing consequential omissions or false alarms. Measure the

current process first. Compare time to acceptable work, important errors, unnecessary escalations, review burden, and maintenance cost over an agreed period.

Supplier agreements should identify the intended use, evidence each party will provide, acceptance criteria, change-notification duties, records available for investigation, and responsibility for evaluating changes. Delivery of software is not acceptance of the configured workflow. NIST's voluntary Generative AI Profile supports explicit supplier expectations for quality, security, provenance, third-party evaluation, and incident responsibilities, but it does not impose a pharmaceutical contracting rule. [32]

Pharmaceutical companies now have enough operating experience to identify recurring failure modes and propose controls. The public evidence base is not yet mature enough to establish the long-term effectiveness and net cost of a complete governance model. This approach is therefore designed as a testable operating method: it helps organizations decide what AI-supported work may proceed while generating evidence about whether the controls improve the work and justify their cost.

Putting the cycle to work with the Lilly examples

Pharmaceutical Technology reports that Lilly connects inspection, audit, and deviation information through a Document Explorer for unstructured records and a Data Explorer for structured data. The account describes views by month, site, and authority, linked to responses and after-action reviews, along with translation, search-processing, and corpus-size tradeoffs. The following application uses those reported capabilities as a setting. The controls, scenarios, and acceptance decisions are my proposals, not descriptions of Lilly's internal implementation or results. [33]

Consider this question: has a particular quality concern appeared elsewhere, and what evidence should a quality team examine before deciding whether wider action is warranted? The permitted output would be a cited evidence packet with unresolved questions. The system would not declare a site compliant, predict an inspection outcome, or approve corrective action. An authorized quality owner would decide the significance of the findings and any response.

Begin with a bounded evaluation collection reconciled against an inventory. Each consequential statement should link to its original passage, document status, site, date, and translation history where applicable. Include superseded responses and relevant records expressed in unfamiliar terminology. Outside a bounded collection, search limits should remain explicit because failure to retrieve a precedent does not establish that none exists.

Suppose the packet omits a decisive earlier finding recorded in unfamiliar terminology. The returned evidence looks consistent, and the reviewer concludes that the concern is isolated. Reconciliation against the test inventory reveals the omission. The evaluation owner preserves the query, retrieved records, configuration, and review decision, then restricts the affected use while the failure is investigated. This is an illustrative challenge, not a reported Lilly incident.

The investigation should distinguish at least three possibilities. If the record was never indexed or the search failed to match its terminology, correct ingestion or retrieval. If the interface concealed search limits or unavailable records, redesign the workflow. If the evidence and limitations were accessible but the reviewer misunderstood their significance, targeted preparation may be warranted. More than one cause may contribute.

Test the correction on fresh cases with similar retrieval problems and on unaffected cases that could reveal new errors. Measure consequential omissions, unsupported conclusions, correct packets wrongly rejected, unnecessary escalation, and review burden. Resolving the original case alone would not justify restoring use.

The quality owner should decide whether the new evidence supports reinstatement, narrower scope, continued restriction, or retirement.

For structured-data exploration, separate verifiable counts from interpretations. A challenge case can present an apparent rise in findings caused by a changed category definition, duplicate record, or different observation period. The calculation should be reproducible from the underlying dataset with filters and exclusions retained. The reviewer should distinguish a change in recorded data from evidence of worsening performance. When the data cannot support comparison across sites, the system should expose that limitation.

For connections across quality systems, test whether records describe the same issue, related issues, or only similar language. Pair an audit observation with a deviation from a different product and a superficially reassuring response that does not address the original concern. The reviewer should reject the unsupported connection, identify what remains unresolved, and avoid treating a completed response as proof of effectiveness.

Where feasible, compare three conditions: the current search-and-review process, AI assistance with existing preparation, and AI assistance after task-specific preparation. Keep the tool, settings, and source collection the same in the two AI groups so better software is not mistaken for better training. Use unfamiliar cases, distribute difficult cases fairly, and score the work without revealing the condition. Report baseline performance and uncertainty alongside any improvement.

How we would know the approach is not helping

The evaluation must permit the conclusion that an added check, workflow step, or training requirement does not help. A simpler interface, deterministic tool, narrower AI task, or existing process may achieve better results. An assessment that mainly produces documentation or teaches people to pass its own cases has not demonstrated workplace value.

The method is not helping if qualified reviewers cannot reproduce its decisions, if important omissions or false accusations do not improve, if the burden exceeds the task's benefit, if users move to unapproved tools, or if monitoring cannot detect deterioration soon enough to act. Those findings should lead to redesign, restriction, or withdrawal rather than another layer of paperwork.

Decide whether the work can proceed

Before allowing AI to support a consequential quality decision, name the task and accountable owner, examine representative results, and agree on what finding would make the team stop or restrict use. Set a date or trigger for reviewing the evidence again.

What evidence would justify authorizing this workflow? What finding would make us restrict or stop it? Answer both before the work proceeds.

Sources

[1] Martha Heller. "How Roche drives AI outcomes in a changing world." CIO, September 16, 2026. Interview with Wafaa Mamilli. Read source

[2] Alex Devereson and Navraj Nagra. "How pharma is rewriting the AI playbook: Perspectives from industry leaders." McKinsey, November 6, 2025. Read source

[3] Boyd Spencer, Parag Patel, and Vivek Arora, with Krithiknath Tirupapuliyur and Raj Rajendran. "Gen AI: A game changer for biopharma operations." McKinsey, January 28, 2025. Read source

[4] Eli Lilly and Company. "Responsible Use of AI." Corporate sustainability disclosure, updated July 1, 2026. Read source

[5] GSK. Annual Report 2025. Manufacturing and supply section. Read source

[6] US Food and Drug Administration. Purolea Cosmetics Lab, Warning Letter 320-26-58, MARCS-CMS 722591. April 2, 2026. Read source

[7] Brian Drapeau. "Is Compliant AI in Pharma Even Possible? The Search for an Honest Answer." Pharmaceutical Technology, June 16, 2026. Read source

[8] Gourav Pandey and Svyatoslav Borshchenko. "What FDA's AI Warning Letter Tells Us About GMP Accountability." Pharmaceutical Technology, July 6, 2026. Read source

[9] 21 CFR 211.25, Personnel qualifications. eCFR text accessed September 18, 2026. Read source

[10] International Council for Harmonisation. ICH Q9(R1), Quality Risk Management. Step 4 version, January 2025. Read source

[11] US Food and Drug Administration. Data Integrity and Compliance With Drug CGMP: Questions and Answers. Guidance for Industry, December 2018. Read source

[12] European Commission. EudraLex Volume 4. Effective-document listing checked September 18, 2026. Read source

[13] European Commission. Stakeholders' Consultation on EudraLex Volume 4: Chapter 4, Annex 11 and New Annex 22. Closed consultation, July 7 to October 7, 2025; status checked September 18, 2026. Read source

[14] European Commission. Draft Annex 22, Artificial Intelligence. July 2025 consultation draft. Read source

[15] US Food and Drug Administration and European Medicines Agency. Guiding Principles of Good AI Practice in Drug Development. January 2026. Read source

[16] US Food and Drug Administration. Computer Software Assurance for Production and Quality Management System Software. Final Guidance, revised February 2026. Applies to medical-device production and quality-management systems under 21 CFR Part 820. Read source

[17] International Society for Pharmaceutical Engineering. ISPE GAMP Guide: Artificial Intelligence. July 2025. Read source

[18] International Society for Pharmaceutical Engineering. GAMP 5: A Risk-Based Approach to Compliant GxP Computerized Systems. Second Edition, July 2022. Read source

[19] Brian Drapeau. "Europe Tried to Ban Generative AI From Critical GMP. The Ban May Not Survive. It Does Not Matter." Pharmaceutical Technology, August 13, 2026. Read source

[20] Brian Drapeau. "Pharma's Next AI Problem Is Authority." Pharmaceutical Technology, August 27, 2026. Read source

[21] US Food and Drug Administration. Part 11, Electronic Records; Electronic Signatures: Scope and Application. Guidance for Industry, August 2003. Read source

[22] Michelle Vaccaro, Abdullah Almaatouq, and Thomas Malone. "When combinations of humans and AI are useful: A systematic review and meta-analysis." Nature Human Behaviour 8, 2293-2303 (2024). Read source

[23] Fabrizio Dell'Acqua and colleagues. "Navigating the Jagged Technological Frontier: Field Experimental Evidence of the Effects of Artificial Intelligence on Knowledge Worker Productivity and Quality." Organization Science, published online March 11, 2026. Read source

[24] Brian Drapeau. "Training: The Real Constraint in CGT Manufacturing." Pharmaceutical Technology, May 27, 2026. Read source

[25] Brian Drapeau / Life Sciences AI Academy. Compliant AI in GMP: Full Technical Blueprint for Risk-Proportionate Learning, System Controls, and Performance Assessment. Version 0.5, September 17, 2026. Available after registration from the Life Sciences AI Academy curriculum page. Read source

[26] Alex Devereson, David Champagne, Erika Stanzl, and Lieven Van der Veken, with Alex Peluffo, Chris Anagnostopoulos, and Jennifer Hou. "From linear gates to learning loops: Rewiring biopharma R&D with AI." McKinsey, June 29, 2026. Read source

[27] Brian Drapeau. "Can Governance-by-Design Make AI Compliant in Pharma? (Part II)." Pharmaceutical Technology, July 16, 2026. Read source

[28] Brian Drapeau. "Pharma's AI ROI Problem Isn't a Tech Problem but a Qualification Problem." Pharmaceutical Technology, August 10, 2026. Read source

[29] Sanofi. "Digital & Artificial Intelligence." Manufacturing and supply section; accessed September 18, 2026. Read source

[30] Jakob Mökander and Luciano Floridi. "Operationalising AI governance through ethics-based auditing: an industry case study." AI and Ethics 3, 451-468. Published online May 31, 2022; 2023 issue. Read source

[31] Erik Brynjolfsson, Danielle Li, and Lindsey Raymond. "Generative AI at Work." Quarterly Journal of Economics 140(2), 889-942 (2025). Read source

[32] Chloe Autio and colleagues. Artificial Intelligence Risk Management Framework: Generative Artificial Intelligence Profile. NIST AI 600-1, July 2024. Read source

[33] Zachary Zubulake. "Lessons For AI Governance in Quality Systems." Pharmaceutical Technology, September 15, 2026. Conference reporting on Karthik Iyer's presentation. Read source