Governing AI in Pharmaceutical Operations

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

By Brian Drapeau

Version 2.2 | September 19, 2026

Brian Drapeau is a consultant with specific expertise in pharmaceutical quality operations, AI governance, and workforce qualification. He founded GxP Frame and the Life Sciences AI Academy.

Executive summary

Pharmaceutical organizations cannot answer whether “AI is validated” in the abstract. The decision belongs to a configured workflow: a defined task, a particular system and version, specified source material, named users, and a bounded level of authority. A useful decision must also account for the quality of the evidence the workflow uses, the ability of people to review it, and the organization’s ability to detect and correct failure.

This paper proposes an AI-Supported Work Authorization Cycle for that decision. It begins by inventorying the use and assigning ownership. It then defines the work and risk, sets an authority envelope, tests the complete human-AI workflow on representative cases, authorizes only the scope supported by evidence, and reconsiders the decision when the workflow changes or fails. A pilot, routine use, expansion, restriction, and retirement are separate decisions.

The method is designed for pharmaceutical manufacturing and quality operations. It works alongside the lifecycle-oriented, risk-based practitioner guidance in GAMP Guide: Artificial Intelligence and GAMP 5, while adding a local authorization decision for each configured human-AI workflow. It does not replace existing GMP responsibilities, validation, data-integrity controls, Part 11 analysis, or quality-unit accountability. It gives those activities a common operating question: what work is authorized now, under what limits, on what evidence, and what finding would require the organization to restrict or stop it?

The paper distinguishes reported industry experience from the proposed operating method. Public examples support the need to manage evidence access, human judgment, and change. They do not prove that this complete model is effective or cost-effective. The accompanying templates are practical records for a local decision; they are not regulatory forms and do not substitute for the organization’s own quality-system requirements.

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.

The method connects evidence adequacy, system authority, human performance, and correction. A model test or course completion alone cannot show that the configured workflow will produce acceptable work.

What industry experience tells us

Reported industry experience supports the need to evaluate work around the technology: Roche connects trusted data, people, process redesign, and governance; Lilly assigns employees accountability for AI outputs; and GSK reports generative AI in manufacturing investigations. These disclosures support the practical concerns behind this method. They do not prove that any complete governance model is effective. [1–5]

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]

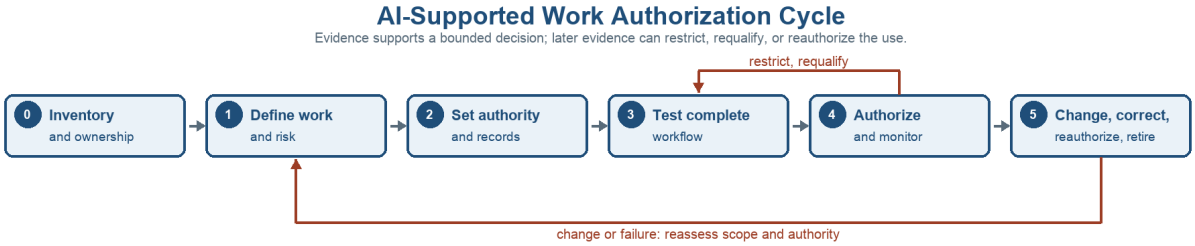
FDA's February 2026 Computer Software Assurance guidance applies to software used in medical-device production and quality-management systems under Part 820; its risk-based intended-use logic is only an analogy here. ISPE's July 2025 GAMP Guide: Artificial Intelligence, read alongside GAMP 5 Second Edition,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

How this cycle relates to GAMP AI

The proposed operating method is not a competing standard or a claim that GAMP requires these records. The next section introduces its AI-Supported Work Authorization Cycle; the Gate Decision Record and Authority Envelope are set out as practical templates in Appendices A and B. The added emphasis is a local authorization decision: for one configured human-AI workflow, define the authority boundary, test representative work, record the accountable decision, and state the conditions that would reverse it. Those elements make the link explicit between lifecycle assurance and the work the organization is permitting.

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. The Gate Decision Record in Appendix A records each decision; the Authority Envelope in Appendix B defines the permitted system boundary that supports 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

You cannot govern a workflow you have not found or assigned.

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. Appendix A provides a Gate Decision Record for showing 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.

Stage 1: Define the work and risk

Authorize the work, not the label on the tool.

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 belongs to that complete relationship. Change any element and the qualified state may change. [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.

Document the interface, connected tools, retrieval sources, configurations, permissions, and points where a person must intervene. Treat a retrieval change, model update, new product, altered interface, or grant of write access as a potential change to the reviewer's job. The product name can stay the same while the qualified state changes. A chronology drafted from a superseded deviation record can look sound and still steer an investigation in the wrong direction.

Set the human baseline at the same time. If the current process already misses records, produces inconsistent classifications, or takes an unacceptable amount of time, show that weakness. The purpose is to determine whether the proposed workflow produces acceptable work and whether its benefit justifies the added risks and controls. A comparison that measures only the AI condition can mistake inherited failures for tool failures or hide failures that both methods share.

Match the evaluation to the technology. For a classifier, test independent data, consequential defect classes, subgroup performance, calibration, and the consequences of false positives and false negatives. A strong average can conceal a poor result on a rare but consequential defect. A visual-inspection classifier can appear reliable across routine batches and still miss the defect family that triggers a product-quality decision.

Generative systems require verification of meaning. A fluent sentence can reverse a negation, remove a qualification, or imply causation where the source shows only sequence. Repeated answers offer limited reassurance when the same missing source or unsupported inference drives each answer. Assign controlled calculations to tested tools, permissions to deterministic rules, evidence identification to retrieval, and organizing or drafting to generative AI within those boundaries. Test each component for its assigned job, 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

Capability is not permission. Grant authority on purpose.

When AI moves from producing content to taking action, define exactly what the system may observe, interpret, recommend, decide, execute, and approve. Capability describes what the technology can do. Permission states what the organization has authorized it to do. Name accessible systems and records,

allowed operations, prohibited actions, approval points, stop conditions, and the person accountable for the result. Appendix B provides a practical template for making those boundaries specific and testable. [20]

Apply Part 11 with equal precision. It does not make every AI prompt or output a Part 11 record. Under FDA's scope guidance, it 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 serve as required signatures. 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 directly govern AI-supported approvals and actions. [21]

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

Enforce these limits technically. An instruction telling an agent not to approve a record provides weak protection when its credentials permit approval. Give the agent its own attributable identity and only the access required for its assigned function. Do not let it inherit all the privileges of the employee who launches it.

Tie approval to the state reviewed. If the underlying record changes after approval and before the agent acts, determine whether the approval still applies. When the change undermines the original basis, require another review. Permission to retrieve records does not confer authority to determine root cause, open a corrective action, or approve closure.

Keep enough of the record for another person to reconstruct the decision: the task and source versions, retrieved material and tool actions, the reviewer's response and approval, and the resulting system state. Treat a model's narrative explanation as context, not proof of how it reached the answer, and keep it separate from verifiable inputs, outputs, actions, and human reasons.

Decision provenance goes beyond logging. 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 need not preserve every internal computation a supplier cannot expose. It does need reliable records sufficient to explain its 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. Preserve both a sound decision and reliable records of how it was made. Retention should follow the task, applicable predicate rules, 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, yet historical records must still support an explanation of the regulated decision. When exact reproduction is required, make it a system-selection condition before approval.

Stage 3: Test the complete workflow

Test the full decision path, because a credible output can still produce a bad decision.

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.

Separate facts, calculations, explanations, and proposed actions. Compare a factual statement with its source, repeat a calculation, and test an explanation against alternatives and contradictory records. Keep chronology separate from causation. If an excursion follows an intervention, the sequence can justify investigating a connection; it cannot establish that the intervention caused the excursion.

Design review for the actual human task. The phrase “human in the loop” identifies a person’s position, not the effectiveness of the review. Effective review depends on competence, evidence access, available time, interface design, and practical authority to reject or escalate. Showing a persuasive AI conclusion before the records can bias the reviewer’s search. For consequential recommendations, require an initial evidence assessment before displaying the AI conclusion, or use another method that reduces anchoring. Lower-risk transformations may need only a direct source comparison. Human-factors research describes misuse as overreliance that can impair monitoring and judgment; a healthcare-focused systematic review found automation bias across clinical decision-support evidence and identified user, task, and system factors that affect it. These findings are not GMP validation evidence, but they support testing review conditions rather than assuming that a person’s presence is protective. [34,35]

Treat workload as a 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. Evaluate and monitor review time, queue size, escalation patterns, and whether staff can report uncertainty and near misses without pressure to approve. A five-minute review window for forty generated outputs turns review into sampling.

Preserve evidence identity during evaluation. Freeze the test collection, source versions, inclusion rules, expected findings, tool configuration, reviewer instructions, and scoring rules before results are examined. Record and justify any change made after results are examined. Reusing visible cases for repeated tuning turns a test into practice. Use fresh cases or a protected holdout set when results will support authorization.

Score more than average accuracy. Identify critical misses that easier successes cannot offset, false connections that could misdirect an investigation, correct findings rejected without basis, and unnecessary escalations that add burden. Measure the time needed to reach an acceptable evidence packet, not merely the time to generate a first draft. Record withdrawals and incomplete work rather than silently excluding them. For small samples, present uncertainty instead of converting a few cases into a precise-looking percentage.

Before evaluation, set the performance needed for the bounded use and the improvement that would justify the added intervention. Each control must earn its place through results.

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

Knowing that AI can make mistakes is where qualification starts, not evidence of readiness. Readiness is shown when someone can identify a consequential defect in representative work, explain why it matters, and accept, reject, correct, or escalate the result appropriately.

Build challenge cases around correct outputs, awkward but sound outputs, fluent dangerous outputs, incomplete records, wrong-product sources, and conditions where rejecting everything is itself an error. Count missed defects and unsupported accusations of error. Decide in advance which serious failures cannot be offset by success on easier items. Ensure assessors agree on acceptable work 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]

Match qualification to responsibility. A person using a bounded tool to reformat approved text needs a different demonstration from a quality approver evaluating an investigation or a system owner granting an agent write access. Training hours and course completion are inputs, not proof 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 fail by blaming the operator for impossible conditions. When the interface hides source status, the record set is incomplete, or volume makes careful review unrealistic, repeated retraining will not repair the workflow. Training cannot recover records the system never makes available.

Give each passing standard a reason. Define the outcomes that 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 lacks a dependable basis for using the result to grant permission.

Revisit qualification when its supporting conditions change. A reviewer qualified on one repository, interface, language, and authority level does not automatically remain qualified after the system searches additional repositories or initiates workflows. Target requalification to the changed function. It need not repeat every earlier activity, but retain an explicit rationale for the evidence used.

Challenge agents before routine use

The worked example below applies these controls to a bounded write-access task, including a partial-success failure, restriction, correction, and retest.

Build agents to 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.

Plan recovery around the action's reversibility. A record change may be reversed; an external notification cannot be unsent. Restoring one database may leave another system inconsistent. Specify reconciliation, compensating action, communication, and the authority to suspend the workflow.

Set operating hours that match the response arrangement. If an agent can take consequential action outside staffed periods, provide a response arrangement for those periods. Otherwise, limit operation to times when responsible people can intervene. User training cannot replace least privilege, state checks, observable execution, and tested recovery.

Worked example: a bounded agent with write access

The following is an illustrative application of the method, not a reported deployment or measured result. It shows how a team could evaluate a narrowly scoped agent after a deviation investigation has been approved. The agent receives the approved deviation identifier, opens one follow-up CAPA task in a linked system, and updates one designated field on the deviation record to show that the task was created. It does not determine root cause, alter the approved investigation, approve or close a CAPA, release product, or send external notifications.

Illustrative Authority Envelope. The service is named CAPA-Action-Agent-v1 and runs under a dedicated service identity, not the credentials of the investigator who launched it. It may read the approved deviation and linked CAPA-task status, create one task with the approved title and due-date pattern, and update the field "follow-up CAPA task created." It may run only during staffed hours. It must verify that the exact approved record version and approval state still apply and that no active task with the same deviation identifier and action type already exists.

Envelope element	Illustrative completed boundary	Verification evidence
Permitted actions	Read the approved deviation; create one linked CAPA task; update one task-created field.	Service-account permissions; action log; linked-record check.
Prohibited actions	No root-cause change, investigation edit, approval, CAPA closure, product disposition, release, or external notification.	Denied permissions; negative tests; daily exception review.

Envelope element	Illustrative completed boundary	Verification evidence
Execution conditions	Exact approval state and record version; no existing active matching task; staffed period; one execution attempt per approved trigger.	Approval-state binding; idempotency key; staffed-hours check; pre-action query.
Stop and escalation	Suspend for duplicate creation, stale approval, changed record, partial success, or concurrent edit.	Alert; incident; reconciliation; system owner.

Illustrative acceptance criterion. Before testing, the team could require zero critical authorization or duplicate-action failures across 60 protected holdout cases. Observing zero such failures in 60 cases does not establish a zero underlying rate: the rule of three gives an approximate one-sided 95% upper bound of 3/60, or 5%. If a 5% critical-failure rate would be unacceptable for the use, 60 cases are not enough evidence to authorize it. The organization must either narrow the use, increase the evidence base, or change the control design. [36]

Illustrative challenge cases. The test set deliberately includes: a timeout after the task is created but before the deviation field is updated; an approval that expired before execution; a deviation record edited after approval; and simultaneous edits by two users. The first case tests whether the agent retries blindly and creates a duplicate. The second and third test whether it acts on stale authority. The fourth tests whether it can detect a changed state instead of overwriting or misleading a reviewer.

Illustrative failure and correction. In one challenge case, task creation succeeds, the field update times out, and the agent retries the entire sequence. A duplicate CAPA task appears. The workflow is restricted because the agent acted twice on a single approved trigger. The investigation finds that the agent recorded neither a durable idempotency key nor a pre-action check for an existing task. The correction adds both checks, preserves the partial-success state for reconciliation, and sends a failure to a human queue rather than retrying the whole workflow. The correction is then tested on fresh partial-success cases and unaffected cases under the Stage 5 method; resolving the original duplicate alone is not evidence that the problem is corrected.

Illustrative gate decision. The process owner, quality-unit representative, system owner, and data owner review the evidence packet. They authorize only a bounded pilot: up to 30 previously approved deviations from one site, during staffed hours, with each action independently reviewable. The accepted residual risk is a missed or delayed follow-up after a failed state check; the pilot accepts that delay because no product or quality decision is changed and a human queue receives the exception. Any duplicate task, action on a stale approval, unauthorized write, inability to reconstruct an action, or failure of the human exception queue is a stop trigger. Routine use, another site, or any broader write permission requires a new decision.

Stage 4: Authorize routine use and monitor

Routine use is a renewed decision based on current conditions, not a permanent entitlement.

A completed evaluation supports only the conditions tested. Before authorizing routine use, require the process and quality owners to review pilot performance, critical failures, unresolved risks, review burden, recovery results, and operating controls. Treat another task, site, source collection, user group, or authority level as a new decision. A clean pilot at one site does not establish that the same controls hold when a second site uses a different record set and staffing model.

Revisit 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 results can inform earlier decisions while authority for protocol and regulatory changes remains human. It is a strategic model, not a GMP control standard. [26]

Monitor 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. A monthly average can conceal the one document type that now sends reviewers to the wrong source. Keep small samples and uncertainty visible.

Interpret overrides and near misses in 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

A governance method earns trust only if it can find failure, correct it, and restrict or retire the work.

Assess what a change could affect before selecting 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. Document which earlier results still apply and which conditions require new testing or requalification.

Deliver approved changes to 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. Use a governed applicability process to 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 relevant records, restrict affected use when warranted, and investigate human and system causes. Retraining fits a missing skill. It will not repair missing records, an interface that hides limitations, a defective permission design, or unreliable equipment. Test corrective action on fresh cases that reproduce the failure mechanism and on unaffected cases that can reveal 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 shows that an action occurred. It does not show 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

Measure the whole job: time to acceptable work, review, correction, escalation, maintenance, and downstream rework. Drafting speed alone is not sufficient. [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, define what should improve and what cannot worsen. Compare the current process with the bounded AI use on time to acceptable work, consequential errors, unnecessary escalation, review burden, and maintenance cost.

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 as described in Stage 5 before 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.

Appendix A: Gate Decision Record

Use this record to connect evidence to an explicit decision at any stage of the cycle. It is a proposed operating template, not a regulatory form. Completing reviews, tests, or training does not itself authorize the work.

Required entry	Record the decision basis
Decision identification	Use ID; workflow title; cycle stage; decision date; decision owner; linked inventory entry.
Work proposed	Intended task, expected output, named users, downstream decision, products/sites/records in scope, and exclusions.
Configured workflow	System, model and version; interface and integrations; source boundary; assigned roles; current permissions.
Evidence reviewed	Risk assessment; technical and supplier evidence; representative-work results; human-role qualification; security/privacy review; recovery testing; open issues.
Decision	Authorize bounded pilot; authorize routine use; continue under conditions; narrow scope; restrict; stop; retire. State the effective date.
Authorized scope and limits	What the workflow may read, draft, recommend, change, execute, or approve; prohibited actions; required human approvals.
Conditions and compensating controls	Required review method, operating hours, volume limits, monitoring, escalation path, and any temporary safeguards.
Accepted residual risk	Remaining risks after controls; why they are acceptable for the authorized scope; compensating controls; accountable owner accepting the residual risk; date and review trigger.
Stop and re-review triggers	Critical failure, performance threshold, changed source or permission, model/configuration change, incident, expiry date, or other trigger.
Records and retention	Regulated record determination; linked evidence; audit/provenance records; retention location and period; investigation access.
Approvals and accountability	Process owner; system owner; data owner; quality-unit representative; security/privacy owner where applicable. Record name, role, date, and required signature or electronic approval.

Appendix B: Authority Envelope

Define permission separately from technical capability. The boundary below should be specific enough to configure, test, monitor, and inspect. A prompt instruction is not an authority control when the technical permission remains available.

Envelope element	Authorized boundary	Enforcing or Verifying Evidence
Identity and version	System/service identity, model and configuration version, environment, integration versions, and accountable system owner.	Approved configuration baseline; change-control record.
Permitted users and roles	Who may initiate, review, approve, administer, or support the workflow; segregation-of-duties limits.	Role matrix; account/role configuration; training or qualification record.
Accessible sources	Named systems, repositories, data classes, sites/products, document statuses, and exclusions.	Access review; source inventory; retrieval or integration configuration.
Allowed operations	Read, retrieve, transform, draft, classify, recommend, create, update, notify, or execute only if approved.	Least-privilege configuration; workflow design; action logs.
Conditional actions	Actions allowed only after a named approval, a state check, a validated confidence threshold where one exists, or a staffed-period restriction.	Approval-state binding; operational checks; tested scenarios.
Prohibited actions	Actions the workflow must never perform, including any approval, release, disposition, record change, notification, or external action outside scope.	Denied permissions; negative test evidence; monitoring alert.
Technical enforcement	Service identity, authentication, access controls, action checks, rate/volume limits, and separation from the initiating user's privileges.	Configuration evidence; security review; test results.
Stop, recovery, and escalation	Who can suspend use; automatic stop conditions; recovery/reconciliation method; contact path outside normal operating hours.	Recovery test; incident procedure; on-call or staffing arrangement.
Change and review triggers	Events requiring reassessment: model, prompt/configuration, source, interface, user role, product/site, volume, or permission change.	Change-control linkage; review date; monitoring plan.

Approval statement: the accountable process and quality owners confirm that the technical permissions and operating controls match this authority envelope. Any expansion beyond the recorded boundary requires a new decision before use.

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

[34] Raja Parasuraman and Victor Riley. “Humans and Automation: Use, Misuse, Disuse, Abuse.” *Human Factors* 39(2), 230–253 (1997). Read source

[35] Kate Goddard, Abdul Roudsari, and Jeremy C. Wyatt. “Automation Bias: A Systematic Review of Frequency, Effect Mediators, and Mitigators.” *Journal of the American Medical Informatics Association* 19(1), 121–127 (2012). Read source

[36] J. A. Hanley and A. Lippman-Hand. “If Nothing Goes Wrong, Is Everything All Right? Interpreting Zero Numerators.” *JAMA* 249(13), 1743–1745 (1983). Read source

Author note: The author founded GxP Frame and the Life Sciences AI Academy.