Sandboxes, Stewards, and the Governance Gap

An Organizational Model for AI in Regulated Manufacturing

Martin Kockx | Senior Manufacturing Technology & Digital Transformation Lead | May 2026

Abstract

Centralized AI governance in regulated manufacturing creates a structural paradox. Restrictive policies intended to manage risk drive AI use underground into a shadow tier — invisible, ungoverned, and disconnected from organizational learning. Survey data confirms the scale: more than half of employees report willingness to use AI tools without employer approval, and the majority access them through personal accounts. Training investments decay without practice environments. Centralized AI teams cannot scale governance across domains they do not understand. The gap between who understands the problem and who controls the tools is where organizational value is lost. In April 2026, the FDA's first warning letter citing inappropriate AI use in pharmaceutical manufacturing turned the shadow tier from a theoretical risk into a documented enforcement trigger.

This paper proposes an integrated organizational architecture with three interlocking elements: sandbox environments for structured experimentation, a competency tier model coupling training with tool access, and AI stewards — named individuals with dual fluency in domain expertise and AI literacy, carrying explicit accountability for AI use within their function. The framework draws on user innovation theory, training transfer research, and collective intelligence frameworks to argue that governance in regulated environments must be distributed to the practitioners closest to the work.

While individual components — regulatory sandboxes, training programs, distributed governance roles — exist independently across regulated industries, this paper argues they are interdependent by design. The integration is the contribution: sandboxes without stewards produce ungoverned experimentation; stewards without sandboxes have nothing to govern; training without either produces negligible transfer. Cross-industry evidence from financial services, nuclear energy, aerospace, and food safety supports the generalizability of the structural argument. The framework introduces an Attributable-Observable-Correctable design discipline for human oversight of AI outputs and a regulated-context specification — the GxP steward — for functions where AI influences quality decisions.

The model is proposed, not validated. The paper engages honestly with its limitations and defines concrete validation criteria for future empirical work.

Section 1: Introduction — The Governance Gap

In April 2026, the FDA issued its first warning letter citing inappropriate use of artificial intelligence in pharmaceutical manufacturing (CDER, 2026). The firm — Purolea Cosmetics Lab, despite the name a manufacturer of drug products under CDER's CGMP authority — had used AI agents to generate drug product specifications, procedures, and master production records, then released them into a quality system without quality unit review. The agency's remedy is now on record: any AI output used in CGMP activities "must be reviewed and cleared by an authorized human representative of your firm's QU." The first regulatory landmine in this space has been mapped. There will be others.

The conventional reading of this letter is that AI requires stricter controls — better policies, tighter access, more validation gates before any AI output can touch regulated work. That reading is incomplete and, in practice,

counterproductive. The Purolea pattern is not the failure of a permissive AI environment. It is the failure of an environment in which AI use was happening but governance was nowhere — neither in the design of the AI use, nor in the review of its output, nor in the system that should have caught both. Restrictive policies do not produce safety in that environment. They produce *invisibility*. The work still needs to get done. The tools still exist. The only variable that changes when access is denied is whether the organization can see the use.

This is the governance gap: the distance between who controls AI access and who actually uses AI to do the work. When that gap is wide, three things happen simultaneously. Practitioners route around restrictions through personal accounts on personal devices — the shadow tier (Section 2). The organization loses the ability to detect, govern, or learn from AI's influence on regulated decisions. And when an enforcement event eventually occurs, the firm discovers that the AI use it could not see was nonetheless attributable to it. Survey data confirms the scale of the gap as it exists today: more than half of employees report willingness to use AI tools without employer approval, more than two-thirds of those using AI at work do so through personal accounts, and only 40% of organizations whose employees use AI maintain official AI subscriptions (BCG, 2025; TELUS Digital, 2025; Harmonic Security, 2025). Shadow AI is influencing GxP work today; the only open question is whether organizations have a mechanism to see, govern, and improve that influence.

The mechanism that produces and sustains the gap is structural. The dominant organizational model for AI in manufacturing is centralized: a data science team, sometimes branded as an AI Center of Excellence, receives requests from operational functions, translates them into technical problems, builds solutions, and returns them to the business. This model has a strong intuitive logic. AI requires specialized skills — statistical modeling, machine learning engineering, prompt design, model evaluation — that most manufacturing practitioners do not possess. Concentrating those skills in a dedicated team appears efficient. In practice, it is a reliable engine for underperformance, and it is also the engine that produces shadow use.

The failure begins with knowledge asymmetry. A centralized data science team, however technically capable, does not understand the operational context in which its solutions must function. A deviation investigation in a biopharmaceutical manufacturing facility is not a generic text-summarization problem. It involves regulatory history, equipment-specific failure modes, batch genealogy, operator experience, and a quality system with specific evidentiary standards. The central team lacks this context. The domain practitioner who requested the solution possesses it but cannot transfer it through a requirements document. The result is a translation tax: the practitioner must convert a rich, contextual operational problem into a simplified specification that the central team can act on, and the central team must build a solution that satisfices against that simplified specification rather than the actual problem. The round trip — domain to data science to domain — strips away the nuance that determines whether a solution works in practice or merely works in a demonstration. The same erosion occurs in process knowledge itself: manufacturing intent — why parameters were set, what was tried and rejected, what the true envelope of acceptable variation is — degrades with each successive handoff from development to clinical to commercial to contract manufacturing organization. By the time a process is running at commercial scale in an external partner's facility, it is optimized for someone else's intent, executed by operators whose understanding derives from documentation that has long since shed its reasoning.

This is not a coordination problem that better project management can solve. It is a structural consequence of separating domain expertise from tool access. The practitioners who understand the problem cannot directly experiment with solutions. The specialists who can build solutions do not understand the problem deeply enough to build the right ones. The shadow tier is what fills the void: practitioners who cannot get capable tools through legitimate channels acquire them through personal channels and continue to do the work. The gap between these two populations — and the shadow use that flourishes in it — is where organizational value is lost and where regulatory risk silently accumulates.

The pattern is visible across industries, but it is especially costly in regulated manufacturing. In GxP environments, context is not merely helpful — it is a compliance requirement. A deviation investigation must trace causal reasoning. A batch disposition must document the evidentiary basis for the decision. When an AI-

assisted solution is built by a central team that does not understand these requirements at the operational level, the result is either a tool that practitioners do not trust, a tool that does not meet regulatory expectations, or — most commonly — a tool that is technically functional but organizationally orphaned because no one in the domain feels ownership over it. The practitioner, faced with that tool and a deadline, falls back to a personal AI account. Purolea did not invent this pattern; it merely produced the first instance the FDA chose to write up.

Eric von Hippel's research on user innovation provides the theoretical frame for understanding why this happens and what the alternative looks like. Across decades of empirical work spanning industrial equipment, medical devices, software, and consumer products, von Hippel demonstrated that a substantial share of commercially significant innovations originates not from producers or centralized R&D functions but from users — the practitioners who work directly with products and processes to solve their own problems (Von Hippel, 1988). Users innovate because they possess two things that central teams lack: intimate knowledge of the problem and direct motivation to solve it. When organizations provide users with capable tools and supportive infrastructure, the rate and quality of innovation increases. When organizations restrict user access to tools and centralize innovation in specialist functions, they systematically suppress the highest-value source of problem-solving (Von Hippel, 2005).

The application to AI in manufacturing is direct. The process engineer who understands alarm patterns on a specific production line, the quality analyst who knows which deviation categories are genuinely investigable versus which are documentation artifacts, the batch reviewer who can distinguish a meaningful trend from instrument noise — these practitioners are, in von Hippel's framework, the lead users (Von Hippel, 1986). They experience needs ahead of the broader organization and have the domain sophistication to evaluate potential solutions. A centralized AI team, no matter how technically skilled, is structurally incapable of replicating this proximity to the problem.

The conventional AI Center of Excellence compounds this structural weakness by gatekeeping rather than enabling. In the typical model, the center of excellence controls tool access, prioritizes use cases, and builds solutions on behalf of the business. Practitioners submit requests and wait. The center of excellence becomes a bottleneck — not because its members are slow, but because demand for AI-assisted problem solving across a manufacturing organization vastly exceeds the capacity of any centralized team to deliver. Meanwhile, practitioners who could be experimenting with solutions to their own problems are idle, waiting for the queue to clear, and the most motivated among them are already using personal AI accounts to make their deadlines. The model proposed in this paper retains a center of excellence — but redefined: a coordinator and connector of distributed practice, not a gatekeeper of access. Subsequent references to the center of excellence carry this redefined meaning.

This paper argues that the solution is not a better centralized team and not stricter restrictions, but a fundamentally different organizational architecture: one that distributes AI capability to practitioners while maintaining the governance infrastructure that regulated environments require. The model has three interlocking elements. First, *sandboxes* — structured experimentation environments where practitioners can work with AI tools on representative data without regulatory exposure, and where the use is logged. Second, a *competency tier model* — a development pathway that couples training with tool access at each level, building practitioner capability from awareness through applied proficiency. Third, *stewards* — named individuals embedded in domain functions with dual fluency in domain and AI, accountable for keeping experimentation visible, attributable, and escalatable. Together, these three elements close the governance gap by collapsing it: domain practitioners gain legitimate access to capable tools inside governed boundaries, the shadow tier loses its rationale, and the organization regains the ability to see the AI influence that is already shaping its regulated work.

Section 2: The Shadow Tier

Before proposing solutions, it is worth examining what happens when organizations fail to provide legitimate paths to AI tools. The answer isn't abstinence. It is invisibility.

Somewhere in most manufacturing facilities today, a quality analyst is pasting deviation language into a personal AI account to draft an investigation summary. A process engineer is feeding alarm data into a free-tier tool to identify patterns no one has time to analyze manually. A batch record reviewer is using a personal subscription to summarize a protocol before a deadline. They are doing it because it works. They are doing it quietly because they know it is not sanctioned. And the organization cannot see any of it.

This is the shadow tier: AI use that occurs on personal devices, through personal accounts, completely outside organizational visibility. It is distinct from both sanctioned tools and restricted tools — it occupies a third category that most governance frameworks do not acknowledge.

Ethan Mollick describes the employees who inhabit this tier as "secret cyborgs" — people who have discovered that AI substantially augments their capability but who hide their use because organizational policy offers no legitimate path to the tools they need (Mollick, 2024). They are not rogue actors. They are rational people responding to an environment that gives them two options: comply and forgo the productivity gain, or use the tools covertly and capture the benefit personally. Most choose the latter. The organizational learning loop — documentation, peer review, knowledge capture, escalation — breaks entirely.

The scale of this phenomenon is not speculative. BCG's 2025 survey of more than 10,000 employees across 11 countries found that 54% would use AI tools even without their employer's approval (BCG, 2025). TELUS Digital's 2025 research found that 68% of employees using generative AI at work access it through personal accounts, and 57% have entered sensitive company information into public AI tools. Even among employees with company-provided AI access, 22% still use personal accounts — suggesting the gap is not just about availability but about usability and capability (TELUS Digital, 2025). Harmonic Security's analysis of over 22 million enterprise AI prompts found that employees at more than 90% of organizations use AI tools, but only 40% of those organizations have official AI subscriptions (Harmonic Security, 2025). The CybSafe and National Cybersecurity Alliance report found that 38% of employees share sensitive information with AI tools without employer knowledge, while 52% have received no training on safe AI use (CybSafe & NCA, 2024).

The conventional governance response frames the shadow tier as a compliance problem. Employees are not following rules. The solution, in this framing, is more rules — stricter policies, tighter controls, better training on acceptable use. But this diagnosis misidentifies the mechanism. Restrictive policies do not eliminate demand for AI tools. They displace it. The work still needs to get done. The tools still exist. The only variable that changes is visibility. When use goes underground, the organization bears all the risk — data leaving controlled environments, outputs entering regulated workflows without traceability — while capturing none of the value.

Von Hippel's user innovation framework predicts exactly this outcome. When organizations suppress practitioner access to capable tools, they do not get safety. They get suppressed innovation (Von Hippel, 2005; Von Hippel, 2017). The shadow tier is what user innovation looks like when it is driven underground: individual practitioners solving real problems with real tools, generating real value, but doing so in ways that produce zero organizational benefit. There is no documentation of what was tried. No peer review of the output. No escalation path when something goes wrong. No knowledge capture for the next person who faces the same problem. Each discovery is local and ephemeral.

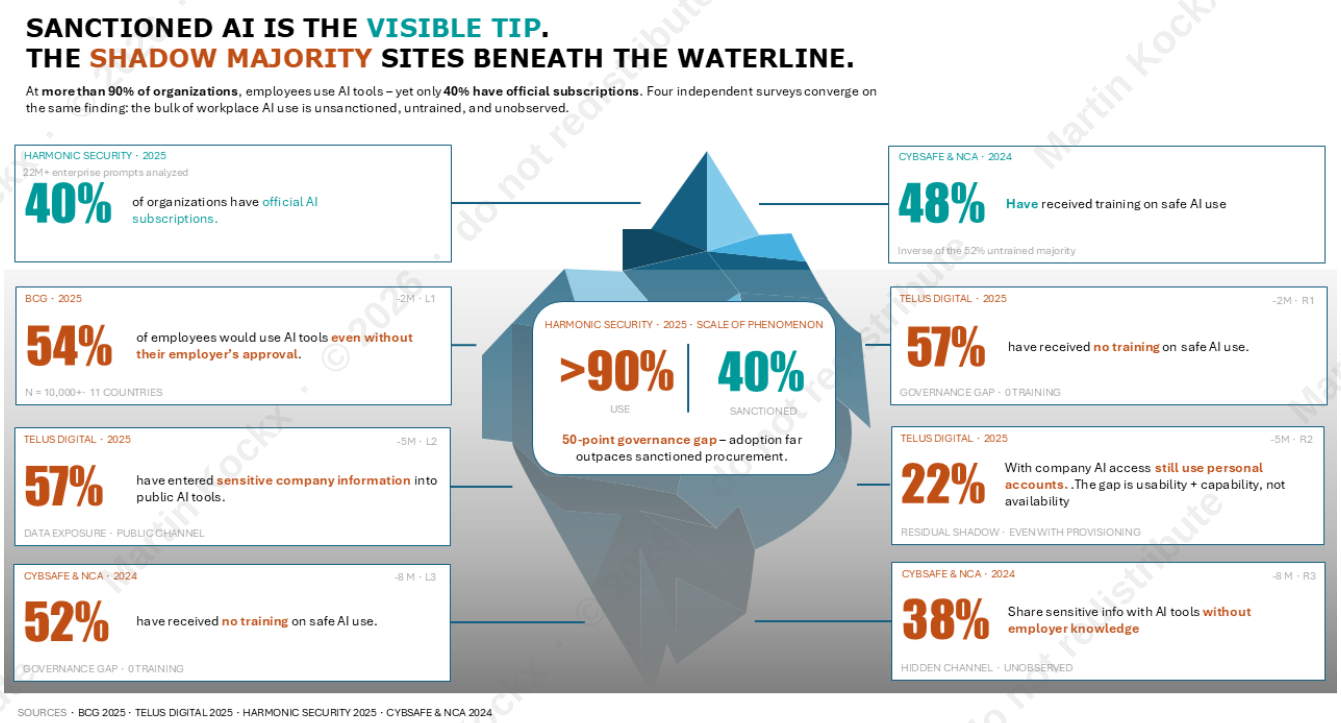

Figure 1. *The shadow tier problem: restrictive governance displaces AI use into invisible, ungoverned channels, breaking the organizational learning loop.*

In regulated manufacturing, the consequences extend beyond lost innovation. In a GxP environment, shadow AI creates invisible influence pathways into regulated decisions. When someone uses an undocumented AI tool to inform a deviation investigation, draft language for a CAPA assessment, or summarize data for a batch record review, the AI's contribution is untraceable. The tool is not validated. The prompt is not documented. The reasoning chain has an invisible dependency that the quality system has no mechanism to detect. Consider: a quality analyst uses a personal AI account to interpret a trend in environmental monitoring data. The AI's summary influences their investigation narrative, which shapes the CAPA, which drives a process change. At no point does the AI tool appear in the investigation record. The entire chain of reasoning has an undocumented dependency — and the organization does not know it exists (EY Switzerland, 2024).

This is not a hypothetical edge case. Given that more than half of employees report willingness to use AI without approval, and more than two-thirds of those using AI at work do so through personal accounts, shadow AI influence on GxP decisions in pharmaceutical manufacturing facilities today is highly probable — though the cited surveys measure willingness and usage, not GxP-specific influence directly. The question is not whether it is happening but whether the organization has any mechanism to detect, govern, or learn from it.

The first FDA enforcement signal arrived in April 2026: a Warning Letter to Purolea Cosmetics Lab — despite the firm name, the letter concerns drug products under CDER's CGMP authority — cited violations under a section explicitly titled *Inappropriate Use of Artificial Intelligence in Pharmaceutical Manufacturing*. The firm had used AI agents to generate drug product specifications, procedures, and master production records without quality unit review, and had skipped process validation because — as the firm told inspectors — "the AI agent never told [them] it was required." The agency's remedy: any AI output used in CGMP activities "must be reviewed and cleared by an authorized human representative of your firm's QU" (CDER, 2026). Shadow AI in regulated manufacturing is no longer hypothetical. It is now a documented enforcement trigger.

Restrictive governance is strategic self-harm, misclassified as risk management.

Section 3: The Sandbox — Structured Experimentation

If the shadow tier represents the failure mode of restrictive governance, the sandbox represents the design response: a structured experimentation environment that makes AI use visible, attributable, and improvable — converting the shadow tier's ungoverned individual experimentation into a governed organizational learning process.

The sandbox concept is not novel. Regulatory sandboxes have been deployed across multiple regulated industries as a mechanism for enabling innovation within governance constraints. The OECD's 2023 policy paper on AI regulatory sandboxes documented approximately 30 jurisdictions with AI-related sandbox programs, noting benefits including accelerated market entry, improved regulatory understanding, and stimulated investment (OECD, 2023). The EU AI Act mandates that all member states establish national or regional regulatory sandboxes for AI, with emphasis on annual reporting and priority access for smaller organizations — a signal that sandbox approaches have moved from experimental to institutionally required (EU AI Act, 2024). The U.S. Nuclear Energy Agency's RegLab project brings together technologists, operators, and regulators in a multi-stakeholder sandbox to examine AI deployment in nuclear power plant operations, with explicit competency development recommendations alongside technical testing (NEA, 2024-2025).

The strongest empirical evidence for sandbox effectiveness comes from financial services. Cornelli et al., in a study published in the *Review of Finance*, found that entry into the UK Financial Conduct Authority's fintech sandbox increased capital raised by 15% over the following eight quarters, increased the probability of raising capital by 50%, and was associated with higher rates of firm survival and patenting. The mechanism identified was reduction of information asymmetries and regulatory costs — the sandbox made innovation legible to external stakeholders (Cornelli et al., 2020/2024). Wang et al.'s systematic review of 20 studies on regulatory sandbox effectiveness confirmed ecosystem spillover effects: funding increases for non-participants once a sector has its first sandbox entrant, suggesting that sandboxes generate systemic rather than merely firm-level benefits (Wang et al., 2025).

For manufacturing organizations, the sandbox serves a dual purpose that distinguishes it from purely regulatory applications. It is simultaneously an **innovation environment** and a **governance mechanism**. The innovation function is straightforward: practitioners gain access to representative data and capable tools in a setting where experimentation carries no regulatory risk. The governance function is less obvious but equally important: because the sandbox is instrumented — all queries, models, and outputs are logged — it converts invisible AI use into visible, attributable, reviewable organizational activity. The sandbox does not merely permit experimentation. It makes experimentation legible.

Three sandbox types address different experimentation needs:

Data sandbox. The foundational experimentation layer. Practitioners work with representative or synthetic datasets that preserve the statistical properties of real manufacturing data — historian time series, alarm logs, batch records — without exposing controlled information. The design principles are straightforward: the data must be realistic enough that experiments conducted on it transfer meaningfully to production analysis and protected enough that governance risk is eliminated by construction. A process engineer learning to apply AI to trend analysis needs data that exhibits the distributions, correlations, and failure modes of actual production data, not a generic tutorial dataset. Synthetic data generation from statistical profiles of real manufacturing data, supplemented by anonymized historical records where regulatory review permits, achieves this balance. The data sandbox is the lowest-barrier entry point: any employee with basic orientation can access it. This matters because the first requirement of distributed innovation is broad access.

Tool sandbox. Approved AI platforms — large language models, coding assistants, agentic frameworks — available for practitioner experimentation on work problems. This is the primary entry point for most practitioners: the environment where they can use AI tools on domain problems within governed boundaries. The tool sandbox serves the majority of the workforce that needs to develop applied AI competency through direct use rather than classroom instruction. For practitioners who advance to building custom applications — developing agents, constructing retrieval-augmented generation systems, prototyping integrations with manufacturing data sources — an **advanced builder tier** within the tool sandbox provides deeper infrastructure access with commensurate training requirements. The tool sandbox is where the shadow tier is absorbed: practitioners who were previously using personal AI accounts on work problems now have a sanctioned, instrumented, supported alternative.

Process sandbox. The highest-fidelity experimentation environment. Simulated GxP workflows — batch record review, deviation investigation, operator decision support — allow practitioners and governance leads to test AI-assisted decisions before they touch anything regulated. A quality analyst can explore how AI summarization handles deviation language without that output entering a quality system. An operator can interact with an AI copilot prototype using synthetic batch data in a staging environment that mirrors production architecture. The process sandbox is where experiments are stress-tested against regulatory requirements: audit trail integrity, decision attributability, data governance compliance. Access is restricted to experienced practitioners and governance leads because the experiments conducted here are the direct precursors to validated deployment.

Four design principles govern all three sandbox types. **Representative**: data and workflows must be realistic enough to generate transferable learning. **Isolated**: there must be no pathway from sandbox to production GxP systems, enforced at network, identity, and application levels. **Instrumented**: all activity is logged — queries, models, artifacts, escalation proposals — creating both institutional memory and governance audit trails. **Escalatable**: there must be a clear, lightweight path from sandbox experiment to formal evaluation, so that successful discoveries carry forward as documented evidence rather than requiring re-justification from scratch.

The escalation pathway is where the sandbox's governance function becomes explicit. A sandbox experiment that demonstrates value does not graduate to production through ad hoc advocacy. It follows a structured escalation discipline: the practitioner documents the finding, an AI steward evaluates whether the experiment merits formal assessment, and the application enters a governance gate process that evaluates technical readiness, business justification, and risk profile. This discipline ensures that the sandbox is not a free-for-all but a structured pipeline — broad entry, instrumented experimentation, governed escalation. The boundaries are what make the sandbox safe enough to be sanctioned, and the instrumentation is what makes it valuable enough to justify the investment.

The sandbox, then, is not merely infrastructure. It is organizational commitment made tangible. It says to practitioners: we trust you enough to let you try, and we have built the environment to make your trying visible, safe, and valuable to the organization.

Section 4: The Competency Tier Model

The sandbox solves the environment problem — where practitioners can experiment — but not the capability problem. An experimentation environment without a development pathway produces uneven, undirected exploration. Conversely, a development pathway without an experimentation environment produces the most common failure mode in corporate AI upskilling: training that decays because there is nowhere to apply it.

The training transfer literature is unambiguous on this point. Baldwin and Ford's seminal 1988 study established that roughly 10% of training content transfers to actual job performance, with skill maintenance curves showing rapid decay when opportunities to apply are absent (Baldwin & Ford, 1988). Blume et al.'s 2010 meta-analysis of 89 empirical studies confirmed that a supportive work environment — meaning the opportunity, encouragement,

and infrastructure to apply what was learned — is one of the strongest predictors of whether training transfers to job performance (Blume et al., 2010). The finding is not controversial in the learning sciences. It has been replicated for decades. Yet organizations continue to invest in AI awareness programs while providing no sanctioned environment where that awareness can become competence.

The mirror failure is equally damaging. Organizations that deploy AI tooling without investing in upskilling find that users underutilize tools they do not understand, misapply them in ways that erode trust, or distrust them entirely and route around them. BCG's 2025 survey found that only 36% of employees were satisfied with their AI training, and 18% of regular AI users reported receiving no training at all (BCG, 2025). McKinsey found that 48% of employees rank training as the most important factor for AI adoption — and nearly half report receiving minimal or none (McKinsey, 2025).

The coupling principle is straightforward: **training without tooling is quickly forgotten; tooling without training is wasted.** Both are expensive. Both are common. And in most organizations, the two investments are managed by entirely different groups who rarely coordinate. The competency tier model addresses this by designing training and tool access as a single integrated system rather than parallel initiatives.

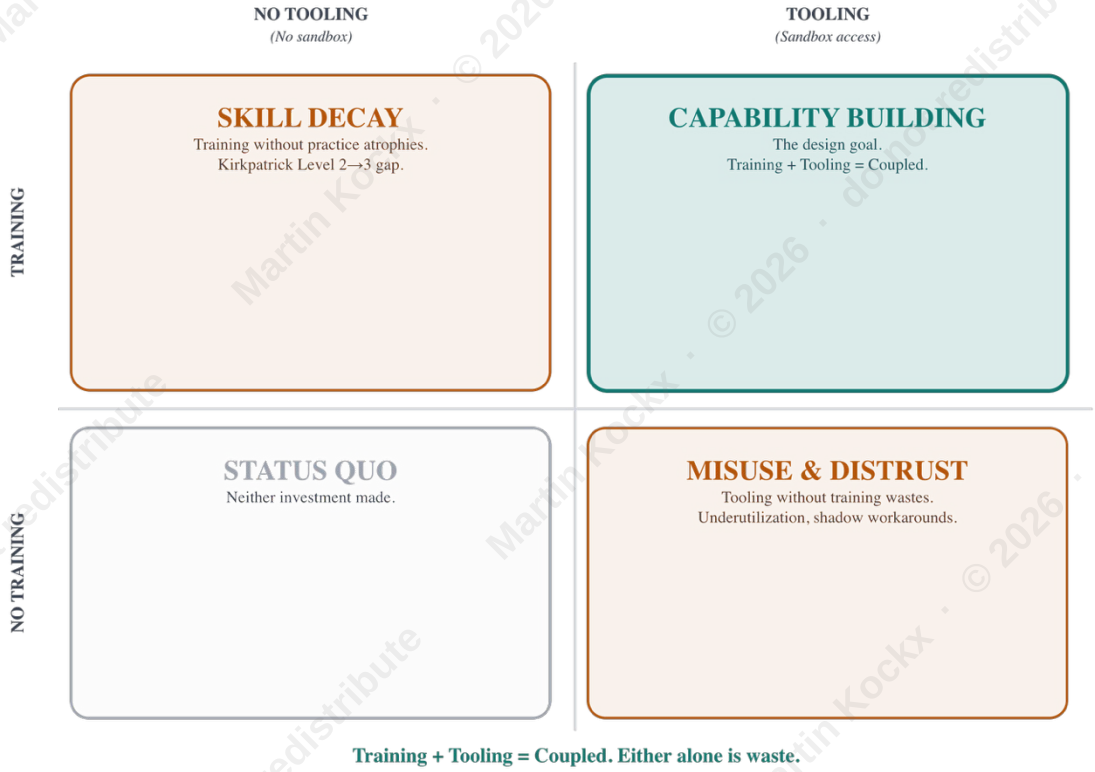

Figure 2. *The coupling principle: training and tooling must be paired. Either investment alone produces waste — skill decay (training without practice) or misuse (tools without preparation).*

The model defines three tiers, each coupling learning with application:

Tier 1: AI Awareness. The broadest tier, targeting all employees. Foundational understanding of what AI can and cannot do, organizational AI governance policies, responsible use principles, and — critically — how to access sandbox environments and learning resources. Tier 1 is not a passive lecture. It is an orientation that ends with sandbox access: the participant leaves knowing not just what AI is but where to try it. This design ensures that the gap between Kirkpatrick's Level 2 (Learning) and Level 3 (Behavior) — the gap where most corporate training dies — is as narrow as possible (Kirkpatrick & Kirkpatrick, 2006). The participant does not learn

concepts and then wait months for tool access. They learn concepts and walk into the data sandbox the same week.

Tier 2: AI Proficiency. Practitioners with demonstrated interest and manager support develop applied skills through intensive training coupled with hands-on projects conducted in sandbox environments. Tier 2 is where the sandbox becomes pedagogy. Instead of a multi-day workshop followed by months of waiting for tool access, the training design uses iterative cycles: learn a concept, apply it in the sandbox, observe the result, refine. The participant builds competence in context from the start — not learning abstract concepts and later trying to map them to their work, but working on their actual domain problems with AI tools under guided conditions. Tier 2 graduates earn authenticated access to production-data sandbox environments and have demonstrated the ability to frame manufacturing problems as AI opportunities, evaluate AI outputs critically, and document findings for organizational learning. The Kirkpatrick "New World" model's Level 3 "required drivers" — processes that reinforce, monitor, encourage, and reward learners to apply what was learned — are built into the sandbox infrastructure itself (Kirkpatrick & Kirkpatrick, 2016).

Tier 3: AI Governance. The steward tier. Selected practitioners with deep domain expertise and demonstrated AI proficiency develop the additional capabilities required for governance: risk assessment methodology, escalation authority, post-deployment monitoring, and the ability to mentor peers. Tier 3 is not a promotion out of the domain — it is an expansion of responsibility within it. Stewards remain embedded in their functional teams, maintaining the domain proximity that makes their governance judgments credible. Their mandate is threefold: monitoring responsible AI use within their function, shepherding promising applications through the escalation pathway from sandbox to validated deployment, and managing continuous post-deployment review as AI systems evolve.

The three-tier structure has a direct precedent in the Six Sigma belt system. The American Society for Quality's certification hierarchy — Yellow Belt (awareness), Green Belt (part-time practitioner), Black Belt (full-time expert) — demonstrated that organizations sustain operational transformations not through concentrated expertise at the top but through distributed competency at multiple levels (ASQ, n.d.). A broad base of Yellow Belts sustains improvement culture. A larger cohort of Green Belts drives process-level improvements. A smaller number of Black Belts leads strategic projects. The AI competency model follows the same architecture: broad awareness creates organizational readiness, applied proficiency generates distributed innovation, and governance expertise ensures that innovation remains safe and sustainable.

The critical insight is that the three elements — sandboxes, the competency tier model, and stewards — are not independent initiatives that happen to coexist. They are interdependent by design. The sandbox gives Tier 2 practitioners a place to practice. Tier 2 training gives practitioners the skills to use the sandbox productively. Tier 3 stewards provide the governance that makes sandbox freedom sustainable. Remove any one element and the others lose their function: a sandbox without trained practitioners is underutilized infrastructure; training without a sandbox is forgotten knowledge; governance without distributed competency collapses into the centralized gatekeeping model that created the problem in the first place.

Section 5: The AI Steward

The preceding sections establish three interlocking problems. Restrictive governance drives AI use underground (Section 2). Training without practice environments produces negligible transfer (Section 4). Sandboxes provide the experimentation infrastructure but create their own governance question: who watches what happens inside them, and who decides when an experiment is ready to leave?

The AI steward is this paper's proposed answer — and its most distinctive contribution: a translation of distributed-governance precedents (data stewards, CRAs) into AI-specific stewardship.

The Role

The AI steward is not a committee, a working group, or a shared mandate. It is a named individual carrying explicit accountability — a secondary role occupying roughly 10–20% of an experienced practitioner's time, owned by one person whose domain function is at stake. The defining selection criterion is not technical AI expertise; that is what training and sandbox access provide. The defining criterion is **dual fluency**: deep domain knowledge *and* sufficient AI literacy in the same person. The central steward question — *is this AI application appropriate, in this context, for this decision?* — cannot be answered by a domain expert without AI literacy, nor by an AI specialist without domain context. Both halves must live in the same head. Candidates should have at least five years of experience in their manufacturing or quality function, demonstrated initiative, comfort with ambiguity, and credibility with both leadership and the shop floor.

The single-name principle matters operationally. Distributed governance only works when accountability is not itself distributed. A "steward function" owned by no one in particular reproduces the central-team failure mode at smaller scale: when something goes wrong, no one is positioned to act because no one's name is on it. Other practitioners contribute. The steward decides.

This selection logic is deliberate. A centralized AI team, however technically capable, cannot understand every process deviation pattern, every alarm interaction, every calibration history across every manufacturing line. The governance judgment requires knowledge that does not centralize. The people closest to the work are the ones best positioned to evaluate whether an AI application makes sense in that work context. This is a structural observation, not a criticism of central teams.

A note on terminology: when this paper calls governance "distributed," it describes the *network*, not the accountability. Each steward role is held by a named individual; the network of those roles is what is distributed. Diffuse responsibility is the failure mode the steward role is designed to prevent.

Three mandates define the role:

Monitor responsible use. The steward ensures that AI tools within their function are used appropriately — within sandbox boundaries, with proper documentation, consistent with organizational policy. This is not surveillance. It is the same contextual judgment a data steward applies when evaluating data quality: someone who understands the domain is watching for misuse that a central team would not recognize.

Shepherd applications through escalation. When a sandbox experiment shows promise, it needs a governed path from prototype to pilot to production. The steward is the bridge. They understand what the experiment does, why it matters to the process, and what risks must be addressed before it scales. Without this bridge, promising experiments remain trapped in perpetual proof-of-concept — interesting but organizationally inert. The steward converts the sandbox from an isolated playground into an innovation pipeline with a real path forward.

Manage post-deployment review. Once an AI application is deployed, it is not finished. Models drift. Process conditions evolve. And in a landscape of vendor-hosted foundation models, system behavior can change without any action by the deploying organization. A vendor updating a foundation model can alter outputs in ways that are invisible unless someone with domain expertise is watching. The steward owns this ongoing judgment: is this application still performing as intended, in the current operating context? In regulated environments, this determination may trigger formal change control obligations — and someone must own the decision about whether a vendor model update constitutes a change that requires revalidation. The steward makes that ownership explicit. Under draft EU GMP Annex 22 (European Commission, 2025), any retraining or vendor-initiated model update is treated as a change requiring formal change control and revalidation; data drift requires retraining, revalidation, or use restriction. The steward owns the determination that triggers this process.

The Data Steward Precedent

The AI steward role is not without precedent. Most regulated manufacturers have some form of data steward program, and the organizational insight that made data governance work is directly transferable. DAMA International's Data Management Body of Knowledge codified the principle: data governance cannot be administered effectively from a central team that does not understand what the data means in operational context (DAMA International, 2017). The solution was to embed governance in domain expertise — to create distributed roles that combine process knowledge with data literacy, giving practitioners the skills and authority to govern data within their function.

What translates to AI governance is the structural pattern: distributed judgment, domain context as a prerequisite for effective oversight, and practitioner credibility as the basis for organizational trust. What does *not* translate cleanly is the pace of change. Data governance standards evolve slowly; data stewards can develop expertise that remains current for years. AI technology moves faster than any standards body. Vendor model updates create a dynamic governance challenge that data stewardship does not face. And where data stewards assess deterministic data quality — is this value correct, complete, timely? — AI stewards must assess probabilistic outputs. The judgment required is fundamentally different: not "is this data accurate?" but "is this model's output reliable enough, in this context, for this decision?"

The CRA Analogy

Regulated industries already deploy distributed governance roles that follow this pattern. In clinical research, the Clinical Research Associate operates as a bridge between sponsor oversight and site-level domain expertise (ICH E6(R2)). CRAs do not centralize trial governance — they distribute it to people who understand both the protocol requirements and the local site context. The AI steward follows the same organizational logic in manufacturing: governance embedded in domain expertise, distributed across functional areas, connected through a coordinating body.

GxP Stewardship and Regulated Accountability

The general steward role provides the structural pattern. In GxP environments, that pattern requires an additional specification: the steward in a regulated function must carry explicit GxP accountability, not implicit alignment with quality requirements. Every employee in a GxP environment is, in some sense, a GxP steward — quality is everyone's responsibility. But quality is also no one's responsibility unless someone is named. The regulatory framework demands a specific accountable person who can answer for whether an AI application is fit for use in a regulated decision context, and whose name is on the decision when an inspector asks who approved it.

For AI applications in functions touching GxP work — quality assurance, validation, batch disposition, deviation management, regulatory affairs — the steward role must therefore be paired with formal GxP accountability. Practically, this can be implemented either by selecting stewards who already hold quality-accountable positions, or by formally extending an existing steward's role to include regulatory accountability for AI applications within their scope. The choice depends on organizational structure, but the requirement does not: the AI steward in a GxP function is not a generic AI advocate. The role carries the same standard of accountability — and the same requirement for documentation, training records, and decision rationale — that any other GxP-touching role would.

The implication is operational. Organizations cannot deploy this model in regulated functions without first answering a specific question: *which named person carries GxP accountability for AI applications in this function?* If the answer is "the function head, broadly" or "the quality team, generally," the model has not yet been implemented. The steward role only exists when one person is named, trained, and authorized.

This is not a theoretical position. Purolea Cosmetics Lab had used AI agents to generate drug specifications, procedures, and master production records to "comply with FDA regulations" — but had not subjected those

outputs to quality unit review. The agency's stated remedy crystallizes one element of the GxP steward model — the named, accountable QU representative — in regulatory language: "any output or recommendations from an AI agent must be reviewed and cleared by an authorized human representative of your firm's QU." The named, accountable QU representative is precisely the role this paper describes. The first enforcement signal in this space is not asking organizations to invent something new. It is asking them to make explicit what regulated quality has always required: a named human is accountable for the decision, regardless of what tool informed it.

The European regulatory direction reinforces this expectation. EU GMP Annex 22 in consultation draft (European Commission, 2025) requires that AI models in critical GMP applications be deployed only with defined SME, QA, IT, and data-science roles, with logged feature attribution, decision justification, and confidence scores supporting human review. The steward role this paper proposes is one operational implementation of that requirement: a single named person carrying the QA-aligned accountability that Annex 22 distributes across multidisciplinary roles.

A skeptical reader could interpret the Purolea letter as grounds to prohibit AI use in CGMP work entirely. The agency's actual remedy refutes that reading: it does not bar AI use — it requires that AI outputs be reviewed by a named, authorized human in the QU. The argument is for accountability, not abstinence.

A practical caveat on capacity: the 10–20% time allocation for the steward role is calibrated for monitoring, escalation, and post-deployment review across a function's AI portfolio. For high-stakes scope — heavy AI use in batch disposition, validation, or other primary GxP activities — the allocation may be insufficient, and the steward role should be paired with, or absorbed into, an existing primary quality-accountable role. The principle is non-negotiable; the allocation is calibrated to load.

Collective Intelligence as Design Principle

The steward network is not merely an administrative convenience — it is an organizational design for collective intelligence. Woolley et al. (2010) established that group performance across diverse tasks is predicted not by the maximum intelligence of any individual member, but by social sensitivity, equality of participation, and the quality of information sharing within the group. Concentrating AI capability in a small expert team optimizes for individual expertise at the expense of collective intelligence. Distributing capability broadly — through sandboxes, competency tiers, and steward networks — optimizes for the factors that actually predict group performance.

Malone (2018) extends this finding to human-computer systems. The most powerful cognitive entities, he argues, will be "superminds" — combinations of people and computers thinking together, organized through hierarchies, markets, communities, and democracies. The steward network is connective tissue for a manufacturing supermind: practitioners experimenting locally, sharing discoveries through the center of excellence, stewards bridging local innovation and organizational governance. The AI steward role makes this architecture operational rather than aspirational (Malone, Rus, & Laubacher, 2020).

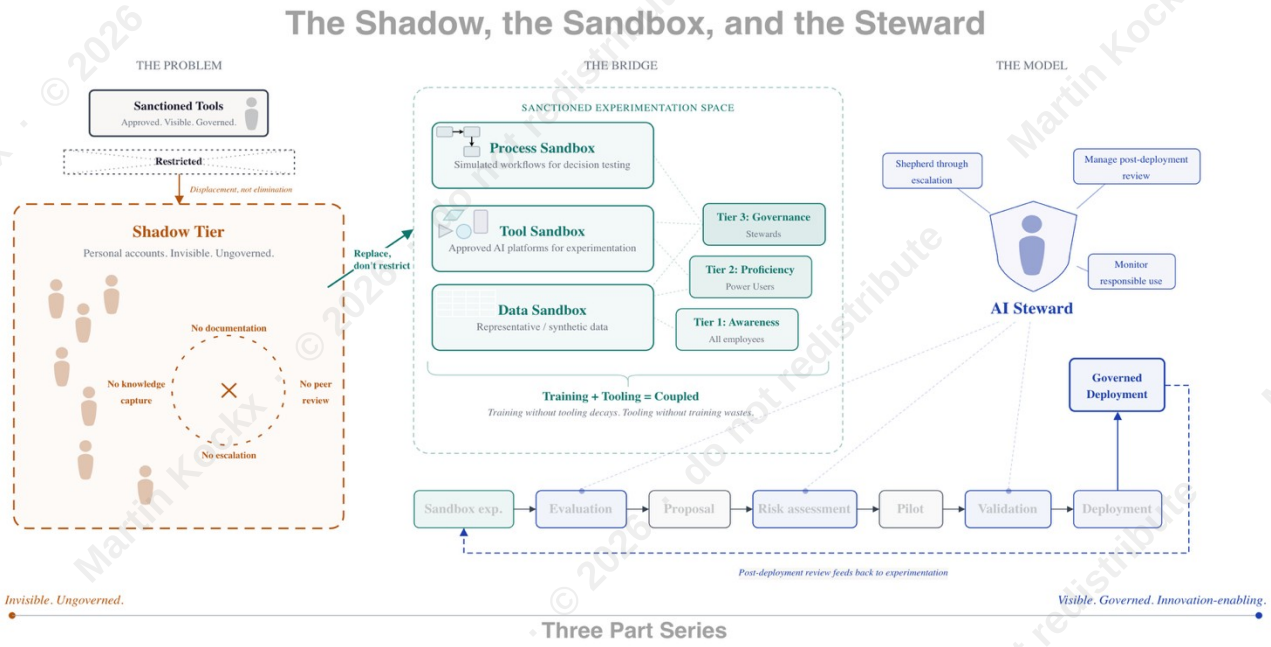

Figure 3. *The integrated framework: from the shadow tier problem (left) through sandboxes and competency tiers (center) to the AI steward and governed deployment (right). The broken learning loop in the problem space is restored through the feedback loop in the governed space.*

Section 6: Governance as Enabler

The word “governance” in most organizational contexts functions as a synonym for restriction. This framing is precisely backward. A well-designed governance framework is what allows an organization to say *yes* to experimentation — because the risks are understood, bounded, and monitored. Without governance infrastructure, the rational organizational response to AI risk is prohibition. With it, the rational response is structured permission.

The sandbox, the competency tiers, and the steward network described in the preceding sections are governance infrastructure — not in the restrictive sense, but in the enabling sense. They create the conditions under which an organization can afford to let its workforce experiment with AI: bounded environments reduce blast radius, trained practitioners reduce misuse, and distributed stewards provide the contextual oversight that central teams cannot. The governance architecture is what converts organizational risk tolerance from a fixed constraint into an expandable capacity.

Risk-Tiered Framework

Not all AI applications carry equal risk, and governance burden should be proportional to consequence. A risk-tiered framework distinguishes at minimum between three categories:

Lower-risk applications — advisory tools, non-regulated process support, documentation assistance — require lighter governance: sandbox experimentation, peer review, steward oversight. These applications can move quickly from experiment to deployment because the blast radius of failure is contained.

Moderate-risk applications — those adjacent to regulated processes, influencing but not directly determining GxP decisions — require deliberate governance: formal risk assessment, documented validation rationale, defined escalation criteria, and post-deployment monitoring.

Higher-risk applications — those directly impacting product quality, batch disposition, or patient safety — require the full governance apparatus: formal validation, regulatory defensibility review, continuous monitoring, and explicit human oversight at every decision point where regulatory defensibility requires it.

The tier determines the governance burden, not a binary approved-or-rejected gate. This distinction matters operationally: organizations that apply maximum governance to minimum-risk applications create the bottlenecks that drive practitioners toward shadow AI. Organizations that match governance to risk create room for the broad experimentation that builds organizational capability.

Across all tiers, a useful design discipline for human oversight is the **Attributable-Observable-Correctable** principle. An AI application is *attributable* when its reasoning chain is traceable — a human can understand why the system reached its conclusion. It is *observable* when its recommendation is visible to a human decision-maker before it is acted upon. And it is *correctable* when a human can override, modify, or reverse the output without heroic effort. These are not binary gates but design criteria: the higher the risk tier, the more rigorously each property must be enforced. Treating attributability, observability, and correctability as continuous design dimensions — rather than yes-or-no compliance checks — gives organizations a practical vocabulary for calibrating human oversight to context. AOC governs the design of *deployed AI applications*; the four sandbox principles introduced in Section 3 (Representative, Isolated, Instrumented, Escalatable) govern the *experimentation environment*. The two sets are complementary, not redundant — different objects at different lifecycle stages. The AOC principle aligns with EU GMP Annex 22's explicit requirements that AI models "log feature attribution, justify decisions based on relevant features, and record confidence scores" within human review processes (European Commission, 2025).

A current regulatory boundary worth naming: under EU GMP Annex 22 as drafted (European Commission, 2025), AI models permitted in critical GMP applications are limited to static, deterministic systems; generative AI, LLMs, and continuously learning models are excluded from critical use and admissible only in non-critical contexts under human oversight. The sandbox-and-steward architecture proposed here applies across both categories, but the higher-risk tier in this section must be read in light of that categorical restriction.

Generation versus Critique — A Usage-Mode Design Principle

Tier and risk classification address *whether* AI may be used in a given context. A second design dimension, frequently overlooked, addresses *how* it is used. The same model, applied to the same problem, carries materially different risk profiles depending on whether it is asked to generate content or to critique content already produced by a human.

In generation mode, the AI produces the substantive output. A hallucinated paragraph, an inverted causal claim, or an invented citation enters the workflow as content the human reviewer must catch. The cost of inattention is direct: the AI's error becomes the organization's record. In critique mode, the AI receives a human-authored draft and is asked to identify weaknesses — gaps in reasoning, missing evidence, inconsistent claims, regulatory considerations the author may have overlooked. A hallucinated critique can be inspected against the underlying draft and rejected. The structural asymmetry matters: generation errors are inherited unless caught; critique errors are rejected unless adopted. The default direction of the human-AI interaction inverts.

For high-stakes, regulated work — deviation investigations, CAPA assessments, validation rationales, change control evaluations — the critique mode is the structurally lower-risk usage pattern. The human retains authorship and reasoning ownership; the AI functions as an opinionated reviewer whose role is to surface what the human may have missed. The discipline this implies is straightforward: stewards should actively encourage critique-

mode usage in regulated functions and treat generation-mode use of AI as a higher-burden case that requires correspondingly stronger AOC controls.

One operational requirement is worth naming explicitly. When AI is used to critique a draft, the critique context window must be isolated from the generation context. If the same model both generated and is now reviewing the output, its critique will be biased toward the framing of its own prior reasoning. This is empirically observable in conversational AI use: ask a model to write something, then ask it to find flaws in what it wrote, and it will under-report problems it would readily identify in another author's text. Separation of the generation and critique contexts — through fresh sessions, distinct prompts, or different model instances — is a practical control that preserves the integrity of the critique signal. Steward training should make this distinction explicit.

This usage-mode design principle does not replace tier or risk classification. It supplements them. Two organizations applying the same risk-tiered framework can produce very different risk profiles based on whether their practitioners are using AI primarily to generate first drafts or primarily to interrogate human-authored ones. The sandbox is the venue where this distinction can be developed into a teachable practice; the steward is the role that ensures it is being applied where it matters most.

Two Operating Speeds

This risk-tiered principle translates into a deployment model with two distinct operating speeds. Non-regulated functions — operations analytics, maintenance planning, internal communications — move fast: experiment broadly, iterate, build organizational AI muscle. These functions serve as proving grounds where the organization learns what works before advancing into regulated territory.

Regulated-adjacent functions move deliberately. AI accelerates the scaffolding of compliance work — generating first drafts of deviation investigations, flagging potential gaps in change control documentation, supporting risk assessment templates — without displacing human judgment at the points where regulatory defensibility requires it. The design constraint is that AI acceleration must preserve attributability, observability, and correctability. Speed without deliberateness does not serve the regulatory goal.

Section 7 addresses the introduction sequence — the third speed governing the path into fully regulated deployment.

Cross-Domain Applicability

The GxP framing is specific to pharmaceutical manufacturing, but the principle — governance as enabler rather than constraint — applies across regulated industries. Energy utilities face analogous challenges under nuclear safety regulations, where the OECD Nuclear Energy Agency's RegLab initiative demonstrates graduated governance for AI deployment (OECD NEA, 2024). Aerospace manufacturers navigate DO-178C certification for software-dependent systems, where EASA has adopted an incremental approach to AI certification across autonomy levels (Frontiers in Aerospace Engineering, 2024; AIAA, 2025). Financial services operate under SR 11-7 model risk management, the most mature risk-based framework for algorithmic oversight in a regulated industry (Federal Reserve & OCC, 2011). In each case, the governance challenge is the same: how to enable innovation within a framework designed primarily for control.

Sandbox Pace as a Design Constraint

One design constraint applies across all of the above: the sandbox must offer tooling close enough to current-state external capability that it can credibly absorb practitioner demand. The shadow tier is not driven solely by access. It is driven by *capability gap*. When practitioners can obtain materially more capable tools through a personal account than through the sanctioned sandbox, the shadow tier will reconstitute itself regardless of policy. Internal AI environments will, structurally, always lag the public frontier — vendor evaluation, security review, validation cycles, and procurement timelines impose a delay that cannot be eliminated. The design objective is therefore not

parity with the frontier but sufficient closeness that the sandbox remains the path of least resistance for legitimate work. A sandbox that ships last quarter's model when last week's is available freely on the open web will not contain shadow use. Organizations operating under this model must commit to a refresh cadence for sandbox tooling that is proportional to the rate of external change, and stewards must be empowered to flag widening gaps as governance risks rather than treating them as procurement preferences.

Section 7: Three-Speed Introduction

Start Where Consequences Are Advisory

The implementation logic follows directly from the governance framework: begin where the blast radius is smallest. AI-assisted analysis of non-regulated data, documentation support, pattern recognition in maintenance logs, energy optimization modeling — these applications build organizational competency without exposing regulated processes to unproven tools.

This is not timidity. It is strategy. Organizations that attempt to deploy AI directly into regulated workflows before building governance maturity and workforce competency reliably fail — not because the technology is inadequate, but because the organizational infrastructure cannot absorb it. Progressive exposure builds the trust, the muscle memory, and the institutional knowledge that make regulated deployment sustainable.

The strategic logic is sequential but not slow. Lower-stakes applications generate three things simultaneously: demonstrated value that justifies continued investment, organizational competency that reduces future deployment risk, and documented evidence — successful experiments, governance interventions, lessons learned — that constitutes the institutional knowledge base for more ambitious deployment. This is intelligence preservation: progressive deployment converts tacit operational knowledge that would otherwise remain locked in individual practitioners into organizational memory that compounds over time. Each advisory application deployed is not merely useful in itself; it is training data for the organization's governance maturity.

The Invisible Influence Problem

A critical reframe is required here. The GxP/non-GxP boundary is not the boundary between safe and unsafe AI use. A scientist using AI to draft a deviation investigation on a personal device is already influencing GxP decisions — the AI output enters the regulated workflow through manual transcription, invisible to any governance framework. The question is not whether AI influences regulated work. It does. The question is whether that influence is visible, governed, and traceable.

This reframe changes the urgency calculus. Organizations that delay AI governance pending "compliant solutions" are not avoiding risk. They are allowing risk to accumulate in the shadow tier, where it is neither visible nor manageable. Standing up governance infrastructure — sandboxes, steward networks, competency programs — in non-regulated functions first is not a conservative approach. It is the fastest path to reducing the actual risk that already exists.

Progressive Exposure

The deployment pattern is generalizable across regulated industries. Start where consequences are advisory. Prove value. Build trust. Advance.

In practice, this means non-regulated functions serve as fast-movers: they experiment broadly, identify what works, and generate documented evidence of value and risk. Regulated-adjacent functions follow as informed adopters, adapting proven approaches with appropriate governance overlays rather than reinventing from scratch. This is the "fast follower" principle applied within the organization: let early adopters in lower-risk functions

demonstrate what works, then adapt for regulated contexts with the governance scaffolding already in place. The knowledge transfer between these speeds is not automatic — it requires the connective infrastructure of steward networks, the center of excellence, and a shared use case library.

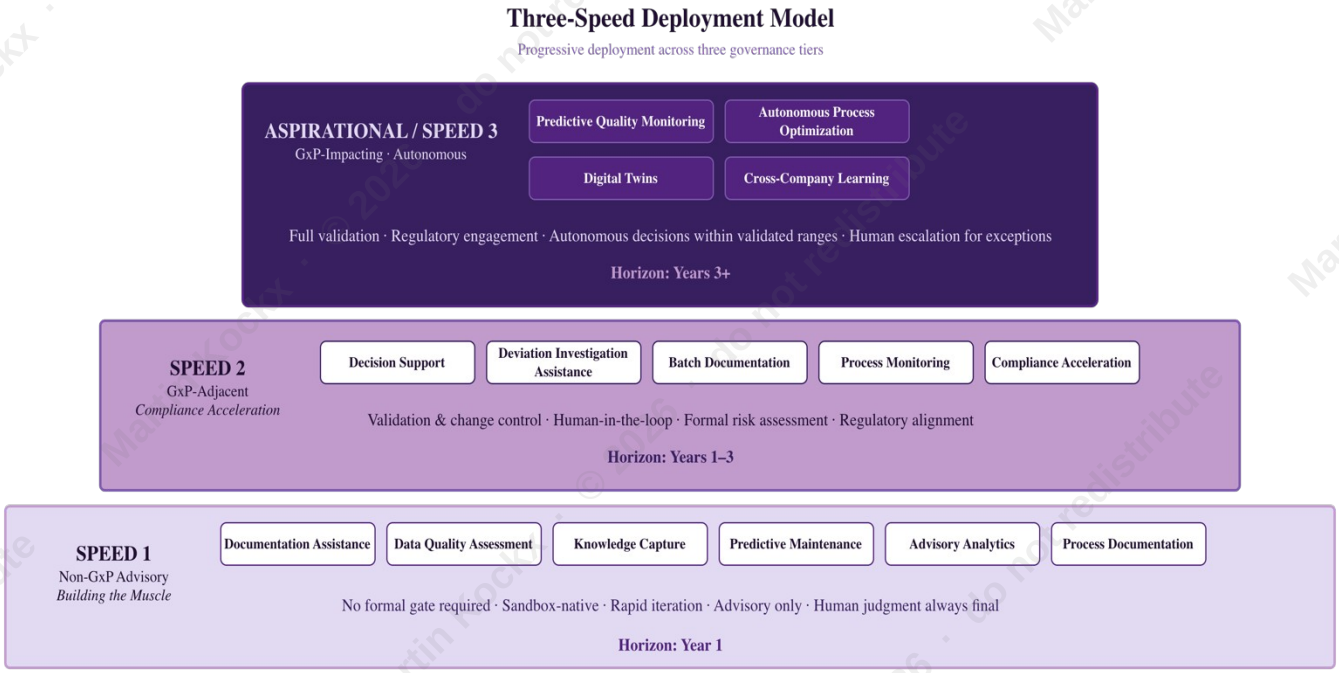

Figure 4. *Three-speed deployment model: organizations begin with non-regulated advisory applications (Speed 1), advance to validated regulated-adjacent deployment (Speed 2), and progress toward full regulated integration (Speed 3) as governance maturity and organizational capability develop.*

Cross-industry evidence supports this graduated approach. EASA's incremental certification pathway for AI in aerospace defines escalating autonomy levels with corresponding governance requirements (AIAA, 2025). The OECD Nuclear Energy Agency's RegLab brings together technologists, operators, and regulators to examine AI deployment from concept to operation through a structured sandbox methodology, with explicit recommendations for competency development alongside technology deployment (OECD NEA, 2024). The EU AI Act mandates regulatory sandboxes across all member states, with annual reporting requirements and priority access for smaller organizations — an institutional recognition that structured experimentation is a prerequisite for responsible deployment, not an alternative to it (EU AI Act, 2024).

The underlying principle is consistent: organizations that build capability progressively, with governance architecture that matures alongside deployment ambition, outperform those that either rush to production or wait for perfect conditions.

Section 8: Discussion — Limitations and Open Questions

The model proposed in this paper — sandbox infrastructure, competency tiers, and AI stewards integrated through a center of excellence — is coherent in principle. Whether it works in practice is an open question. This section engages honestly with the limitations of the proposal and the gaps that separate a logical framework from a validated one.

Untested at Scale

No organization has implemented the full integrated architecture as described in this paper. April 2026's FDA Warning Letter to Purolea Cosmetics Lab establishes one element of this argument — named QU review of AI outputs in CGMP activities — as regulatorily required. But *required* is not the same as *proposed*: the integrated architecture this paper defines, including the steward role with its dual-fluency selection criteria, time allocation, and three specific mandates, remains untested.

Partial implementations exist across pharmaceutical and adjacent regulated industries and provide structural precedent. Several major pharmaceutical organizations operate AI Centers of Excellence with internal sandbox tooling broadly aligned with the principles in this paper — sanctioned experimentation environments, BYO-key access patterns, and varying degrees of distributed support across functions. Anecdotal accounts from practitioners inside these programs suggest the philosophical model is sound but that execution quality varies substantially: some COE implementations function primarily as gatekeepers rather than connectors, others lack the named-steward layer that ties domain accountability to AI use, and few have explicit competency-tier structures coupling training to tool access. Beyond pharma, fintech regulatory sandboxes have demonstrated measurable economic benefits (Cornelli et al., 2020/2024), data steward programs are well-established (DAMA International, 2017), and AI training programs of varying sophistication operate across industries. The components are real and increasingly deployed; the *integration* is what remains novel. The claim that these components reinforce each other — that sandboxes without stewards produce ungoverned experimentation, that stewards without sandboxes have nothing to steward, that training without either produces negligible transfer — is a design argument, not yet an empirical finding. Whether organizations that operate parts of the model achieve measurably different outcomes from those that operate all of it is a question only longitudinal field study can answer. The novelty of this paper is in the integration, not the components. That is both a contribution and a limitation.

Steward Effectiveness Unmeasured

The AI steward role is proposed, not tested. The regulatory expectation that a named human reviews AI outputs in CGMP activities now exists (CDER, 2026); the specific implementation defined here — selection criteria, mandate scope, time allocation — does not yet have empirical evidence behind it. If an organization were to implement this role in 2027, what would constitute evidence that it is working by 2029? Concrete validation criteria should include:

- **Pipeline conversion rate:** The ratio of sandbox experiments that progress to pilot versus those that stall. A functioning steward network should produce a measurably higher conversion rate than an ungoverned sandbox.
- **Time to deployment:** The elapsed time from initial sandbox experiment to production deployment. Steward facilitation should compress this cycle.
- **Shadow AI reduction:** Measurable through IT monitoring of unsanctioned tool usage. If the steward network and sandbox infrastructure are working, shadow AI should decrease as sanctioned channels absorb demand.
- **Issue interception rate:** The number of steward-identified issues that prevented production failures — a leading indicator of governance value that is only visible when someone is watching.
- **Knowledge capture quality:** Documentation completeness, peer review participation, and use case library growth. A steward network that produces no documented learnings is not functioning.

These metrics are specific enough to be measured but remain untested. Whether the steward role produces these outcomes, and at what organizational cost, is unknown.

Resourcing Realism

The model assumes organizations will invest protected time — 10–20% of experienced practitioners' capacity — for the steward role, plus center-of-excellence overhead, plus sandbox infrastructure, plus tiered training across three competency levels. The full architecture is not free, and the resourcing assumption deserves direct scrutiny. Automation and manufacturing technology groups in most pharmaceutical organizations are squeezed as overhead: persistent headcount pressure, contractor freezes, and a structural tendency to fund only what is directly tied to a release date. The model proposed here adds responsibilities to the population that is most frequently asked to absorb more without more.

Two distinct sustainability problems follow.

The first is steward time itself. The 10–20% allocation competes with the most persistent force in manufacturing: production pressure. When a line is down, when a batch is at risk, when delivery commitments are threatened, secondary roles are the first casualty. The steward role is designed to be sustainable at this allocation, but that allocation must survive contact with quarterly production targets, staffing shortages, and the organizational instinct to redeploy experienced people to immediate operational needs. Under what conditions is the time investment sustainable? Likely only when senior leadership treats the steward role as a strategic investment rather than an administrative overhead — when steward contributions are visible in performance reviews, when steward time is protected by the same mechanisms that protect safety training and compliance activities, and when the center of excellence has sufficient organizational authority to advocate for its network. If the steward role can be quietly deprioritized by a functional manager under production pressure, it will be.

The second is the realism of the full architecture itself. A "best in class" framing of sandbox + competency tiers + steward network + center of excellence is aspirational for the majority of pharmaceutical sites. Most organizations will not deploy the model as described. They will deploy a partial version — perhaps a sandbox without a steward network, or a steward role without protected time, or training without tooling. The honest question is whether the model can degrade gracefully. A *minimum viable implementation* might consist of: one steward per functional area at 10% allocation; a single shared sandbox environment with current-state tooling and basic logging; a Tier 1 awareness program coupled to sandbox access; and a small central function that connects rather than gatekeeps. Whether this minimum form retains enough of the integration logic to deliver the claimed benefits — measurably reduced shadow use, governed escalation, defensible regulatory posture — is itself an open empirical question. The risk of the full-architecture framing is that organizations either implement the full model at unsustainable cost or, more commonly, conclude the model is "too much" and revert to the centralized pattern that produced the shadow tier in the first place. Both outcomes are failure modes the design must contend with.

Throughput Fatigue

The Attributable-Observable-Correctable principle (Section 6) requires a reviewer who actually exercises judgment on AI output. The most subtle failure mode of the model is not technical — it is cognitive. When AI compresses a task that previously took two days into one hour, reviewer rigor does not automatically scale to match. Reviewers who once worked through a deviation investigation from first principles are now in the position of reading a finished narrative and validating it. The pattern is familiar from other domains: pilots monitoring autopilot, radiologists overseeing diagnostic AI, drivers in advanced driver-assistance systems. The 99% of cases where the output is correct establishes the trust posture. The 1% where it is not is the one that surfaces in audit.

Throughput fatigue is the predictable consequence. When AI is reliably right, the cost of careful review feels disproportionate to the apparent rarity of error. Engagement decays. The reviewer's signoff becomes a procedural step rather than a substantive judgment. *Correctable* in the AOC sense remains technically available but is rarely exercised. By the time an audit exposes the case where the AI was wrong and the human signed off without engaging, the failure mode is fully established and the documentary trail makes it difficult to reconstruct what review actually consisted of.

The model proposed in this paper does not solve this problem. The sandbox provides the venue where the practice of AI-assisted work is developed; the steward provides the named accountability for that practice within a function. Neither mechanism, by itself, guarantees that the reviewer of a specific output remains substantively engaged on the day of review. Mitigations are design problems for the application itself: forcing engagement through structured prompts that require the reviewer to articulate specific judgments rather than approving holistically, varying the form of AI output to prevent reviewer adaptation to a fixed template, sampling review processes to detect signoff-without-engagement patterns. Organizations adopting this model should plan for throughput fatigue as a foreseeable failure mode rather than a surprise. The steward role provides the place where evidence of fatigue would be detected; the responsibility for designing engagement-preserving workflows lies with the function that deploys the AI application.

Self-Referential Training Risk

A second failure mode is longer in horizon but plausibly more consequential. AI systems deployed in regulated functions will, over time, develop site-specific history: the deviation investigations they have helped draft, the CAPA assessments they have summarized, the trend reports they have generated. When future model outputs are conditioned on this history — through retrieval-augmented generation, fine-tuning on internal records, or prompt context constructed from prior outputs — the system inherits its own past judgments. Errors that escaped review the first time become part of the training context for the next analogous case. The model does not flag this; it has no mechanism to distinguish its own prior conclusions from external ground truth.

The risk compounds in two directions. First, low-quality outputs that entered the record without correction are reinforced as patterns. The model converges toward producing future outputs that resemble past mistakes, because past mistakes are evidence that this is what the organization accepts. Second, the failure becomes invisible: a reviewer comparing an AI output against past organizational records will find consistency, not because the output is correct but because the records were themselves shaped by earlier instances of the same model. The standard of comparison has been silently contaminated by the entity being evaluated.

This is not a hypothetical concern. The pattern is well-established in machine learning research as model collapse — degradation that occurs when generative systems are trained on their own outputs (Shumailov et al., 2024). The pharmaceutical instantiation differs only in operational context: the contaminating training signal arrives through site history rather than recursive synthetic data generation, but the dynamic is the same. The steward role provides one of the few organizational mechanisms positioned to detect this drift, but only if the role explicitly includes auditing the *quality* of AI outputs that enter the regulated record, not merely the *fact* that they were reviewed at the time. Periodic adversarial review — selecting closed AI-assisted records and asking a fresh reviewer to evaluate them independently of the original signoff — is one practical control. Whether organizations will invest in this kind of meta-review at the cadence the risk requires is unknown.

Automation Bias and the Limits of Human Oversight

The two failure modes above share a common substrate: *automation bias*, the well-documented tendency of human reviewers to over-trust automated output and under-engage with their own judgment when an automated system is in the loop (Parasuraman & Manzey, 2010). The phenomenon is not a deficiency of any particular reviewer. It is a stable property of human-automation systems, observed across domains from aviation to medicine to military targeting. The reviewer who is told to validate an AI output is in a fundamentally different cognitive position than the reviewer who is asked to produce the output from scratch. The former is searching for errors in a coherent artifact; the latter is constructing reasoning from inputs. The former task is harder than it appears and produces predictable miss patterns: confirmation rather than scrutiny, holistic approval rather than item-by-item evaluation, and signoff that rests on the AI's apparent confidence rather than on the reviewer's independent analysis.

A pattern increasingly visible in pharmaceutical AI discussions frames human-in-the-loop review as the durable safety mechanism for AI deployment in regulated work. The argument of this paper is that human-in-the-loop is necessary but not sufficient, and that treating it as a primary control reproduces the regulatory failure mode it is meant to prevent. An organization that relies on HITL signoff without engineering for engagement is purchasing a documentary record of review, not the act of review itself. The Purolea Warning Letter illustrates the regulatory consequence: an organization that had AI in the loop and humans nominally in the workflow, but no infrastructure ensuring substantive review, produced compliance failures the agency now treats as enforcement-worthy.

The sandbox-and-steward architecture is a hedge against automation bias, not a guarantee against it. It works to the extent that it places named, domain-expert practitioners in positions where their judgment is exercised on AI applications repeatedly and where their reviews are visible to peers and to governance leadership. It does not work in the limit case where production pressure, throughput fatigue, and routine signoff have eroded the substantive content of human review. The honest framing is that AI governance in regulated environments is a *system design problem* — encompassing the AI application, the workflow it operates in, the reviewer's cognitive ergonomics, and the organizational signals that determine whether substantive review is rewarded or treated as an inefficiency. The model proposed here addresses the organizational layer of that system. The remaining layers — application design, workflow engineering, cognitive ergonomics — are out of scope for this paper but cannot be assumed away. Future work, particularly empirical evaluation of the steward model in operation, must include direct measurement of reviewer engagement, not merely of review event counts.

Regulatory Maturity Prerequisite

The model assumes a baseline of regulatory sophistication that not all organizations possess. Organizations still struggling with fundamental data integrity — 21 CFR Part 11 compliance, ALCOA+ principles, basic electronic record governance — may lack the organizational infrastructure on which AI governance depends. AI governance is not a substitute for data governance; it is an extension of it. An organization that cannot reliably manage electronic signatures is not ready for AI-assisted batch disposition, regardless of the governance framework layered on top.

This creates a maturity prerequisite that the paper does not fully characterize. What is the minimum organizational maturity level required for this model to be viable? At minimum: functioning data governance, established change control processes, a quality management system capable of absorbing novel technology, and leadership willing to invest in capability before demanding returns. Organizations below this threshold need foundational work before they need AI stewards.

Generalizability and Boundary Conditions

The model is presented with pharmaceutical manufacturing as its primary domain, but the structural argument — distributed governance for AI in regulated environments — should apply more broadly. Nuclear energy faces analogous challenges: the OECD NEA RegLab has already adopted a sandbox methodology with multi-stakeholder governance and explicit competency development recommendations (OECD NEA, 2024). Aerospace manufacturers navigate AI certification under DO-178C, where the fundamental challenge is governing non-deterministic systems within deterministic regulatory frameworks (Frontiers in Aerospace Engineering, 2024). Food safety organizations operating under FSMA face AI adoption patterns that parallel pharma, including data privacy concerns, workforce adaptation challenges, and regulatory barriers (Trends in Food Science & Technology, 2025). Financial services operate under SR 11-7, the most mature model risk management framework in any regulated industry, and fintech sandboxes provide the strongest empirical evidence for sandbox effectiveness (Cornelli et al., 2020/2024; Wang et al., 2025).

Where the model likely does *not* apply is in environments where production is fully deterministic and the regulatory framework is prescriptive rather than risk-based. Highly standardized manufacturing with minimal process variability may not generate the diversity of AI use cases that justifies steward infrastructure. The model

is designed for complexity: distributed domain expertise, process variability that creates both risk and opportunity, and regulatory frameworks requiring human judgment rather than algorithmic compliance.

What Success Looks Like

If an organization implemented this model in 2027, measurable outcomes by 2029 should include: a documented reduction in shadow AI usage; a functioning use case library with cross-functional contributions demonstrating an operational sandbox-to-production pipeline; steward-identified governance interventions that prevented downstream failures; measurable improvement in time-to-deployment for applications traversing the escalation pathway; training completion at all three competency tiers with Kirkpatrick Level 3 evidence of behavioral change for the steward cohort (Kirkpatrick & Kirkpatrick, 2006); and cross-functional knowledge sharing that is documented and attributable rather than assumed.

Open Questions for Future Work

Several questions remain beyond the scope of this paper and merit dedicated investigation:

Steward role evolution. As AI capability increases — as models become more reliable, as agentic systems assume more complex workflows — does the steward role become more important or less? If AI systems eventually self-monitor for drift and flag their own confidence degradation, the monitoring mandate may shift from human oversight to human audit of automated oversight. The steward role may evolve from active governance to meta-governance: governing the governance systems rather than the applications directly. Any such evolution toward agentic or self-monitoring AI in regulated functions must contend with EU GMP Annex 22's current exclusion of non-deterministic models from critical GMP applications (European Commission, 2025). Whether that boundary holds, softens, or hardens in revisions after 2026 will shape what the steward role can govern.

The agentic-validation paradox. A more fundamental question sits behind Annex 22's categorical exclusion. Validation as traditionally practiced in regulated manufacturing assumes a system whose behavior is sufficiently deterministic that its outputs can be characterized, qualified, and re-qualified against a fixed acceptance specification. Generative and agentic AI systems do not satisfy this assumption by construction: identical inputs produce non-identical outputs, and the underlying model can be replaced beneath the application by a vendor decision. The conventional validation framework does not have a natural seam for these systems. The regulatory response so far has been to exclude them from critical applications — a categorical solution that is operationally clean but that, as this paper has argued, displaces use rather than eliminating it. The shadow tier will absorb the work that Annex 22 will not let into the sandbox.

The harder design question is whether *continuous oversight* — instrumented use, named accountability, periodic adversarial review, post-deployment monitoring — can substitute for point-validation when point-validation is structurally impossible. Other regulated industries have already faced versions of this question. Biologics manufacturing accepted decades ago that cell lines are not deterministic systems and developed control strategies built around measuring output rather than fixing input — a model of validation that lives with variability rather than against it. The analogy is imperfect (a cell line is not an LLM) but the structural insight is portable: governance regimes can be designed for systems whose individual outputs are not predictable, provided the aggregate behavior is bounded and observable. Whether pharmaceutical AI governance evolves toward this model, holds the current static-deterministic line, or develops a hybrid pathway with explicit categories for differently-validatable AI is one of the most important open questions for the next decade of regulated AI use. The model proposed in this paper is compatible with any of those evolutionary paths — but the steward's mandate, the sandbox's instrumentation requirements, and the post-deployment review cadence will all look different depending on which path the regulatory regime takes.

Minimum viable organization. What is the smallest organizational unit for which this model is viable? A 50-person manufacturing site may not generate sufficient AI adoption to justify a steward network and center of excellence. The overhead-to-value ratio is scale-dependent, and the minimum viable scale is unknown. Smaller organizations may require shared steward resources across sites, or simplified governance models that preserve the distributed judgment principle without the full infrastructure.

Multi-site and partner dynamics. Manufacturing increasingly operates through networks of internal sites and contract manufacturing organizations. How does the steward model extend across organizational boundaries? A steward at a contract manufacturing partner operates under different incentives, different information access, and different organizational authority than an internal steward. Whether the model can function in a multi-organizational context — and what modifications it requires — is an important design question that this paper does not address.

Vendor governance. The post-deployment review mandate assumes stewards can meaningfully monitor vendor-initiated changes. As AI moves toward opaque vendor-hosted services — where model architecture, training data, and update schedules are proprietary — the steward's ability to exercise informed judgment may erode. The model may require contractual mechanisms (change notification requirements, performance guarantee baselines) that extend steward visibility into vendor behavior.

These are not objections to the model. They are the questions that implementation will force, and that future work — empirical, not theoretical — must answer.

References

AIAA. (2025). DO-178 compliance considerations for artificial intelligent software. *AIAA SciTech Forum 2025*. https://doi.org/10.2514/6.2025-2511

American Society for Quality (ASQ). (n.d.). Six sigma belts, levels & roles. *ASQ Quality Resources*. https://asq.org/quality-resources/sixsigma/belts-executives-champions

Baldwin, T. T., & Ford, J. K. (1988). Transfer of training: A review and directions for future research. *Personnel Psychology, 41*(1), 63–105. https://doi.org/10.1111/j.1744-6570.1988.tb00632.x

Blume, B. D., Ford, J. K., Baldwin, T. T., & Huang, J. L. (2010). Transfer of training: A meta-analytic review. *Journal of Management, 36*(4), 1065–1105. https://doi.org/10.1177/0149206309352880

Boston Consulting Group (BCG). (2025). *AI at work 2025: Momentum builds, but gaps remain*. https://www.bcg.com/publications/2025/ai-at-work-momentum-builds-but-gaps-remain

Center for Drug Evaluation and Research (CDER), U.S. Food and Drug Administration. (2026, April 2). *Warning Letter 320-26-58 to Purolea Cosmetics Lab* (MARCS-CMS 722591). https://www.fda.gov/inspections-compliance-enforcement-and-criminal-investigations/warning-letters/purolea-cosmetics-lab-722591-04022026

Cornelli, G., Doerr, S., Gambacorta, L., & Merrouche, O. (2020/2024). Regulatory sandboxes and fintech funding: Evidence from the UK. BIS Working Papers No. 901 (2020); published in *Review of Finance, 28*(1), January 2024. https://doi.org/10.1093/rof/rfad017

CybSafe & National Cybersecurity Alliance. (2024). *Oh, behave! The annual cybersecurity attitudes and behaviors report 2024*. https://www.cybsafe.com/press-releases/study-almost-40-of-workers-share-sensitive-information-with-ai-tools-without-employers-knowledge/

DAMA International. (2017). *DAMA-DMBOK: Data management body of knowledge* (2nd ed.). Technics Publications. https://dama.org/learning-resources/dama-data-management-body-of-knowledge-dmbok/

European Commission. (2025). *EudraLex Volume 4 — Good Manufacturing Practice Guidelines: Annex 22, Artificial Intelligence (consultation draft, 7 July 2025)*. Directorate-General for Health and Food Safety. https://health.ec.europa.eu/document/download/5f38a92d-bb8e-4264-8898-ea076e926db6_en

European Union. (2024). *Regulation (EU) 2024/1689 of the European Parliament and of the Council laying down harmonised rules on artificial intelligence (Artificial Intelligence Act)*. Official Journal of the European Union. https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX:32024R1689

EY Switzerland. (2024). *GxP and AI tools: Compliance, validation and trust in pharma*. https://www.ey.com/en_ch/insights/life-sciences/gxp-and-ai-tools-compliance-validation-and-trust-in-pharma

Federal Reserve System, Board of Governors & Office of the Comptroller of the Currency. (2011). *Supervisory guidance on model risk management (SR 11-7)*. https://www.federalreserve.gov/supervisionreg/srletters/sr1107.htm

Frontiers in Aerospace Engineering. (2024). ML meets aerospace: Challenges of certifying airborne AI. *Frontiers in Aerospace Engineering*. https://www.frontiersin.org/journals/aerospace-engineering/articles/10.3389/fpace.2024.1475139/full

Harmonic Security. (2025). *What 22 million enterprise AI prompts reveal about shadow AI in 2025*. https://www.harmonic.security/resources/what-22-million-enterprise-ai-prompts-reveal-about-shadow-ai-in-2025

International Council for Harmonisation (ICH). (2016). *Guideline for good clinical practice E6(R2)*. https://www.ich.org/page/efficacy-guidelines

Kirkpatrick, D. L., & Kirkpatrick, J. D. (2006). *Evaluating training programs: The four levels* (3rd ed.). Berrett-Koehler Publishers.

Kirkpatrick, J. D., & Kirkpatrick, W. K. (2016). *Kirkpatrick's four levels of training evaluation*. ATD Press.

Malone, T. W. (2018). *Superminds: The surprising power of people and computers thinking together*. Little, Brown and Company.

Malone, T. W., Rus, D., & Laubacher, R. (2020). *Artificial intelligence and the future of work*. MIT Work of the Future Research Brief RB17-2020. https://workofthefuture-taskforce.mit.edu/research-post/artificial-intelligence-and-the-future-of-work/

McKinsey & Company. (2025). *Superagency in the workplace: Empowering people to unlock AI's full potential at work*. https://www.mckinsey.com/capabilities/tech-and-ai/our-insights/superagency-in-the-workplace-empowering-people-to-unlock-ais-full-potential-at-work

Mollick, E. (2024). *Co-intelligence: Living and working with AI*. Portfolio/Penguin.

Nuclear Energy Agency (NEA), OECD. (2024–2025). *International RegLab project reports on AI use in nuclear power plant operations*. https://www.oecd-nea.org/jcms/pl_117030/international-reglab-project-reports-on-ai-use-in-nuclear-power-plant-operations

OECD. (2023). *Regulatory sandboxes in artificial intelligence* (OECD Digital Economy Papers No. 356). https://www.oecd.org/en/publications/regulatory-sandboxes-in-artificial-intelligence_8f80a0e6-en.html

Parasuraman, R., & Manzey, D. H. (2010). Complacency and bias in human use of automation: An attentional integration. *Human Factors, 52*(3), 381–410. https://doi.org/10.1177/0018720810376055

Shumailov, I., Shumaylov, Z., Zhao, Y., Papernot, N., Anderson, R., & Gal, Y. (2024). AI models collapse when trained on recursively generated data. *Nature, 631,* 755–759. https://doi.org/10.1038/s41586-024-07566-y

TELUS Digital. (2025, February 26). *Enterprise employees are entering sensitive data into AI assistants more than you think*. https://www.telusdigital.com/about/newsroom/telus-digital-survey-reveals-enterprise-employees-use-of-shadow-ai

Trends in Food Science & Technology. (2025). Advancing food safety behavior with AI: Innovations and opportunities in the food manufacturing sector. *Trends in Food Science & Technology*. https://www.sciencedirect.com/science/article/pii/S0924224425001864

Von Hippel, E. (1986). Lead users: A source of novel product concepts. *Management Science, 32*(7), 791–805. https://doi.org/10.1287/mnsc.32.7.791

Von Hippel, E. (1988). *The sources of innovation*. Oxford University Press.

Von Hippel, E. (2005). *Democratizing innovation*. MIT Press. https://direct.mit.edu/books/book/2821/Democratizing-Innovation

Von Hippel, E. (2017). *Free innovation*. MIT Press. https://mitpress.mit.edu/9780262035217/free-innovation/

Wang, Y., Zhou, et al. (2025). Effectiveness of regulatory sandboxes in financial services: A systematic review. *Regulation & Governance*. https://doi.org/10.1111/rego.70129

Woolley, A. W., Chabris, C. F., Pentland, A., Hashmi, N., & Malone, T. W. (2010). Evidence for a collective intelligence factor in the performance of human groups. *Science, 330*(6004), 686–688. https://doi.org/10.1126/science.1193147