

Managed Autonomy & Operational Trust

How responsibility is progressively delegated to Digital Workers

Fabio A. Thiers, MD, PhD

ConverGen AI, New York, NY, USA

ABSTRACT

As AI-enabled systems begin performing accountable healthcare work, the question is not whether a system is autonomous but how much responsibility an organization should delegate, for which work, on what evidence. Healthcare already answers this for human workers through graded supervision, credentialing, and evidence-based entrustment. Managed Autonomy applies the same discipline to Digital Labor: progressive delegation to a Digital Worker, task by task, gated on local evidence, and reversible. Authority is granted at the level of the task, because Retrieval, Judgment, and Action fail differently and require different controls. The Digital Operations System (DOPs) makes this repeatable through the Digital Worker Record, which maintains the authority state, and an operating loop of supervision, telemetry, incident review, and change control that determines whether authority should expand, hold, narrow, or be withdrawn.

Keywords: healthcare AI; managed autonomy; progressive delegation; task authority; evidence gate; human-digital collaboration; operational trust; Digital Operations System (DOPs); Digital Workers; Digital Worker Record.

Key points

- **Autonomy is delegated, not possessed.** It is a grant of responsibility made by the organization, not a property the underlying model or agent brings with it.
- **Authority is task-specific.** A Digital Worker may hold a broad role, but responsibility is delegated one task at a time and can differ across tasks in the same workflow.
- **Retrieval, Judgment, and Action require different controls.** They fail differently, so each carries its own control set and its own evidence gate.
- **Delegation changes the human-digital composition.** Machine independence and human control are separate dimensions, and the effectiveness of their combination should be demonstrated rather than assumed.
- **DOPs operationalizes progressive delegation.** Authority state, evidence, and revocation criteria live in the Digital Worker Record and move through an operating loop rather than a governance review.
- **Managing automation as work is more repeatable than managing it as software.** Each new Digital Worker enters an established management system instead of requiring a new *ad hoc* control structure.

Autonomy is delegated, not possessed

Reliable healthcare AI depends on the managed work system, not the model alone. The resulting management question is not whether a Digital Worker is autonomous, but which tasks it may perform independently, under what conditions, and on what evidence. Binary labels such as autonomous versus supervised, or agent versus copilot, obscure that question. Human-factors research has long treated automation as graded and function-specific, spanning information acquisition, analysis, decision selection, and action implementation, each of which can be automated to a different degree within the same system.¹ More recent human-centered AI frameworks go further, separating machine automation from human control rather than treating them as opposite ends of one continuum.²

In institutions, responsibility is delegated rather than possessed. A new employee does not receive full signing authority on the first

day because the person is intelligent. A resident physician, pharmacist, claims manager, or revenue-cycle specialist works within a scope shaped by training, credentials, policy, supervision, and accumulated evidence. Research on trust in automation supports the same discipline: reliance should be calibrated to demonstrated performance under specific conditions, and both overreliance and underreliance are design failures with operational cost.^{3,4}

Therefore, the key question is not whether a Digital Worker is autonomous. It is which tasks it may perform, under what conditions, on what evidence, and subject to what reversal mechanism. Managed Autonomy is the approach applied to continuously answer those questions after initial deployment.

Healthcare organizations already know how to progressively delegate responsibility

Healthcare organizations progressively delegate responsibility during graduate medical education, advancing trainees through graded supervision. Today's formalization of the *entrustable professional activity* defines a unit of work that may be given to a trainee only when evidence supports execution at a defined level of supervision.^{5,6} Entrustment decisions are made for specific activities, at specific supervision levels, on the basis of observed performance rather than seniority or general ability.^{7,8} Credentialing and privileging extend the same logic across a career, making authority specific to procedures and settings, supported by documented performance, and revocable.⁹

This analogy must be handled carefully. A Digital Worker is not a person, not a licensed professional, not an employee, and not legally equivalent to human labor. What transfers here is the institutional mechanism, not the professional status. Granting bounded, evidence-based, revocable authority for a defined activity is ordinary workforce management applied to a new kind of worker, and the progression is recognizable on both sides. A trainee moves from observation to supervised practice, to greater responsibility with escalation, to independent work within privileges, and back to reassessment when performance changes. A Digital Worker moves from shadow mode to supervised execution, to exception-based autonomy, to bounded autonomy, and back to narrowing when evidence changes.

Authority is granted task by task

Delegation decisions go wrong when they are made at the level of the workflow name. A Digital Worker holds a role, and that role contains tasks. A prior authorization role may include record retrieval, documentation assembly, payer-policy comparison, medical-necessity drafting, portal submission, status monitoring, and exception escalation. Some tasks may tolerate high authority. Others require human approval. Others should be prohibited for Digital Labor altogether. The delegation decision is determined by the task, not the role.

Tasks fall onto three surfaces, and the distinction is operationally decisive because the three fail in different ways. **Retrieval** fails when the wrong, stale, incomplete, or low-authority information is used. **Judgment** fails when governing criteria are misapplied, ambiguity is not recognized, or conclusions are insufficiently supported. **Action** fails when an unauthorized, incorrect, or poorly bounded change is made to external state.

Each surface therefore carries a different control set. **Retrieval** is controlled through source hierarchy, freshness gates, least-privilege read access, and provenance capture. **Judgment** is controlled through a reviewable basis, exposed uncertainty, supervisor-agreement thresholds, and escalation criteria for ambiguity. **Action** is controlled through action gates, volume caps and rate limits, rollback, and a tested kill switch. Because the controls differ, the evidence justifying expansion also differs, which is why the surfaces must be gated separately. Regulatory practice points the same way: the FDA's clinical decision support framework turns on the function performed and its intended use, including whether a professional can independently review the basis for a recommendation.¹⁰

The canonical formulation is therefore simple. A Digital Worker may hold a broad organizational role, but authority is granted task by task, and Retrieval, Judgment, and Action require different evidence and different controls. The same worker may be highly autonomous in Retrieval, supervised in Judgment, and tightly constrained in Action (Fig. 1).

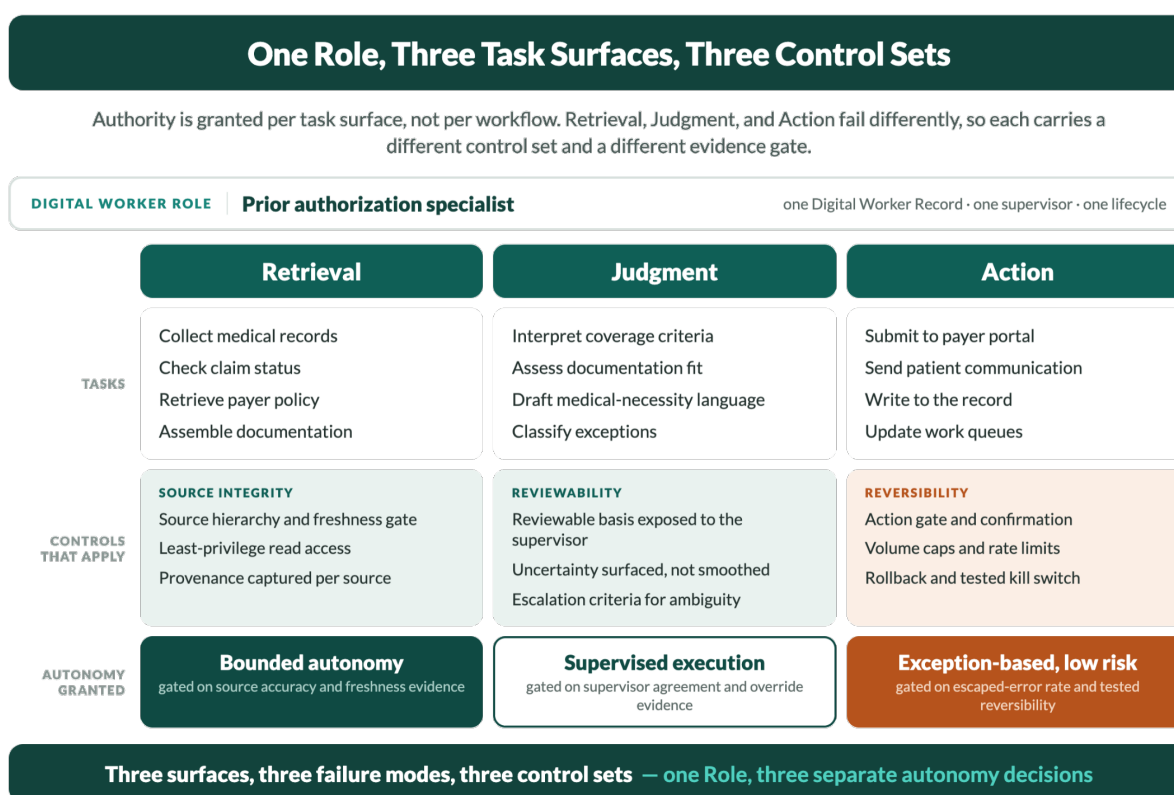


FIGURE 1 - Task surfaces within a single Digital Worker role. Retrieval fails on wrong or stale information, Judgment on misapplied criteria, and Action on unauthorized or poorly bounded change. Each surface carries its own controls and evidence gate, so one worker may hold a different autonomy band on each.

The autonomy ladder is progressive delegation

Delegation should be staged, and each stage has a recognizable institutional counterpart. In **shadow mode**, comparable to observation and parallel practice, the worker classifies, drafts, or recommends without changing operational state; the objective is to characterize errors, missing context, exception patterns, review burden, and workflow fit before anything depends on it. In **supervised execution**, the worker operates in production, but a human approves every material action; this tests whether the worker can hold production rhythm and whether people can review its output meaningfully rather than perfunctorily. In **exception-based autonomy**, comparable to entrusted routine work with escalation, the worker independently completes predefined low-risk cases and escalates ambiguity or elevated

consequence. In **bounded autonomy**, comparable to independent work within defined privileges, the worker acts inside an explicit envelope of tasks, risk classes, context-completeness thresholds, and volume, subject to monitoring, rollback, and revocation (Fig. 2).

Two properties of the ladder matter more than its rungs. First, movement is bidirectional: authority should expand, hold, narrow, pause, or be revoked as evidence changes. An organization that cannot revoke authority cleanly has not delegated it responsibly. Second, the top of the ladder is not the goal. Some tasks may remain permanently in supervised execution because the work is clinically sensitive, legally constrained, or genuinely dependent on judgment. That is not failure. The objective is reliable operations at acceptable risk, not maximum independence.

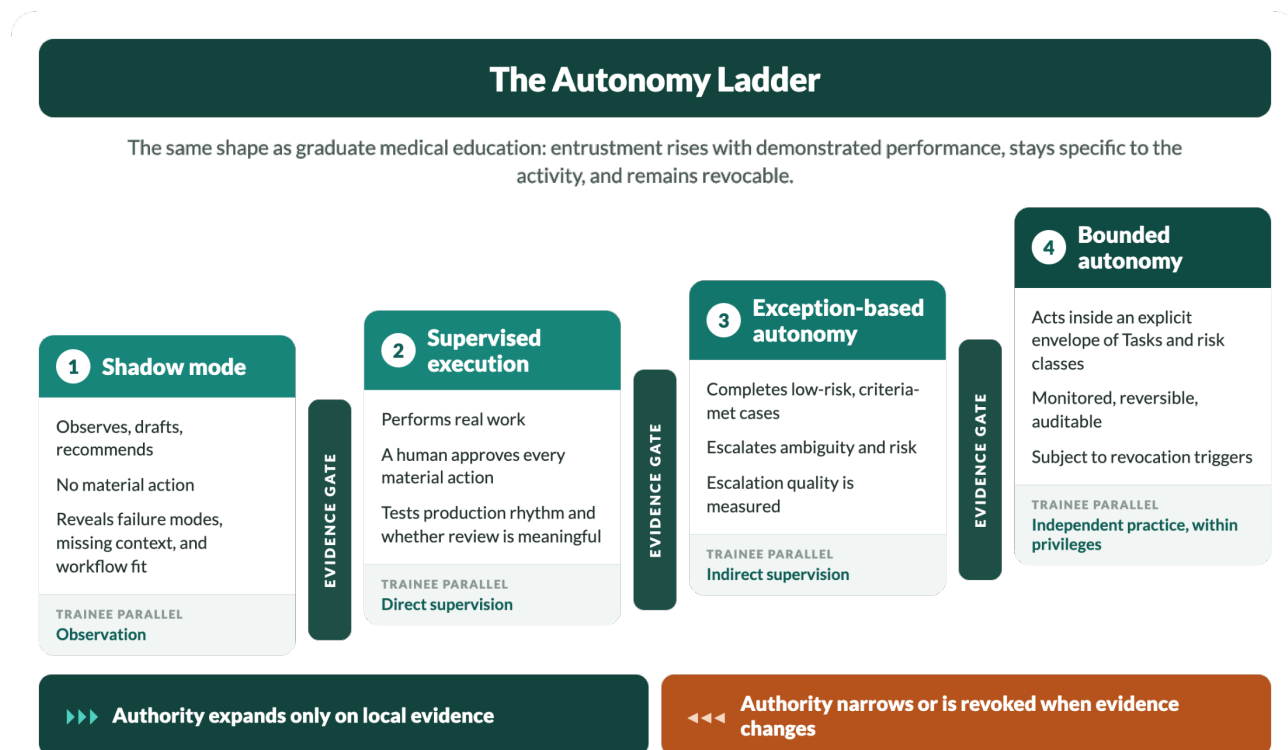


FIGURE 2 - The autonomy ladder and its institutional precedent. Authority progresses from shadow mode to supervised execution, exception-based autonomy, and bounded autonomy, one task surface at a time. Movement is bidirectional, and the top is not the goal.

Evidence gates drives automation changes

Evidence is the currency of delegation. A Digital Worker should not receive greater responsibility because it is impressive in demonstration, because a vendor claims high accuracy, or because a general benchmark result is strong. Recent healthcare-agent benchmarks report strong subtask competence alongside weak end-to-end completion, which is precisely the gap between capability and dependable work.^{11,12} Model performance also fails to transfer automatically across sites and over time: external validations of widely deployed models have found substantial local shortfalls, and recurring local validation is increasingly recommended over one-time approval.^{13,14} Drift and environmental change — payer policies, portals, EHR builds, staffing, populations — erode yesterday's evidence.^{15,16} Frameworks for early-stage evaluation of AI-driven decision support reach the same conclusion: assessment must be staged, setting-specific, and attentive to human factors.¹⁷

Authority should therefore depend on local evidence: performance in the organization's own workflow, on its own case mix, with its own sources and supervisors. The gate for each transition should be defined before the transition is requested; it should be specific, and it should be evaluated per task surface rather than for the role as a whole.

An evidence gate is not an abstract assessment and should not create a separate management object. It is recorded against the Digital Worker Record, which already holds role definition, operational identity, authority, supervision, context, evidence, incidents, performance, and lifecycle. For each task, the record

includes the current authority band and task type, the accountable supervisor, the applicable evidence threshold, and the performance history behind it. Keeping this within the record avoids a second ledger that competes with the object the organization already maintains.

Managed Autonomy is a process inside DOpS

Progressive delegation is not a separate governance layer added on top of Digital Labor. It runs on the capabilities a Digital Operations System (DOPs) already provides, and it can be described as an operating loop (Fig. 3). The **Digital Worker Record** holds role, task scope, authority state, supervision, evidence thresholds, and revocation criteria. **Context, identity, access, and execution control** establish the operating envelope. **Supervision, Telemetry and Evidence** generate the evidence needed to evaluate performance. **Incident and change review** determine whether existing evidence remains valid. **Autonomy management** then expands, holds, narrows, pauses, or revokes authority, while the **Learning** layer incorporates validated corrections into future work.

These capabilities are interconnected rather than sequential steps in a project plan. Evidence influences whether authority should change; authority determines the supervision required; supervision produces corrections; corrections update context; updated context changes what the evidence means. Managed Autonomy is one operating loop inside that system, and its distinguishing feature is that every pass writes back to the same record. The same evidence system that supports expansion must also support narrowing or revocation.

Managed Autonomy Is an Operating Loop Inside DOPs

Progressive delegation is not a separate governance layer. It runs on the Digital Operations System capabilities the organization already operates, and every pass writes back to the Digital Worker Record.



FIGURE 3 - Managed Autonomy as an operating loop within DOPs. Authority is set, work is bounded, humans supervise, evidence accumulates, change is assessed, and authority is adjusted before the loop runs again. Each step is an existing DOPs capability that writes to the Digital Worker Record.

Managed Autonomy changes the division of labor

A Human-Digital Organization composes work deliberately across human and Digital Workers, specifying which work is human-led, digital-led, shared, supervised, or prohibited. Managed Autonomy is the mechanism through which that composition changes over time. The two questions should not be collapsed: composition is the allocation of work, asking who should perform the task, while Managed Autonomy is the allocation of authority within it.

Two findings should discipline how that division is set. The first is that capability is not the same as being a usable teammate. Research on joint human-agent activity identifies the properties automation must have to participate in shared work: it must be predictable, directable so that a person can redirect or halt it mid-activity, and able to signal status, intent, and its own limits.¹⁸ A Digital Worker that cannot be interrupted partway through a task, or whose current state a supervisor cannot read, is not a usable teammate regardless of how well it performs in isolation. Those properties belong on the evidence gate alongside accuracy.

The second is that combining a human and a Digital Worker does not reliably beat either one working alone. A systematic review and meta-analysis of human-AI combinations found that they frequently underperform the better of the two on their own, with the outcome depending heavily on how the work was divided and on the baseline strength of each party.¹⁹ Composition is therefore an empirical question. The right division of labor for a given task must be measured in the organization's own setting rather than assumed from the premise that human review can only add value.

The distinction becomes concrete in a single workflow. At deployment, the Digital Worker retrieves records while a person interprets criteria and submits. Months later, retrieval runs autonomously, the worker drafts the medical-necessity judgment for approval, and low-risk submissions clear under exception-based authority. Some tasks may move further toward digital-led execution while others remain permanently human-led because consequence, ambiguity, or professional accountability requires it. The division of labor is not a static allocation made at design time; it is a managed position that moves with evidence.

Managing Digital Workers is more scalable than managing automation as software

The practical case for Managed Autonomy is not only safety. It is repeatability. Under the traditional model, the workflow or script is the management object. Authority is embedded in configuration. Each automation tends to carry its own permissions, exception logic, monitoring, maintenance, and ownership. Human oversight is added locally, lifecycle management is technical, and changes are handled workflow by workflow. Scaling produces a growing collection of automation projects, each with its own control structure.

Under the Digital Worker model, the persistent role is the management object. Tasks and authority are explicitly defined, the supervisor is named, authority is evidence-gated, identity and access are role-based, supervision patterns are reusable, incidents attach to the worker record, and operational history survives changes of model, prompt, or vendor. The same DOPs mechanisms can be applied to the next role. Managed Automation scales because the organization reuses a common management system rather than designing a new control structure for every automation.

The trainee analogy is more than a safety argument here; it is the scalability argument. Healthcare organizations do not build a new supervision architecture for every resident or nurse. They reuse role definitions, credentialing, supervision structures, escalating privileges, documentation expectations, incident processes, evaluation standards, and remediation mechanisms. DOPs brings the same repeatable institutional structure to Digital Labor. This is consistent with the broader economic finding that information technology produces durable gains only when organizations redesign work, decision rights, and management practice around it, and that the returns arrive once those complementary organizational changes are in place.^{20,21}

Oversight must be designed as meaningful work

Placing a human in the loop is not, by itself, a control. Automation can leave human supervisors with a smaller but more difficult residue of exceptions, weakening oversight if review capacity and expertise are not redesigned accordingly.²² Automation-bias research shows that reviewers over-rely on automated recommendations when verification is difficult, workload is high, or the system appears authoritative.^{23–25} A reviewer who sees a polished recommendation without sources, uncertainty, or consequences, and who has no efficient way to correct the system, supplies the appearance of oversight without its substance. Interaction-design guidance for AI-enabled systems makes the same operational point: systems should make clear what they can do, why they did what they did, and how a person can efficiently correct or dismiss the output.²⁶ Early deployments of AI-drafted patient messages illustrate the consequence: value depended on editing burden, workflow fit, and clear accountability for what was sent, not on fluency.²⁷

Oversight quality should therefore be designed and measured like any other operational process. The supervisory interface should show what the worker did, what evidence it used, what it could not verify, why it is uncertain, what action it proposes, and what will happen if approved. Review burden should be tracked, because an approval queue that exceeds human attention converts supervision into rubber-stamping. Overrides and corrections should be captured as structured evidence, both to improve the worker and to test whether the humans in the system remain a strong control rather than a weak one.

Progressive delegation also changes human work rather than removing it. As tasks move up the ladder, staff time shifts from routine execution toward supervision, exception handling, context stewardship, and improvement. Those responsibilities need defined ownership and protected capacity; treating them as informal additions to already strained roles is the most common way a delegation plan fails in operation.

Delegation must respect blast radius

A human worker's error is naturally limited by caseload, working hours, and manual pace. A Digital Worker can repeat a flawed action quickly across many cases. A stale source, prompt defect, permission error, or interface change can create correlated failure: many instances of the same escaped error before anyone notices. Delegation decisions should therefore weigh not only the severity of a single error but the number of cases that could be affected before detection. Layered defenses are the classic response to systemic failure, and they apply directly here.²⁸

The practical controls, such as volume caps, rate limits, phased rollout, site-by-site expansion, version pinning, kill switches, and monitoring alerts tied to automatic narrowing of authority, are not generic governance recommendations. They are configurations of the DOpS execution-control and autonomy-management layers, attached to a named role and a named supervisor, and testable. The emerging literature on agentic systems converges on the same requirements: agent identifiers, real-time monitoring, activity logging, attributability, and interruptibility are proposed as baseline practices for systems that act with limited supervision.^{29,30} Digital Operations gives those abstractions an institutional home: identity lives in the Digital Worker Record,

monitoring in its telemetry, and interruptibility in revocation mechanics that are tested rather than assumed.

Operational trust is the key DOpS output

Operational trust is often discussed as if it were a feeling about a technology. In organizations, trust is a system of evidence. A hospital trusts its medication process because formularies, order checks, pharmacist review, administration records, and adverse-event processes surround it. Passengers trust aviation because maintenance logs, certification, air-traffic control, and incident investigation surround flight. Digital Labor will be trusted the same way: not because stakeholders inspect every log, but because behavior is bounded, failures are discoverable, responsibility is assigned, and lessons are retained.

The practical requirement is an evidence chain that can answer a fixed set of questions after any material act: which Digital Worker acted, under which identity; which task was performed and whether it fell within granted authority; what context was used, from which sources, at what freshness; what reviewable basis supported the conclusion; what was done and what downstream systems recorded; which human saw what and decided what; which model, prompt, workflow, and policy versions were active; what the outcome was; and what changed after review. Management-system frameworks for AI point toward the same architecture of documented risk management, monitoring, and controlled change.^{31–33} DOpS is the infrastructure that continuously produces and maintains that evidence, written to the Digital Worker Record of a role that has a supervisor and a current authority band.

Stakeholders need different views of the record

Trust has audiences, and they should not all see the same data. A supervisor may need full-fidelity logs and case replay. A compliance officer needs scope, access, and approval evidence. An executive needs portfolio answers: which Digital Workers hold which authority, on what evidence, and how incidents changed those decisions. Clinicians need assurance that clinical-adjacent tasks remain under professional control, with a reviewable basis for anything approaching judgment.¹⁰ Regulators need evidence of lawful scope, human accountability, privacy safeguards, and nondiscrimination, including whether a worker performs differently across patient populations.³⁴ Insurers need exposure that can be described, bounded, and reconstructed, typically as aggregated signals rather than raw operational detail.^{35,36}

Evidence architecture is therefore also access architecture. The organization should be able to produce each view from one record without creating broad secondary exposure of patient data in the name of accountability.

The executive test before expanding authority

Before approving any expansion of Digital Worker authority, an executive team should be able to answer a short set of questions with evidence rather than assurances. The test is deliberately narrow. It is not a judgment about whether an AI system deserves autonomy. It asks whether a specific task, on a specific surface, has earned a change in what is allowed to act (Fig. 4).

The Executive Test Before Expanding Authority

Seven questions answered from the Digital Worker Record, not from assurances — not whether the work can be done reliably, but whether the operation has earned a change in what is allowed to act.

- 1 **TASK SURFACE** **Exactly which Tasks would gain authority?**
EVIDENCE The Role decomposed into Retrieval, Judgment, and Action Tasks, with the change specified per surface
- 2 **EVIDENCE GATE** **What local evidence supports this level?**
EVIDENCE Escaped-error rate, near misses, override patterns, escalation quality, and drift within the envelope
- 3 **SUPERVISION** **What does human review actually catch?**
EVIDENCE Reviewer agreement and override data, review burden, interface adequacy, and correction incorporation
- 4 **BLAST RADIUS** **How many cases could one defect touch before detection?**
EVIDENCE Volume caps, rate limits, phased rollout, monitoring thresholds, and tested kill-switch and rollback
- 5 **REVERSIBILITY** **Can authority be narrowed cleanly?**
EVIDENCE A tested revocation mechanism, defined narrowing triggers, and a work-continuity plan
- 6 **TEAM FIT** **Is the worker directable and legible to its supervisor?**
EVIDENCE Demonstrated mid-task interruption, readable status and intent, and measured pairing performance
- 7 **RECORD** **Where is this decision written down?**
EVIDENCE An autonomy entry in the Digital Worker Record: Tasks, band, evidence, approver, and review date

✓ **Authority may expand only when every row is answered from the record**

FIGURE 4 - The executive test before expanding authority. Seven questions that should be answerable from the Digital Worker Record rather than from the memory of the implementation team. Authority is granted to tasks on a surface, never to a workflow as a whole.

Close: responsibility should move deliberately

Human-Digital Organizations will not scale by deciding whether individual AI systems are autonomous. They will scale by progressively delegating defined tasks to Digital Workers through a common management system. Some tasks will become highly autonomous. Others will remain supervised or human-led. What matters is that authority follows demonstrated performance, remains bounded by role and task, and can expand, hold, narrow, pause, or be revoked as evidence changes.

DOPs makes that process repeatable. It preserves the Digital Worker Record, enforces task-specific authority, supports supervision, captures evidence, manages incidents and change, and adjusts authority over time. The result is not simply safer automation. It is a more scalable operating model for Digital Labor, because every new Digital Worker enters an established system for supervision, evidence, authority, learning, and lifecycle management rather than requiring one to be invented.

The practical objective is therefore not maximum autonomy, but an organization in which responsibility can move deliberately between humans and Digital Workers without losing accountability.

Glossary of key concepts

- **Managed Autonomy.** The evidence-based process through which authority for specific Digital Worker tasks is granted, maintained, expanded, narrowed, paused, or revoked.
- **Progressive delegation.** The staged transfer of responsibility from human supervision toward greater Digital Worker authority as local performance evidence accumulates.
- **Task authority.** The permission granted to a Digital Worker to perform a defined task under specified conditions.
- **Retrieval.** A task surface concerned with obtaining the correct information from the correct source at the correct time.

- **Judgment.** A task surface requiring interpretation of facts under an applicable clinical, administrative, contractual, legal, or organizational standard.
- **Action.** A task surface that changes external or organizational state.
- **Evidence gate.** The predefined body of local operational evidence required before authority for a task may change.
- **Human-digital composition.** The allocation of a task across human and Digital Workers as human-led, digital-led, shared, supervised, or prohibited.
- **Oversight quality.** The degree to which human supervision is informed, effective, adequately resourced, and able to correct or stop Digital Worker activity.
- **Blast radius.** The number of cases, systems, or downstream processes that one defect could affect before detection.

Live Article information

This PDF is a dated snapshot of a continuously maintained Live Article. Material updates undergo editorial review, and substantive public revisions are released when warranted.

The current maintained article and evidence base are available through the ConverGen Digital Operations Advisor: www.convergen.ai/advisor/

Peer-review status: This public edition has not undergone external peer review.

Competing interests: The author is the founder and CEO of ConverGen AI, which develops and commercializes a Digital Operations System (DOPs) for healthcare organizations.

AI-assisted tools: AI-assisted tools supported research synthesis, drafting, editing, and figure production. The author reviewed and takes responsibility for all content.

Author ORCID: 0009-0008-1814-1983.

References

- Parasuraman, R., Sheridan, T. B. & Wickens, C. D. A model for types and levels of human interaction with automation. *IEEE Trans. Syst. Man Cybern. A* 30, 286–297 (2000). doi:10.1109/3468.844354.
- Shneiderman, B. Human-centered artificial intelligence: reliable, safe & trustworthy. *Int. J. Hum. Comput. Interact.* 36, 495–504 (2020). doi:10.1080/10447318.2020.1741118.
- Lee, J. D. & See, K. A. Trust in automation: designing for appropriate reliance. *Hum. Factors* 46, 50–80 (2004). doi:10.1518/hfes.46.1.50_30392.
- Parasuraman, R. & Riley, V. Humans and automation: use, misuse, disuse, abuse. *Hum. Factors* 39, 230–253 (1997). doi:10.1518/001872097778543886.
- ten Cate, O. Entrustability of professional activities and competency-based training. *Med. Educ.* 39, 1176–1177 (2005). doi:10.1111/j.1365-2929.2005.02341.x.
- ten Cate, O. Nuts and bolts of entrustable professional activities. *J. Grad. Med. Educ.* 5, 157–158 (2013). doi:10.4300/JGME-D-12-00380.1.
- ten Cate, O. et al. Curriculum development for the workplace using entrustable professional activities (EPAs): AMEE Guide No. 99. *Med. Teach.* 37, 983–1002 (2015). doi:10.3109/0142159X.2015.1060308.
- Kennedy, T. J. T., Regehr, G., Baker, G. R. & Lingard, L. Point-of-care assessment of medical trainee competence for independent clinical work. *Acad. Med.* 83, S89–S92 (2008). doi:10.1097/ACM.0b013e318183c8b7.
- Centers for Medicare & Medicaid Services. Conditions of Participation: Medical Staff. 42 CFR §482.22 (2024). <https://www.ecfr.gov/current/title-42/chapter-IV/subchapter-G/part-482/subpart-C/section-482.22>. Accessed 10 Aug 2026.
- U.S. Food and Drug Administration. Clinical Decision Support Software: Guidance for Industry and Food and Drug Administration Staff (2026). <https://www.fda.gov/regulatory-information/search-fda-guidance-documents/clinical-decision-support-software>. Accessed 10 Aug 2026.
- Bedi, S. et al. HealthAdminBench: evaluating computer-use agents on healthcare administration tasks. Preprint at arXiv:2604.09937 (2026). doi:10.48550/arXiv.2604.09937.
- Liu, R. et al. PhysicianBench: evaluating LLM agents in real-world EHR environments. Preprint at arXiv:2605.02240 (2026). doi:10.48550/arXiv.2605.02240.
- Youssef, A. et al. External validation of AI models in health should be replaced with recurring local validation. *Nat. Med.* 29, 2686–2687 (2023). doi:10.1038/s41591-023-02540-z.
- Wong, A. et al. External validation of a widely implemented proprietary sepsis prediction model in hospitalized patients. *JAMA Intern. Med.* 181, 1065–1070 (2021). doi:10.1001/jamainternmed.2021.2626.
- Sahiner, B., Chen, W., Samala, R. K. & Petrick, N. Data drift in medical machine learning: implications and potential remedies. *Br. J. Radiol.* 96, 20220878 (2023). doi:10.1259/bjr.20220878.
- Kore, A. et al. Empirical data drift detection experiments on real-world medical imaging data. *Nat. Commun.* 15, 1887 (2024). doi:10.1038/s41467-024-46142-w.
- Vasey, B. et al. Reporting guideline for the early-stage clinical evaluation of decision support systems driven by artificial intelligence: DECIDE-AI. *Nat. Med.* 28, 924–933 (2022). doi:10.1038/s41591-022-01772-9.
- Klein, G., Woods, D. D., Bradshaw, J. M., Hoffman, R. R. & Feltoovich, P. J. Ten challenges for making automation a “team player” in joint human-agent activity. *IEEE Intell. Syst.* 19, 91–95 (2004). doi:10.1109/MIS.2004.74.
- Vaccaro, M., Almaatouq, A. & Malone, T. When combinations of humans and AI are useful: a systematic review and meta-analysis. *Nat. Hum. Behav.* 8, 2293–2303 (2024). doi:10.1038/s41562-024-02024-1.
- Brynjolfsson, E. & Hitt, L. M. Beyond computation: information technology, organizational transformation and business performance. *J. Econ. Perspect.* 14, 23–48 (2000). doi:10.1257/jep.14.4.23.
- Brynjolfsson, E., Rock, D. & Syverson, C. The productivity J-curve: how intangibles complement general purpose technologies. *Am. Econ. J. Macroecon.* 13, 333–372 (2021). doi:10.1257/mac.20180386.
- Bainbridge, L. Ironies of automation. *Automatica* 19, 775–779 (1983). doi:10.1016/0005-1098(83)90046-8.
- Goddard, K., Roudsari, A. & Wyatt, J. C. Automation bias: a systematic review of frequency, effect mediators, and mitigators. *J. Am. Med. Inform. Assoc.* 19, 121–127 (2012). doi:10.1136/amiajnl-2011-000089.
- Lyell, D. & Coiera, E. Automation bias and verification complexity: a systematic review. *J. Am. Med. Inform. Assoc.* 24, 423–431 (2017). doi:10.1093/jamia/ocw105.
- Khera, R., Simon, M. A. & Ross, J. S. Automation bias and assistive AI: risk of harm from AI-driven clinical decision support. *JAMA* 330, 2255–2257 (2023). doi:10.1001/jama.2023.22557.
- Amershi, S. et al. Guidelines for human-AI interaction. *Proc. 2019 CHI Conf. Hum. Factors Comput. Syst.* 1–13 (2019). doi:10.1145/3290605.3300233.
- Garcia, P. et al. Artificial intelligence-generated draft replies to patient inbox messages. *JAMA Netw. Open* 7, e243201 (2024). doi:10.1001/jamanetworkopen.2024.3201.
- Reason, J. Human error: models and management. *BMJ* 320, 768–770 (2000). doi:10.1136/bmj.320.7237.768.
- Shavit, Y. et al. Practices for Governing Agentic AI Systems. OpenAI (2023). <https://openai.com/index/practices-for-governing-agentic-ai-systems/>. Accessed 10 Aug 2026.
- Chan, A. et al. Visibility into AI agents. *Proc. Conf. Fairness Accountability Transparency* 958–973 (2024). doi:10.1145/3630106.3658948.
- National Institute of Standards and Technology. Artificial Intelligence Risk Management Framework (AI RMF 1.0). NIST AI 100-1 (2023). doi:10.6028/NIST.AI.100-1.
- International Organization for Standardization. ISO/IEC 42001:2023 Information technology — Artificial intelligence — Management system (2023). <https://www.iso.org/standard/42001>. Accessed 10 Aug 2026.
- U.S. Food and Drug Administration. Marketing Submission Recommendations for a Predetermined Change Control Plan for Artificial Intelligence-Enabled Device Software Functions (2025). <https://www.fda.gov/regulatory-information/search-fda-guidance-documents/marketing-submission-recommendations-predetermined-change-control-plan-artificial-intelligence>. Accessed 10 Aug 2026.
- U.S. Department of Health and Human Services. Nondiscrimination in Health Programs and Activities: Final Rule. *Fed. Regist.* 89, 37522–37703 (2024). <https://www.federalregister.gov/documents/2024/05/06/2024-08711/nondiscrimination-in-health-programs-and-activities>. Accessed 10 Aug 2026.
- Mello, M. M. & Guha, N. Understanding liability risk from using health care artificial intelligence tools. *N. Engl. J. Med.* 390, 271–278 (2024). doi:10.1056/NEJMhle2308901.
- Stern, A. D., Goldfarb, A., Minssen, T. & Price, W. N. II. AI insurance: how liability insurance can drive the responsible adoption of artificial intelligence in health care. *NEJM Catal. Innov. Care Deliv.* 3, CAT.21.0242 (2022). doi:10.1056/CAT.21.0242.