DIAS™ EXECUTIVE RESEARCH SERIES · THE INTEROPERABILITY TARIFF™ · PAPER NO. 2

When the Loop Runs Itself

How AI agents earn the right to act,
governed by blast radius — not accuracy.

The thesis in one line — *agentic AI does not just suggest, it acts; the moment an autonomous agent writes back into the chart, the interoperability tariff stops being paced by a human and starts running on autopilot. Governing and metering that loop is the central enterprise challenge of healthcare AI in 2026.*

Philip Morales
Founder & Chief Transformation Officer
Digital Intelligence Advisory Services (DIAS™)

Version 1.3 · July 2026 · For Public Distribution

CONTENTS

Executive Summary

Healthcare AI is crossing from assistance to action. Through 2025, large language models summarized charts, drafted notes, and offered decision support — always with a human between the model and the record. In 2026 that boundary is dissolving. Agentic AI — systems that autonomously reason, plan, and execute multi-step tasks under defined human oversight — can now log into systems, retrieve data, and complete work on their own.

The signal is unmistakable: the FDA deployed an agency-wide agentic AI platform in December 2025, ARPA-H launched a program to field the first FDA-authorized agentic system for clinical care, and Deloitte reports that over 80% of health system and health plan executives expect generative and agentic AI to deliver moderate-to-significant value in 2026. The pressure is structural, and it is not primarily a headcount problem. Of an estimated $1.6 trillion in total annual U.S. healthcare waste, administrative complexity is the largest addressable share at roughly $350 billion — rules-heavy, high-volume work that consumes clinical labor without requiring clinical judgment.

This paper makes one argument. The earlier paper in this series, *The Interoperability Tariff*, showed that the durable cost of EHR-integrated AI is the toll paid every time data crosses a boundary — extract, normalize, infer, and especially write back. Agentic AI changes the economics and the risk in a single move: it removes the human who was pacing that loop. An agent that writes back runs the loop continuously, at machine speed, across many systems. The tariff no longer compounds at the speed of staff — it compounds at the speed of compute.

Two things must therefore be engineered together: a governance model that graduates autonomy as validation earns it, and a cost model that meters every crossing the agent makes. This paper offers both — the Write-Back Autonomy Ladder™ and the agentic extension of the tariff — plus the regulatory context, a segment view for providers, payers, and pharma, honest counterpoints, and a staged path to start.

1. The Agentic Shift: From Copilot to Coworker

Agentic systems differ from the chatbots most people use today in one decisive way: they take action. Where generative AI produces content reactively, agentic AI decomposes a goal, coordinates steps across disparate systems, and adapts based on outcomes — enabled by models that can log into systems, retrieve data, and complete multi-step tasks rather than merely answering questions.

The institutional momentum is real and recent. The FDA introduced a secure, agency-wide agentic AI system in December 2025 — built in GovCloud, with embedded human oversight, and explicitly not trained on regulated-industry data — building on its Elsa assistant, voluntarily used by more than 70% of FDA staff. In healthcare delivery, Deloitte finds over 80% of health system and health plan executives expect generative and agentic AI to deliver moderate-to-significant value in 2026, and identifies ambient clinical intelligence (agents that listen, generate structured notes, update the EHR, and flag gaps) as the most mature agentic category to date.

The driver is not novelty. Nor, on the current evidence, is it scarcity. Mercer's 2024 analysis projects a nationwide shortage of roughly 100,000 healthcare workers by 2028 — real, but modest, and accompanied by projected national surpluses of registered nurses and physicians, with the deficits concentrated in nursing assistants and other support roles. A labor shortage of that size would not, by itself, make autonomy a necessity. What makes it a necessity is the composition of the work. Of the estimated $1.6 trillion in total annual U.S. healthcare waste, administrative complexity is the single largest addressable category at roughly $350 billion, and administrative burden is reported to consume about 20% of institutional budgets. That spend is overwhelmingly rules-heavy, high-volume, and low-judgment — precisely the work that agents perform well and that licensed clinicians perform expensively. The case for agentic AI is not that there are too few people. It is that too many of them spend their days on work that does not require a person. Autonomy is a reallocation of judgment, not a replacement for it.

2. The Write-Back Leap

Most agentic value in healthcare is not in reading — it is in acting. An agent that drafts a note is useful; an agent that posts the note, updates the problem list, submits the prior-authorization request, or adjusts the order is transformative. That step — writing back into the system of record — is where autonomy meets the EHR, and it is the precise point at which cost and risk inflect.

> **FRAMEWORK · The Write-Back Leap**
>
> In the Interoperability Tariff model, every workflow runs a four-step loop — extract → normalize → infer → write-back — and each step is a metered border crossing. A human copilot paces that loop: one encounter, one review, one write-back at a time.
>
> Agentic AI removes the pacer. The loop now runs continuously, in parallel, across many systems and sites — so the tariff is no longer charged at the speed of staff but at the speed of compute. **Autonomy is not a new cost line; it is a throttle removed from the existing one.**

This reframes the executive question. It is not “should we adopt agents?” — the trend answers that. It is “at what level of autonomy may an agent write back to this workflow, and how do we meter and govern each crossing it makes?”

3. The Cost Inflection: The Meter on Autopilot

Agentic workflows are token-hungry by design. Gartner estimates that agentic models consume between 5 and 30 times more tokens per task than a standard chatbot, because each task becomes a multi-step reasoning loop rather than a single call. Run that loop autonomously, thousands of times a day, across every site and EHR platform, and the cost curve steepens sharply.

It is tempting to assume falling model prices will absorb this. They will not. Gartner projects inference will become roughly 90% cheaper by 2030, but that deflation applies to the *model* layer. The tariff — integration, normalization, write-back validation, certification, and maintenance — is labor- and license-driven, and agentic autonomy multiplies the number of crossings rather than the price of each. Cheaper tokens times far more loops is not a saving; it is a larger bill arriving faster.

> **FRAMEWORK · Govern and Meter Every Crossing**
>
> Treat each agent as a tariff meter. Before granting write-back autonomy to any workflow, instrument three things: tokens and tool-calls per completed task, the number of border crossings per loop, and the multiplication factor across sites, platforms, and workflows. **Autonomy without metering is an open-ended liability; autonomy with metering is a managed unit cost.**

4. The Risk Inflection: Autonomy Meets the Chart

Writing back is already the hardest integration — it requires validation, conflict resolution, and adherence to each EHR's business rules, and 95% of organizations still depend on legacy HL7 v2 interfaces for it. Autonomy raises the stakes on three fronts at once:

- **Clinical safety and liability.** Peer-reviewed work is blunt about the open questions: who is responsible if an autonomous agent causes harm, and how should clinicians share authority with a system that acts independently? Diagnostic and medication-related actions carry the highest scrutiny.
- **Security and shadow AI.** Healthcare breaches average $7.42 million — the highest of any industry — and IBM found 97% of AI-related breaches occurred where proper access controls were absent, with 63% of organizations lacking formal AI governance. An autonomous agent with write credentials is a powerful new attack surface.
- **Regulatory exposure.** True agentic systems for clinical decision-making face the most stringent (Class III) pathway, and the rules are crystallizing fast — see the landscape below.

The 2026 Regulatory Landscape for Agentic Write-Back

Instrument	What It Requires	Implication
FDA / EMA joint AI principles (Jan 14, 2026)	First transatlantic alignment on AI validation and good practice	Validation is now table stakes, internationally
FDA PCCP (adaptive AI devices)	Transparent update protocols and post-market monitoring	Plan for change control before deployment
EU AI Act / MDR Article 14	Human oversight, traceability, and human override for high-risk AI	Full autonomy is constrained by design
CMS-0057-F (Jan 1, 2027)	FHIR-based prior-authorization APIs across major payers	Agentic prior-auth has a hard deadline
U.S. state AI laws (emerging)	Disclosure, oversight, and governance obligations	Governance frameworks must be auditable

Blast Radius: Same Error Rate, Very Different Consequences

The math of chained autonomy is unforgiving. A single agent that is 95% reliable sounds production-ready; chain five such steps and end-to-end reliability falls to roughly 77% ($0.95^5 \approx 0.774$), because errors compound at every handoff. This is why multi-agent demos impress while multi-agent deployments quietly underperform. But raw accuracy is the wrong target. What matters is *blast radius* — a term borrowed from site-reliability and security engineering for the scope of damage when something fails. A summarization agent that errs is an annoyance; a claims, eligibility, or write-back agent that errs produces member harm, mispriced risk, or a PHI-exposure and audit finding. Same 5% error — very different consequences.

Design around blast radius, not accuracy. In regulated healthcare, evaluations are not a QA checkbox at the end of a build — they are the control system that governs autonomy. Three principles follow. Weight the evaluation budget by blast radius, and test the high-consequence seams hardest. Evaluate the handoffs, not just the endpoints, because most failures live in the seam between agents, not inside any single model. And treat the evaluation suite as the contract you hand to compliance and the auditor — the standing evidence that, when an agent is wrong, the damage is bounded.

This is what the Ladder operationalizes. A workflow earns a higher rung only where its blast radius is bounded and its evaluations prove containment; the audit-and-rollback spine and confirmation gates exist to keep the radius small when a step does fail. The winners in regulated AI will not be the teams with the most agents — they will be the ones who can prove exactly what happens when one is wrong, and that the blast radius is contained.

5. The Write-Back Autonomy Ladder™

Autonomy is not binary. The right question is not whether an agent may act, but how much it may act in a given workflow, and what controls ride alongside. The literature already points the way: human-in-the-loop is operationalized through confirmation gates, uncertainty-triggered pauses, and rollback controls, and tiered-oversight models use higher-tier (and human) review triggered by risk and confidence — not confidence alone. The framework below turns that into an executive-ready ladder.

> **FRAMEWORK · The Write-Back Autonomy Ladder**
>
> A workflow earns each rung only after the prior rung is validated in production. **Autonomy is granted, audited, and revocable — never assumed.**

Tier	What the Agent Does	Human Control	Typical Fit
T0 — Read	Retrieves and summarizes; no write	N/A (read-only)	Analytics, search
T1 — Draft	Prepares the write-back; human posts it	Mandatory sign-off on every action	Notes, coding suggestions
T2 — Confirm	Acts on explicit confirmation	Confirmation gate per action	Prior-auth drafting, orders
T3 — Act & Notify	Acts within a bounded policy; notifies	Post-hoc review; instant rollback	Routine, low-risk updates
T4 — Autonomous	Acts within a validated envelope	Overseer agent + audit + escalation	Narrow, high-volume, proven

Two controls are non-negotiable at every rung above T1. First, an **audit-and-rollback spine**: every agent action is logged with its rationale and is reversible. Second, **complexity-adaptive escalation**: the agent escalates to a higher tier or a human not only when its confidence is low, but when the assessed risk is high — because the dangerous cases are often the ones where a model is confidently wrong. For the top rung, a dedicated overseer agent continuously monitors the acting agents — the same supervisory pattern ARPA-H is funding for its FDA-authorized clinical agent program.

6. Where It Lands: Providers, Payers, Pharma

Segment	First Agentic Write-Back Workflows	Governing Tier to Start
Providers	Coding suggestions, ambient note posting, order drafting, prior-auth preparation	T1–T2, graduating by workflow risk
Payers	FHIR prior-authorization under CMS-0057-F, utilization-management drafting, status write-back	T2 with confirmation gates
Pharma	Pharmacovigilance triage, safety-signal drafting, regulatory-submission assembly	T1 with mandatory expert sign-off

The pattern is consistent: start where volume is high and clinical risk is bounded, hold autonomy low, instrument heavily, and let validated performance — not enthusiasm — buy each step up the ladder.

7. Honest Counterpoints

“The EHR vendors' native agents will handle this.”

For single-vendor, single-workflow cases, native agents can reduce integration cost. But most health systems run multiple EHRs after M&A, and payer and pharma flows span organizations — so the multi-platform write-back problem, and the need for a vendor-neutral autonomy and audit layer, persists. Native agents are a participant in the architecture, not a replacement for governing it.

“Regulation will block autonomy anyway.”

Regulation constrains — it does not prohibit. EU Article 14 mandates oversight and override; the FDA's own deployment is autonomous-with-guardrails; ARPA-H is explicitly funding an FDA-authorized clinical agent. The winning posture is graduated autonomy with auditable controls, which is exactly what the rules reward.

“Tokens are getting cheap, so cost will solve itself.”

Token deflation is real and is the wrong thing to optimize for. Agentic loops multiply crossings 5–30× and run without a human pacer; the durable cost is the integration tariff, which does not deflate. Cheaper tokens make more autonomy affordable, which increases total crossings — the opposite of a self-solving problem.

8. How to Start: A 90-Day Path

1. **Weeks 1–3 — Pick one bounded loop.** Choose a high-volume, low-clinical-risk write-back workflow (coding suggestions or prior-auth drafting). Define the autonomy tier ceiling up front.
2. **Weeks 4–8 — Build the spine first.** Stand up audit logging, rollback, confirmation gates, and the tariff meter before granting any autonomy. Run the agent at T1 (draft-only).
3. **Weeks 9–12 — Earn the next rung.** Validate accuracy, safety, and unit cost in production. Graduate to T2 only on the evidence, and only for that workflow.
4. **Ongoing — Add an overseer, then scale.** Introduce a supervisory agent and complexity-adaptive escalation before contemplating T3–T4, and expand workflow by workflow under the same discipline.

Conclusion

Agentic AI is the defining healthcare-technology shift of 2026, and its value lives in write-back — the moment AI stops suggesting and starts acting in the record. That same moment removes the human who paced the interoperability tariff, so cost and risk both inflect at once. The organizations that win will not be those with the most autonomous agents; they will be those that govern autonomy as a ladder and meter every crossing as a unit cost.

Grant autonomy the way you grant clinical privileges: in tiers, on evidence, with audit, and revocable at any time.

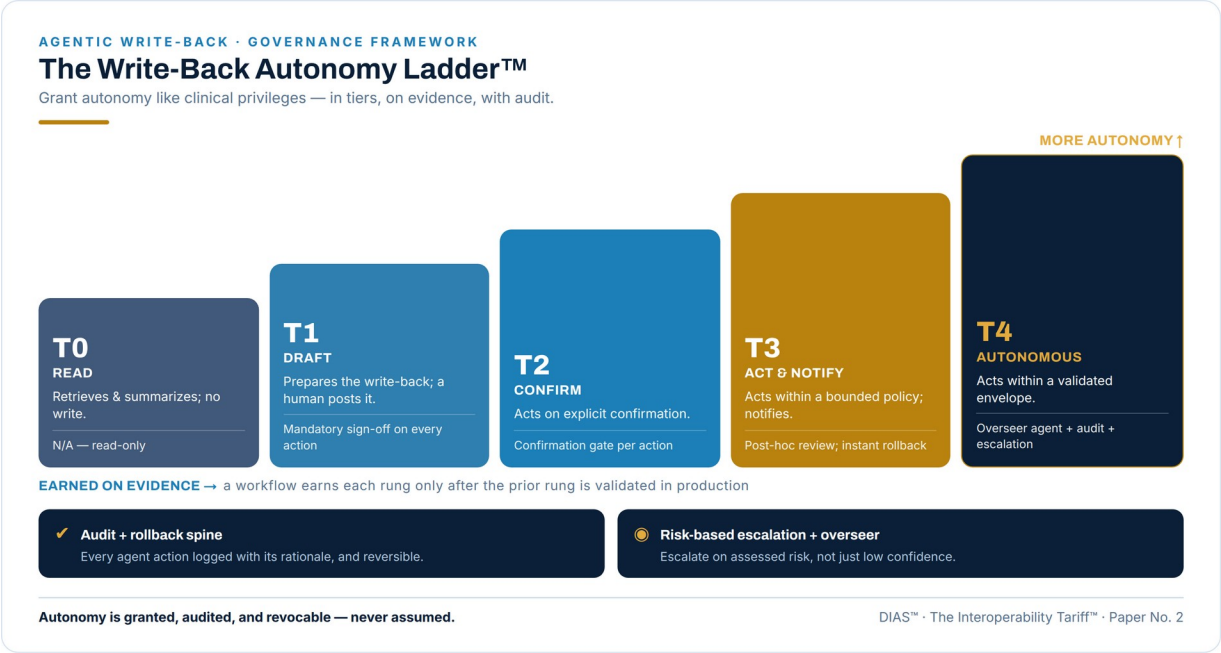

Figure 1 · The Write-Back Autonomy Ladder™ — earning authority to act, one governed rung at a time (T0→T4).

Sources, Methodology & Validation

Methodology

Every quantitative and regulatory claim in this paper is drawn from a primary or recognized secondary source dated 2025–2026 and is cited below. Framework elements attributed to the author (the write-back leap, the Write-Back Autonomy Ladder™, and the govern-and-meter principle) are original analytical contributions, not sourced claims, and are labeled as such. Cost-trajectory statements are directional and depend on each organization's workflow mix, site count, and platform mix.

All regulatory claims were verified against primary government sources. Quantitative market, adoption, and implementation figures were verified against the best available 2025–2026 industry, peer-reviewed, or recognized secondary sources. Tiers — **Verified** (confirmed against the issuing organization's own publication), **As reported** (accurately attributed to the source that reported it), and **Directional** — are marked in each reference entry. **Of 20 sourced claims: 12 verified against a primary or issuing source; 8 accurately attributed as reported.** The "hospital FHIR API use 56%→84%" figure has been withdrawn rather than reported — it could not be traced to an issuing source, consistent with its treatment in Paper No. 1. Analytical contributions — the Write-Back Autonomy Ladder™, the agentic extension of the Interoperability Tariff™, the $0.95^5 \approx 0.774$ arithmetic, and the blast-radius framing — are original or independently verifiable and are excluded from the count.

Version 1.3 (July 2026) corrections: the ladder has five rungs (T0–T4); the ~$1.5–1.6T figure describes total U.S. healthcare waste across all six categories, with administrative complexity the largest addressable share at ~$350B; and the workforce figure is Mercer's 2024 projection of ~100,000 by 2028, superseding the 2021 projection of a 3.2M shortfall among lower-wage support occupations. Cross-references are aligned to published paper titles.

Sources

[Verified] U.S. FDA press release, December 1, 2025 ("FDA Expands Artificial Intelligence Capabilities with Agentic AI Deployment"). Agency-wide agentic AI deployed in a secure GovCloud environment with embedded human oversight; models do not train on staff input or regulated-industry data; use is entirely optional. Elsa, deployed May 2025, is voluntarily used by more than 70% of FDA staff.

[As reported] Oral Health Group (Feb 2026), synthesizing FDA data. ~1,250 AI-enabled devices authorized (mostly narrow imaging); clinical agentic decision-making faces the Class III pathway; administrative burden ~20% of institutional budgets.

[As reported] Dataiku (Apr 2026). FDA and EMA published joint good-AI-practice principles on Jan 14, 2026 — first transatlantic alignment on AI validation; "agentic shift" and large action models.

[As reported] npj Digital Medicine (Mar 2026). FDA PCCP requires transparent update protocols and post-market monitoring; open questions on autonomous-AI liability and shared authority; human-in-the-loop emphasis.

[As reported] Human-in-the-loop and oversight (peer-reviewed, 2026). Confirmation gates, uncertainty-triggered pauses, and rollback; tiered agentic oversight with complexity-adaptive, risk-based escalation.

[As reported] Manatt Health Policy Tracker; Fierce Healthcare (2026). ARPA-H ADVOCATE — a 39-month program to field the first FDA-authorized agentic clinical-care system, including a supervisory "overseer" agent; teams selected June 2026.

[Verified] Deloitte, *2026 US Health Care Outlook Survey* (n=120 US C-suite health system and health plan executives). Over 80% expect generative and agentic AI to deliver moderate-to-significant value across clinical, business, and back-office functions in 2026; ambient clinical intelligence identified as the most mature agentic category to date.

[Verified] Shrank, Rogstad & Parekh, *JAMA* (2019); restated to current CMS expenditure. Total U.S. healthcare waste across six categories estimated at $760–935B (2019), ≈$1.6T at 2025 spending levels; administrative complexity is the largest addressable category at roughly 22% (≈$350B). Administrative waste is a subset of total waste, not a synonym for it.

[Verified] Mercer, *Future of the U.S. Healthcare Industry: Labor Market Projections by 2028* (2024). Nationwide shortage of ~100,000 healthcare workers projected by 2028, concentrated in nursing assistants and support roles; supersedes Mercer's 2021 five-year projection of a 3.2M shortfall among lower-wage support occupations.

[Verified] Gartner (March 25, 2026). Agentic models consume 5–30× more tokens per task than a standard GenAI chatbot; inference on a one-trillion-parameter LLM projected to cost providers over 90% less in 2030 than in 2025. As token consumption rises faster than token costs fall, overall inference costs are expected to increase.

[Verified] IBM *Cost of a Data Breach 2025*. Healthcare breach average $7.42M (highest of any industry); 97% of AI-related breaches lacked proper access controls; 63% of organizations lack formal AI governance.

[As reported] CapMinds / Taction (2026). Write-back is materially harder than read (validation, conflict resolution, business rules); 95% of organizations still use HL7 v2.

[Verified] CMS-0057-F, published 89 FR 8758 (Feb 8, 2024); effective April 8, 2024. Operational prior-authorization provisions generally due January 1, 2026; major API requirements generally due January 1, 2027. EU AI Act / MDR Article 14 human-oversight requirements.

[Analytical, excluded from the count] Multi-agent reliability & blast radius (2026). End-to-end reliability of chained agents compounds multiplicatively ($0.95^5 \approx 0.774$; independently verifiable). “Blast radius” is a standard site-reliability and security term for the scope of impact of a failure, applied here to agentic handoffs. Analytical framing, not a sourced statistic.