A FACTORYOS WHITEPAPER · COMPTRIO

Your Company Is Not One Trust Domain

Why "private AI" usually isn't private enough, and the missing layer that makes business AI safe to actually use

By Shawn Snarski

June 2026 · comptrio.com · factoryos@comptrio.com

Executive summary

Every firm that holds other people's secrets faces the same standoff with AI. A medical practice, a law or accounting office, a wealth manager: the tools are too useful to ignore and too leaky to adopt. The industry's answer has been "private AI": keep the data out of the public cloud. That solves a real problem. It also solves the *wrong half* of the problem.

Keeping your data out of someone else's cloud stops outsiders from reading it. It does nothing about the second wall, the one between your own people. The bookkeeper isn't cleared for the partner-comp file. The front desk isn't cleared for another patient's chart. The new hire doesn't see the board minutes. None of it is written down as a policy. It's just how the place runs.

The fix is to make that access map -- the one your firm already runs on -- the way the AI itself is set up, so each person reaches only their authorized slice. This matters because the moment you point one ordinary AI assistant at "all the company's files," you build a machine that ignores every one of those distinctions: it answers each person identically, from the entire corpus, across the walls you spent years maintaining.

And it does so for a structural reason, one no perimeter can reach. Ordinary search respects folders -- if you can't open the folder, you can't find the file. AI retrieval doesn't: it finds by *meaning*, not permission -- so unless an authoritative partition gates that search, it reaches across every wall at once, never consulting your org chart. The gap is fixable -- this paper is about how -- but not by keeping data off the cloud.

This paper names that gap, **internal data sovereignty**, explains why nearly every AI product on the market ignores it, gives you a vendor-neutral checklist to test any tool against, and is candid about where we stand. At bottom it is one idea most of the industry has never applied to AI: **zero trust**, enforced where the AI actually reads.

> **The one sentence the gap comes down to:** *keeping your data off the cloud keeps the world out. It does nothing to keep the wrong colleague out.*

---

1. The trap every firm walks into

The pitch for business AI is genuinely compelling: point it at everything your company knows, and let anyone ask questions in plain language. The problem reveals itself the instant you take it seriously in a regulated firm.

"Everything your company knows" includes the things that must never cross certain desks. Patient histories. Privileged matter files. Client financials. Board minutes. The comp spreadsheet. A firm's entire professional

standing rests on the fact that these are *compartmentalized*. Knowing a secret and being *cleared* for it are two different things.

So the firm is offered a binary. Adopt the convenient cloud tool and accept that your clients' secrets now live on infrastructure you don't control, or abstain and watch competitors pull ahead. Most of the market has spent two years arguing about that binary. It's the wrong argument, because it only concerns one of your two confidentiality walls.

---

2. The blind spot: a company is not one trust domain

Here is the assumption hiding inside almost every AI product, stated plainly: **a company is one thing, with one set of permissions, in or out.**

No real firm works that way. A company is a federation of trust boundaries:

- The intern does not get the resident surgeon's access.
- The bookkeeper does not get the CFO's.
- The associate does not get the file behind the ethical wall.
- Operations does not see HR. Support does not see the board. Almost no one sees C-level compensation, M&A, or the cap table.

You already enforce this, with separate drives, locked folders, "need-to-know," and years of habit. It is not a nice-to-have. For many firms it is a *legal obligation*. HIPAA's "minimum necessary" standard,[1] GDPR's data-minimization principle,[2] and the legal profession's ethical walls[3] are all, at bottom, rules about which *insider* may see which slice of the data.

Then a single AI assistant is dropped on top of the whole corpus, and it cheerfully dissolves every one of those walls, because nobody told it the walls were there.

There are really two walls, and most of the market only builds the first one:

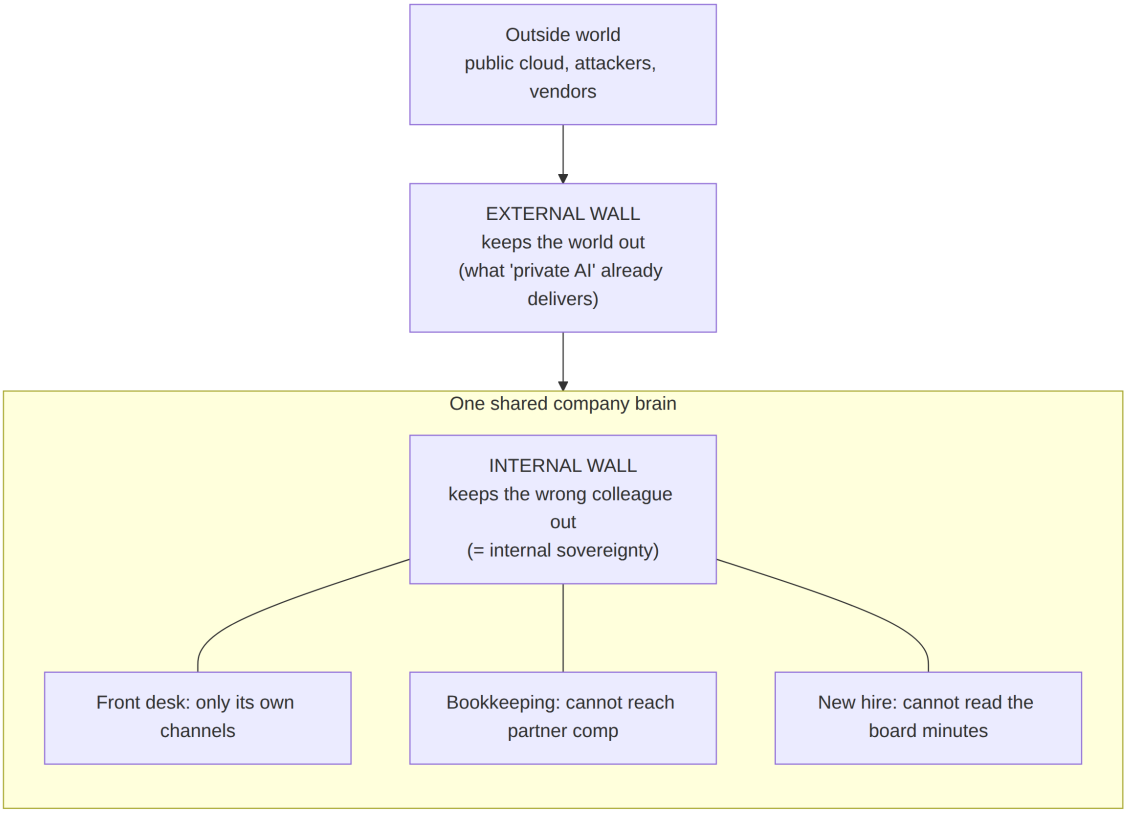

Figure 1. The two walls. *Private AI builds the outer wall. The inner wall, between your own people, is the one almost every AI product leaves out.*

---

3. Two honest attempts, and why each leaves the seam open

The market's serious answers fall into two camps. Both are real improvements. Neither closes the internal wall.

Camp A: sovereign appliances ("keep the data off the cloud"). These put the model on a box on your premises. The data never leaves the building, which is exactly right and not nothing. But ask the next question. *How does this box decide that two different people should see two different things?* Mostly, it doesn't. The access boundary is the **network** or the **hardware itself**, so whoever can reach the box can reach what's on it. A sophisticated appliance can bolt a permission layer onto the shared box -- but the moment it does, that layer is the same coarse-or-derivative model we meet in Camp B, now running on-prem, and the interior seam reopens with it. Strip that permission layer away, and the appliance breaks both ways: a shared box means everyone on it sees everything, and a box-per-person means no shared company brain at all, just a row of disconnected islands, with no single corpus in which the CFO sees the comp file and operations does not.

Camp B: private "chat over your documents" software. These add real multi-user support, which Camp A lacks. But the permission model is either **coarse** or **derivative**. Coarse means access is granted at the level of a whole workspace or group, not the document and the retrieval. Derivative means it mirrors the permissions of

the source systems it connects to: your shared drive, your wiki, your chat tool. Derivative permissions are only ever as correct as the upstream sharing settings that no one has audited since 2021. They are not *authoritative*. They are an echo of someone else's housekeeping. This is a failure mode the largest vendor in the category documents openly. Microsoft ships a deployment "blueprint" for mitigating Copilot *oversharing*, precisely because Copilot surfaces whatever a user's existing permissions already expose, and most organizations' permissions are looser than anyone believes.[4] This is not a theoretical worry. When one data-security firm measured it across more than 550 million records, **16% of an organization's business-critical data was overshared** -- an average of **802,000 files per company** -- and **83% of those exposures were to insiders**, the precise wall an AI assistant then reads straight through.[5]

Approach	What it gets right	Where the access boundary lives	The seam it leaves open
Camp A (sovereign appliance)	Data never leaves the building	The network / the hardware	Everyone who can reach the box sees everything; one-box-per-person means no shared brain
Camp B (chat over your docs)	Real multi-user access	Inherited from your other apps' sharing settings	Permissions are coarse, or derivative and only as correct as years-old sharing nobody audits
Internal data sovereignty (the missing layer)	Both walls, inside one shared brain	The data itself: per channel and role, on every retrieval	Closes the seam: authoritative, default-deny, structural

Table 1. Where each approach draws the line, and where the line leaks.

> *Camp A keeps the world out but treats your whole firm as one room. Camp B lets your firm into many rooms but borrows its locks from systems you no longer trust. The interior wall, authoritative, owned, and enforced, is what neither ships.*

4. Where the wall actually breaks

A classic search respects folders. If you can't open the folder, you can't find the file. An AI assistant pointed at a shared corpus without an authoritative partition doesn't work that way. It finds information by *meaning*, through **semantic search** across everything it has ingested, and unless that search is gated by an authoritative wall, it answers from relevance alone and consults no org chart. Ask it to "pull together what we know about the Hartwell matter," and it reaches for whatever is most relevant, wherever that lives: across the privilege wall, the PHI boundary, the comp file, if those are the closest match.

This is also why a perimeter mindset isn't enough. You can lock the network and still leak, because the leak happens *inside*, at the retrieval step, long after the perimeter was cleared.

So the breach doesn't take a hacker. It takes an employee asking a reasonable question:

- The bookkeeper asks the assistant to "summarize the partners' arrangements," and partner compensation lands in the answer.
- A front-desk staffer asks about "the Riggs appointment," and a different patient's history gets pulled in for context.
- An associate asks about a deal they're walled off from, and the wall, which existed only as a folder permission nobody wired into the assistant, never gets consulted.

No breach to detect, no attacker to blame. Your own tool, doing exactly what it was built to do, with data it was never cleared to combine. A retrieval system without an authoritative internal partition is a confidentiality incident waiting for the right prompt, and the prompt that sets it off will be a perfectly innocent one.

None of these cases involves AI at all -- and that is exactly why they matter. The failure they share -- an insider reaching records they had no business reaching, with no outside attacker in the story -- is already illegal, already caught, already fined. That is the baseline an AI assistant inherits; it doesn't create this risk.

Action (year)	Penalty	What happened
Yakima Valley Memorial Hospital (2023)	**$240,000**	23 security guards accessed 419 patients' records with no job-related purpose[6]
Montefiore Medical Center (2024)	**$4.75M**	An employee accessed and sold 12,517 patients' data; the safeguards to catch it were missing[7]
Gulf Coast Pain Consultants (2024)	**$1.19M**	A former contractor reached 34,310 patients' records, partly because their access was never terminated[8]
Memorial Healthcare System (2017)	**$5.5M**	Employees impermissibly accessed 115,143 records, and the system couldn't prove who saw what[9]

Table 2. Recent HIPAA penalties for insider access -- human insiders, no AI and no outside attacker in any of them. They show the baseline failure an AI retrieval layer would automate, not instances of it. The fines span 2017 to 2024; a standing pattern, not a one-off.

And it is the common case, not the edge case. In healthcare, insiders are still behind roughly **30% of breaches**, about one in three, even now that external attackers and ransomware dominate the headlines, and internal actors remain the primary source of the two patterns that map exactly to the risk here: privilege misuse and ordinary mistakes.[10] Meanwhile a healthcare breach is still the costliest of any industry, for the fourteenth year running, averaging **$7.42 million** in 2025 even after a sharp drop from the year before; in the US, where these firms

operate, the average breach across all industries just set an all-time high of **$10.22 million**.[11] The courts are beginning to name AI directly, too: a November 2025 class action accused a health system of deploying an AI "ambient scribe" (Abridge) that recorded patient visits without consent -- a consent claim, not a retrieval-wall one, but a sign the legal weather is turning toward AI specifically.[15]

And ordinary use leaks a second way, out through the *other* wall -- the table-stakes one. Give your people no sanctioned, private AI and they will reach for an unsanctioned one -- a study of 1.6 million workers found **11% of everything employees paste into ChatGPT is confidential**[12] -- and a 2026 update, measuring more broadly, finds **39.7% of all AI interactions now expose sensitive data**.[13] IBM now puts **shadow AI inside one in five breaches (20%)**, adding **$670,000** to the average breach cost and compromising customer data at a higher rate (65%) than the norm -- and of the organizations that suffered an AI-related security incident, **97% had no AI access controls in place at all**. That last figure belongs to the inner wall: access controls on the AI itself are exactly what is missing -- and the governance that would mandate them barely exists yet, with **63% of organizations either lacking an AI governance policy or still drafting one**.[14] Both walls fail the same way: not a hacker, but a well-meaning employee, doing ordinary work, with a tool never built to keep your firm's secrets apart.

An AI assistant that retrieves across the internal wall doesn't invent a new risk. It industrializes the one regulators are already fining, and hands it a search box.

---

5. The missing layer: internal data sovereignty (zero trust, where the AI actually reads)

The fix is a layer the generic tools skip entirely. Call it **internal data sovereignty**: an authoritative, role-based partition *inside* one shared company brain, enforced at **every retrieval**, with the access boundary being the data itself. Not the network. Not the box. Not an echo of another app's settings.

If that sounds like zero trust, it is. Zero trust starts from a single assumption: no actor is trusted by default, and every access is verified, every time. Its first principle is **default-deny**. The industry has spent a decade applying that idea to networks, devices, and logins. Applying it to the place an AI actually reads from -- the vector index, the knowledge graph, the memory store -- is only now forming as a category, and most early implementations still inherit the derivative or coarse permissions this paper has been describing. Internal data sovereignty is that idea taken to its authoritative conclusion: zero trust finally reaching the retrieval path.

It also reframes what "sovereignty" has to mean. The usual definition stops at company scope: your data stays inside your walls. That is necessary and not enough. The instant departments, roles, and individual users start bleeding into one another inside a single AI, the boundary has to be far more granular than "the company." It has to run per channel, per role, per person, enforced on the retrieval itself.

Done properly, it has four non-negotiable properties:

Property	What it means	The failure it prevents
Authoritative, not derivative	The permission system *is* the source of truth -- one place to set and inspect -- not a mirror of your shared drive's settings	Inheriting stale, unaudited sharing rules from your other apps
Enforced at every retrieval	The check runs on vector search, graph traversal, and memory lookup, every path, not just a folder UI	A single unchecked retrieval path quietly bypassing the wall
Default-deny	Access is granted, never assumed; new data is private until shared	Data left readable only because no one got around to locking it
Structural, not procedural	The wall is *how the system is built*, not a rule you hope people follow	A curious employee, or a clever prompt, talking its way around the rule

Table 3. The four properties that separate a real internal wall from a hopeful one.

> *The difference between a procedural wall and a structural one is the difference between "employees are instructed not to look" and "the system cannot show them." Only the second survives a curious employee, a clever prompt, and an auditor.*

What makes the wall structural rather than hopeful is *where* the check sits. Authorization runs before any query reaches the index, the graph, or the memory store, so the system forms no request it isn't cleared to make. It does not retrieve across the whole corpus and trim the answer afterward -- which is exactly what a permission layer bolted onto a shared appliance does, and exactly why such a layer stays only ever as good as its trimming. Enforcing the boundary *before* retrieval, not after, is the difference between a system that **cannot** surface a record and one merely asked to filter it back out.

This is not exotic. It is simply the wall your firm *already enforces in the physical world*, finally expressed in the one place AI actually reads from.

---

6. How to test any vendor: a checklist you can use today

You do not need FactoryOS to apply this. Bring these questions to any "private," "on-prem," or "secure" AI vendor. The ones built on the old one-trust-domain assumption will struggle with the second half of the list, and that tells you what you need to know.

On keeping the world out (table stakes):

Ask your vendor	What a confident, passing answer sounds like
Does *any* data leave the building, including telemetry, logs, or prompts used to improve a model?	"Nothing leaves. No telemetry, no logs, no training exhaust; it is air-gap compatible."
Is anything we type ever used to train a model we don't own?	"Never. Your data trains nothing you don't own."
Do we own the system outright, or rent it per seat? Is there a remote kill-switch or a forced-upgrade clause?	"Owned outright, one-time purchase. No per-seat rent, no kill-switch, no forced upgrades."

On keeping the wrong colleague out (the part most tools fail):

Ask your vendor	What a confident, passing answer sounds like
Is the access boundary the **network or the box**, or the **data**? Can two people on the same system see *different* things?	"The boundary is the data itself. Two people on one system see only their own authorized slice."
Is your permission model **authoritative**, or inherited from our other apps' sharing settings?	"Authoritative. The system is the source of truth, not a mirror of stale shared-drive settings."
Is access enforced on **every retrieval path** -- keyword, vector and semantic search, knowledge graph, memory -- through **one shared check**, or a separate filter bolted onto each? When you ship a new retrieval feature, does it inherit the wall by construction, or does someone have to remember to add the check?	"One check every path routes through, before any search runs. A new path inherits the wall automatically; there is no second copy of the check to forget."
Is it **default-deny** or default-allow?	"Default-deny. New data is private until it is explicitly shared."
Can the system express *"one shared company brain, where each person reaches only their authorized channels"* in a single install, not by buying separate boxes?	"Yes: one install, one shared corpus, partitioned per channel and per role."
When someone is offboarded, what can they no longer reach, and is that **instant**?	"Everything they were granted, revoked instantly at the boundary, not eventually."

On keeping judgment human:

Ask your vendor	What a confident, passing answer sounds like
For consequential actions, does a person review and approve before anything happens, with the reasoning and the underlying data in front of them?	"Yes. An approval gate sits at every consequential action, with the justification and the data shown."

Print this page. The questions that earn a confident answer from your vendor are the ones that protect your clients.

---

7. What the wall looks like in practice

The four properties above aren't abstract; they describe a buildable architecture. The access boundary lives on the data itself, organized into channels: whoever owns permissions sets which people reach which channel, there is a private channel for every person, and the default is deny. A person reasons across every channel they hold as one brain; the wall only keeps out the channels they were never given. Every retrieval path -- keyword, vector, graph, memory, calendar -- funnels through one authorization check before any search runs, so a new capability inherits the wall by construction; there is no second copy of the check to forget. And every access is written to an append-only audit log, so *who saw what, and when* has an answer that doesn't depend on anyone's memory -- and that you can ask any vendor to produce. That log earns its keep precisely because an internal leak sets off no alarm -- no attacker, no breach to detect -- and the average healthcare breach already runs **279 days before it is even identified and contained**.[11] Nine months of not knowing is the gap an append-only access record closes.

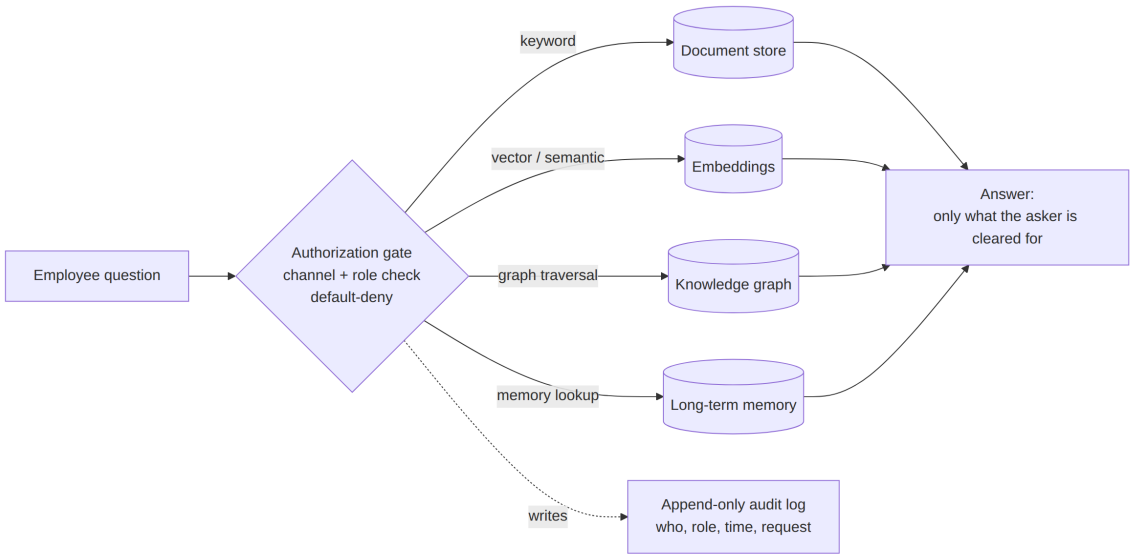

Figure 2. One gate, every path. *Keyword, vector, graph, and memory retrieval all pass the same default-deny check before any search runs, and every access is written to the audit log.*

A disclosure, in the spirit of the candor this topic deserves: this paper comes from Comptrio, maker of **FactoryOS** -- a platform built from the ground up around internal data sovereignty, with the channel as its only security boundary. We didn't reach this idea neutrally; we built a product on it, and the paper inevitably describes the problem in the terms our architecture answers best. So read us as believers with a stake in the outcome. The narrower claim is the one we'll stand behind: the premise doesn't depend on us. The sources are public, the failure modes are documented by the largest vendors in the category, and the checklist in Section 6 is vendor-neutral -- bring it to anyone, us included.

One honest limit belongs here, because it's a property of the idea, not of any product. The wall is only ever as right as the access map it's given: enforce a map perfectly and you enforce its mistakes perfectly too. *Authoritative* doesn't mean automatically correct -- it means one map, configured for this purpose, that fails closed; a derivative model only scatters the same duty across every app you connect. The architecture removes the chance of *accidental* leakage across retrieval; it does not remove the duty to decide who belongs where. Someone has to own that map and keep it honest.

The point all of it serves is one sentence no cloud assistant and no single-box appliance can honestly write:

> **Your bookkeeper's AI can't surface the partner-comp file, because the system was built so it can't, not because someone was told not to.**

---

8. The bottom line

If your firm holds other people's secrets, keeping data off the public cloud is **table stakes**: necessary, and not enough. The wall that actually defines a professional firm is the one *between its own people*, and it is the wall almost every AI product silently tears down the day it's installed.

Internal data sovereignty is that wall, rebuilt where AI actually reads. It is zero trust applied to the knowledge layer, and it is the difference between an assistant your compliance officer fears and one they can sign off on.

The distinction worth keeping is between *shouldn't* and *can't*. A procedural wall asks people not to look; a structural one is built so the system cannot show them. Done right, a retrieval layer cannot return a record from a channel the asker isn't cleared for -- not in an answer, not in a log, not in the UI -- and the same wall binds the agents. People can still leak; no architecture stops a colleague reading over a shoulder. But the *machine's* leak -- a retrieval system surfacing a record across an internal wall -- is the one this closes by construction. No public dataset counts that incident yet; it is the automated next step the failures in this paper foreshadow, not one they already measure. Closing it before it earns its own line in the breach reports is the whole point.

Published by Comptrio, maker of FactoryOS | comptrio.com | factoryos@comptrio.com

Sources

1. U.S. Department of Health & Human Services, Office for Civil Rights, "Minimum Necessary Requirement" (45 CFR Section 164.502(b), Section 164.514(d)). https://www.hhs.gov/hipaa/for-professionals/privacy/guidance/minimum-necessary-requirement/index.html
2. Regulation (EU) 2016/679 (GDPR), Article 5(1)(c): personal data shall be "adequate, relevant and limited to what is necessary in relation to the purposes for which they are processed" ("data minimisation"). https://gdpr-info.eu/art-5-gdpr/
3. American Bar Association, Model Rules of Professional Conduct, Rule 1.10 (Imputation of Conflicts of Interest) and the screening definition at Rule 1.0(k). https://www.americanbar.org/groups/professional_responsibility/publications/model_rules_of_professional_conduct/rule_1_10_imputation_of_conflicts_of_interest_general_rule/
4. Microsoft, "Mitigate oversharing to govern Microsoft 365 Copilot and agents" (Microsoft 365 Copilot Blog / Microsoft Learn). Copilot surfaces content according to each user's existing permissions, which is why Microsoft publishes an oversharing-remediation blueprint to run before deployment. https://techcommunity.microsoft.com/blog/microsoft365copilotblog/mitigate-oversharing-to-govern-microsoft-365-copilot-and-agents/4448744
5. Concentric AI, Data Risk Report (2026 update), based on more than 550 million records across the technology, financial, energy, and healthcare sectors: 16% of an organization's business-critical data is overshared -- an average of 802,000 files per organization -- and 83% of those at-risk files are overshared with internal users and groups, not outsiders. https://concentric.ai/too-much-access-microsoft-copilot-data-risks-explained/
6. U.S. HHS Office for Civil Rights, "Snooping in Medical Records by Hospital Security Guards Leads to $240,000 HIPAA Settlement" (Yakima Valley Memorial Hospital, 2023). 23 security guards accessed the records of 419 patients without a job-related purpose. https://www.hhs.gov/hipaa/for-professionals/compliance-enforcement/agreements/yakima/index.html
7. U.S. HHS Office for Civil Rights, "HHS' Office for Civil Rights Settles Malicious Insider Cybersecurity Investigation for $4.75 Million" (Montefiore Medical Center, 2024). An employee accessed and sold the PHI of 12,517 patients; safeguards to detect or prevent it were missing. https://www.hhs.gov/hipaa/for-professionals/compliance-enforcement/agreements/montiefore/index.html
8. U.S. HHS Office for Civil Rights, Notice of Final Determination, Gulf Coast Pain Consultants, LLC ($1,190,000 civil monetary penalty, 2024). A former contractor impermissibly accessed the ePHI of approximately 34,310 individuals; cited failures included not terminating former workforce access. https://www.hhs.gov/hipaa/for-professionals/compliance-enforcement/agreements/gulf-coast-pain-consultants-nfd/index.html
9. U.S. HHS Office for Civil Rights, Settlement Agreement with South Broward Hospital District d/b/a Memorial Healthcare System ($5.5 million, 2017). The PHI of 115,143 individuals was impermissibly accessed by its own employees; the case turned on missing audit controls. https://www.hhs.gov/hipaa/for-professionals/compliance-enforcement/agreements/hitech-sa/index.html
10. Verizon, 2025 Data Breach Investigations Report, healthcare snapshot: insiders accounted for roughly 30% of healthcare breaches (external actors 67%, partners 4%, multiple 1%), and internal actors remain the primary actors behind the Privilege Misuse and Miscellaneous Errors patterns. https://www.verizon.com/business/resources/infographics/2025-dbir-healthcare-snapshot.pdf
11. IBM / Ponemon Institute, Cost of a Data Breach Report 2025: healthcare remained the costliest industry for the 14th consecutive year at an average of $7.42 million (down from $9.77 million the prior year); the global cross-industry average fell to $4.44 million, even as the US average rose to $10.22 million, the highest of any region in the report's history. The same report puts healthcare's mean time to identify and contain a breach at 279 days, the longest of any industry. https://www.ibm.com/reports/data-breach
12. Cyberhaven Labs, analysis of 1.6 million knowledge workers: 11% of the data employees paste into ChatGPT is confidential. https://www.cyberhaven.com/blog/4-2-of-workers-have-pasted-company-data-into-chatgpt
13. Cyberhaven, 2026 update: 39.7% of AI interactions expose sensitive data. This measures a broader category (all AI interactions) than the 2023 paste-specific figure, and is cited only to show the direction of travel, not as a like-for-like increase. https://www.cyberhaven.com/blog/sensitive-data-flowing-into-ai-tools
14. IBM, Cost of a Data Breach Report 2025: The AI Oversight Gap (research by Ponemon Institute): one in five organizations (20%) reported a breach involving shadow AI; high levels of shadow AI added USD 670,000 to the average breach cost; such incidents compromised customer PII (65%) and intellectual property (40%) at higher-than-average rates; and of organizations that suffered

an AI-related security incident, 97% reported not having AI access controls in place; the same report finds 63% of organizations either have no AI governance policy or are still developing one. https://newsroom.ibm.com/2025-07-30-ibm-report-13-of-organizations-reported-breaches-of-ai-models-or-applications,-97-of-which-reported-lacking-proper-ai-access-controls

15. Class action filed against Sharp HealthCare (Nov. 26, 2025) alleging that an AI "ambient" clinical-documentation tool (Abridge) recorded patient-clinician conversations without consent, in violation of California's CIPA (all-party wiretapping) and CMIA. This is a consent/recording claim, not an internal-retrieval-wall case; it is cited only as evidence that AI-specific privacy litigation is emerging. https://www.fisherphillips.com/en/insights/insights/new-class-action-targets-healthcare-ai-recordings