TIDIO | Labs

Shadow AI

FROM HIDDEN HABIT TO DELIBERATE STRATEGY

TITLE	Shadow AI. From Hidden Habit to Deliberate Strategy
AUTHOR	Bart Turczynski
FOREWORD	Tytus Golas
EDITION	First edition, 2026
PUBLISHER	Tidio for Tidio Labs
RIGHTS	© Tidio, 2026. All rights reserved.

ISBN 978-83-68606-18-8

CONTENTS

If your people are already using AI without approval, then “doing nothing” is not a neutral choice: it’s a choice to keep the risk and forgo the upside.

Preface

Preface

Somewhere in most companies right now, an employee is pasting a customer email, a contract clause, or a block of source code into a chatbot the IT department never approved. This is **shadow AI**, and it isn't fringe behavior: depending on how you measure it, between roughly 50% and 90% of employees are already doing it, while only a minority of organizations have a policy, a sanctioned alternative, or any visibility into what's happening.[1,2,3]

That reframes the decision every lower-adoption organization faces. If your people are already using AI without approval, then "doing nothing" is not a neutral choice: it's a choice to keep the risk and forgo the upside. The real question isn't **whether** to engage with AI, but **how**: **tolerate** the ungoverned status quo, **build** something in-house, or **buy** a vetted tool. This briefing works through all three:

- First, the diagnosis: what shadow AI is, how widespread it has become, and what it risks.
- Second, the prescription: the economics of tolerate-build-buy, a staged path for moving deliberately, and how to measure the result.
- Third, the frontier: AI agents (systems that don't just advise but **act**), the most hyped and least mature category in enterprise technology, where the same discipline applies in sharper form.

The recurring answer is the same at every stage: meet the demand deliberately, keep scope narrow, and enforce your guardrails by design rather than by hope.

Tytus Gołas

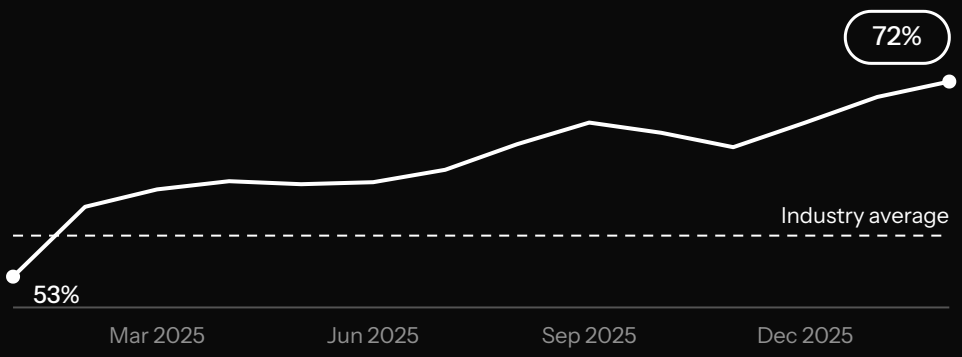
Co-founder and CEO, Tidio

MORE SALES, FEWER TICKETS.

The leading AI that resolves customer issues on its own

Lyro is the AI agent for customer service. It recommends products like an experienced shopping assistant, captures leads, closes sales, and resolves customer issues end to end. No army of engineers required.

Lyro's average resolution rate vs. the industry average



Market data: Contentsquare; Lyro's resolution rate represents the average across all customers.

Resolve 67% of customer queries with an AI agent that acts, not just chats. Lyro gets to work immediately, using your product catalog, knowledge base, and conversation history. Issues are handed off to people only when necessary.

TIDIO / Labs

Invite Lyro to join your team

Part One: The Diagnosis

What Shadow AI Actually Is

Shadow AI is the AI-era descendant of a much older problem: **shadow IT**, which Gartner defines as technology outside the ownership or control of the IT organization.[4] For decades, employees have quietly adopted their own apps, devices, and cloud services faster than IT could vet them. Generative AI is the same pattern at far greater speed, but with a sharper edge: the risk isn't just an unmanaged app. It's the company's own data being typed into a system that may store it, learn from it, or expose it.

In practice, "shadow AI" covers several distinct behaviors:

- **Unsanctioned use of public chatbots.** ChatGPT, Gemini, Claude, Copilot, and others used without approval. ChatGPT dominates: it's by far the most-used workplace AI tool.[5]
- **Personal accounts used for work,** the single largest blind spot. Cyberhaven's telemetry found that 73.8% of workplace ChatGPT accounts are personal, non-corporate accounts that lack enterprise security and privacy controls.[5] Browser-security firm LayerX puts the share of generative-AI access happening through personal accounts at roughly two-thirds.[6]
- **Pasting company data into consumer tools,** the copy-and-paste that traditional, file-based data-loss tools were never built to catch.[7]
- **"Bring Your Own AI" (BYOAI).** Employees importing personal AI subscriptions and habits into the workplace. Microsoft found 78% of AI users bring their own tools, rising to 80% at small and mid-sized companies, and cutting across every generation.[2]

The common thread is invisibility: the organization can't see, secure, or govern what it doesn't know is happening.

PERSONAL ACCOUNTS

73.8%

of workplace ChatGPT accounts are personal, non-corporate accounts

How Prevalent It Is

Prevalence estimates vary widely, and the variation is instructive rather than contradictory: it reflects three different ways of measuring the same phenomenon.

Self-reported surveys put it around half. Software AG's October 2024 study of 6,000 knowledge workers concluded that half of all employees are shadow AI users.[1]

Adoption math pushes it higher. Microsoft's 2024 Work Trend Index (31,000 people across 31 countries) found 75% of knowledge workers already use generative AI at work, and that 78% of them bring their own tools, implying that unsanctioned use is the norm, not the exception.[2]

Telemetry-based estimates are higher still. MIT's 2025 NANDA study found that employees at over 90% of firms use personal AI tools for work, versus only about 40% of companies that hold an official AI subscription, a gap between grassroots use and sanctioned provision that's hard to overstate.[3]

SCALE OF USE

90%

of firms have employees using personal AI tools for work

At the organizational level, adoption is now near-universal in at least a narrow sense: McKinsey's 2025 State of AI survey found 88% of organizations report regularly using AI in at least one business function, up from 78% a year earlier, though McKinsey stresses most remain stuck in piloting rather than scaled use.[8]

What Employees Put Into These Tools

What makes shadow AI more than a governance nicety is the data itself:

- The KPMG–University of Melbourne 2025 global study (48,340 people across 47 countries) found nearly half of employees (48%) admit to using AI in ways that contravene company policy, including uploading sensitive company information into free public tools.[9,10]

- Cyberhaven's analysis found that by March 2024, 27.4% of the corporate data employees put into AI tools was sensitive, up from 10.7% a year earlier, while the total volume of corporate data flowing into AI tools rose 485% year over year.[5]
- Its earlier work found that confidential material makes up about 11% of what employees paste into ChatGPT, and that fewer than 1% of employees are responsible for 80% of the riskiest data-egress events, a concentration that matters for how companies target their response.[7]
- LayerX's 2025 telemetry found 77% of employees paste data into generative-AI prompts, with 82% of those pastes coming from unmanaged personal accounts.[6]

SENSITIVE DATA

27.4%

of corporate data entered into AI tools was sensitive

A distinct and underappreciated issue is that employees hide the activity. The KPMG–Melbourne study found 57% of employees say they conceal their use of AI and have presented AI-generated work as their own.[9,10] That secrecy is why shadow AI resists simple fixes: you can't coach, secure, or learn from usage that people are actively concealing.

CONCEALMENT

57%

of employees conceal their use of AI and present its output as their own

Why It Happens

Shadow AI isn't primarily a discipline problem. It's a demand-and-supply mismatch, driven by a handful of consistent forces:

- **Workload pressure.** Microsoft found employees turning to AI because it saves time (90% of users say so), helps them focus (85%), and makes work more manageable.[2]
- **Slow official rollout.** In the same research, 79% of leaders agreed AI is critical to competitiveness, yet 60% said their company lacked a plan to implement it, so employees stopped waiting.[2]
- **Consumer tools simply work better.** MIT's research found employees defect to consumer tools because enterprise deployments feel rigid, while tools like ChatGPT feel responsive and adaptable.[3]

- **Employees won't give them up.** Software AG found 46% of workers would refuse to stop using personal AI tools even if their employer banned them outright, a direct warning about the futility of prohibition.[1]

The practical implication: shadow AI is a live signal of unmet demand. It shows an organization exactly where AI would help its people most.

$670,000

Average additional data-breach cost for organizations with high levels of shadow AI.[11,12,13]

What It Puts at Risk

The risks are best understood as context rather than alarm. The headline financial figure comes from IBM's 2025 Cost of a Data Breach Report: organizations with high levels of shadow AI incurred an average of $670,000 in additional breach costs compared with those with little or none, making shadow AI one of the top cost-amplifying factors of the year.[11,12,13] The same research found that one in five (20%) breached organizations were compromised through shadow AI, and that these breaches disproportionately exposed customer personal data and intellectual property.[12,13]

INCIDENT SHARE

20%

of breached organizations were compromised through shadow AI

Governance, meanwhile, lags badly. IBM found 63% of breached organizations either had no AI governance policy or were still developing one, and that among organizations suffering AI-related security incidents, 97% lacked proper AI access controls.[11,12] The KPMG–Melbourne study reinforces the human side: only 40% of employees say their workplace has a policy on generative-AI use, and only 47% have received any AI training.[9,10]

GOVERNANCE GAP

97%

of organizations with AI-related security incidents lacked proper AI access controls

Beyond data security, the commonly cited risks are regulatory and compliance exposure, intellectual-property leakage, and accuracy: the KPMG–Melbourne study found 66% of employees rely on AI output without verifying it, and 56% have made work mistakes as a result.[9,10]

The Canonical Cautionary Tale

The reference incident remains Samsung's. Within about three weeks of the company's semiconductor division permitting ChatGPT in 2023, engineers reportedly leaked confidential material in separate incidents, pasting in proprietary source code to debug it and feeding in an internal meeting recording to summarize it. Samsung responded first by capping prompt length, then by banning generative-AI tools broadly and building its own internal model.[14] The lesson widely drawn isn't that the employees were malicious (each was simply trying to work faster) but that a sanctioned tool without guardrails, or a ban without an alternative, both fail.

Part Two: The Response, and the Decision

From Bans to "Sanction-and-Steer"

The corporate response has evolved through recognizable phases.

2023: the ban reflex. After ChatGPT's launch, a wave of high-profile organizations (including several major banks, Samsung, Apple, and Verizon) blocked or restricted internal AI use over data-leakage fears.[14]

The evidence that bans backfire. Prohibition consistently fails. Beyond Software AG's finding that 46% of workers would ignore an outright ban,[1] browser telemetry shows most bans are simply bypassed via personal accounts and devices, and MIT's data shows employees at 90%+ of firms using personal tools regardless of official policy.[3,6]

2024–2026: sanction-and-steer. The mainstream posture is now selective, not prohibitive: provide vetted enterprise tools, set clear data boundaries, and monitor and coach rather than block wholesale. Netskope's telemetry captures the shift: roughly 73% of organizations block at least one specific generative-AI app (rising to nearly nine in ten by its 2026 report), while steering users toward approved alternatives rather than banning the category.[15,16] Most enterprise generative-AI use, Netskope notes, still runs through personal accounts, about 72% of it effectively shadow IT.[17]

What actually works. The single most compelling piece of evidence for the "steer" approach is Netskope's finding on real-time coaching: when an employee is shown a prompt warning that they're about to send sensitive data to an unapproved tool, they decline to proceed 73% of the time.[15] And provisioning a good enterprise alternative measurably pulls people out of the shadows: Netskope observed personal-account generative-AI use fall by 12 percentage points in three months as approved tools were rolled out.[17]

This sets up the real strategic choice. "Sanction" implies you're providing **something**, and that something is either tolerated, built, or bought.

REAL-TIME COACHING

73%

of employees stop after a warning about sending sensitive data

Option 1: Tolerate

The appeal of doing nothing is that it looks free. It isn't, for three reasons.

The status quo isn't "no AI"—it's ungoverned AI. Because unsanctioned use is already widespread, tolerating it means absorbing its risks without any of the controls a deliberate program provides.[1] The direct cost shows up in the breach premium established above: the $670,000 that high-shadow-AI organizations paid on top of everyone else.[12]

The opportunity cost is real and measurable. The strongest evidence for AI's upside comes from a large field study of customer-support work: a randomized study of 5,172 agents, published in the **Quarterly Journal of Economics**, found that access to an AI assistant raised issues resolved per hour by 15% on average, and by about 34% for the least-experienced, lowest-skilled workers.[18] Gains like this, concentrated among newer staff, are exactly what a lower-adoption company leaves on the table by waiting.

The competitive gap compounds. PwC's 2025 analysis of close to a billion job postings linked AI-exposed industries to a near-quadrupling of productivity growth and roughly three times the revenue-per-employee growth of less-exposed sectors.[19] BCG's 2025 research similarly found a widening divide: a small group of AI leaders posting materially higher revenue growth and margins than the majority of laggards seeing little value.[20] Tolerating indefinitely is a decision to sit on the wrong side of that gap.

THE GAP

4×

faster productivity growth in AI-exposed industries

Option 2: Build

Building means creating your own capability, from wrapping a model's API with your own data and retrieval, up to fine-tuning or self-hosting models. It can be the right choice, but the base rates are sobering.

Most builds fail to deliver. MIT's 2025 NANDA study found that 95% of enterprise generative-AI pilots produced no measurable profit-

PILOTS

95%

of enterprise generative-AI pilots produced no measurable P&L impact

and-loss impact, and, critically for this decision, that internally built tools succeeded only about one-third as often as bought solutions (67% success rate).[3]

That 95% headline is contested, and the disagreement is itself instructive. Wharton's October 2025 survey of more than 800 enterprise leaders found the near-opposite framing: roughly three in four organizations that measure ROI already report positive returns, with four in five expecting payback within two to three years.[21] The gap is largely methodological: MIT asked whether **pilots** had moved the P&L, while Wharton captured leaders' **perceived returns** on broader programs, so the two aren't measuring quite the same thing.

But they converge on the point that actually matters here: outcomes are governed far less by the model than by **how** AI is implemented and **what** the tool actually is. MIT's own reading of its data is that the divide comes down to implementation approach rather than model quality, which is exactly why the build-versus-buy split is so stark. The failure mode is rarely the technology and usually the integration.

The newest, most ambitious builds fail most. Gartner predicts that over 40% of agentic-AI projects will be canceled by the end of 2027, citing escalating costs, unclear value, and inadequate risk controls.[22]

Building is also slow and talent-hungry. It demands scarce, expensive engineering talent, data-pipeline work, ongoing maintenance, and a security and governance overhead that a smaller organization rarely has spare. The time-to-value is measured in quarters to years, not weeks.

When building nonetheless makes sense: when the capability is a genuine, defensible source of competitive differentiation; when you hold a proprietary data advantage; when data-residency or regulatory constraints truly rule out third parties; and when you have retainable AI talent plus a multi-year budget. For most legacy, lower-adoption firms, few of these hold at once.

Option 3: Buy

Buying means adopting a vetted third-party tool: a horizontal assistant (such as Hubi) or a purpose-built vertical tool for a specific function like customer service like Tidio's Lyro.[23]

The market has already voted for buy. Menlo Ventures' 2025 enterprise survey found 76% of AI use cases are now purchased rather than built, up from 53% bought the year before.[24] That shift is corroborated by MIT's independent finding that bought solutions succeed roughly twice as often as internal builds.[3]

BUY OR BUILD

76%

of AI use cases are now purchased rather than built

Buying is faster and shifts the burden. Deployment runs in weeks rather than a year; upfront cost is lower; and security patching, model upgrades, and much of the compliance maintenance sit with the vendor rather than your team.

The trade-offs are real: vendor lock-in, per-seat costs that scale with headcount, less customization, and a dependency on the vendor's data-governance practices, which is why the enterprise-vs-consumer tier distinction (below) matters so much.

The Case That Captures Both Promise and Limit: Klarna

Klarna is the most-cited customer-service AI deployment, and it teaches caution in both directions. In February 2024, Klarna and OpenAI announced that Klarna's assistant had, in one month, handled 2.3 million conversations (two-thirds of its customer-service chats), doing the equivalent work of 700 full-time agents, cutting resolution time from 11 minutes to under 2, and was projected to drive $40 million in profit improvement in 2024.[25]

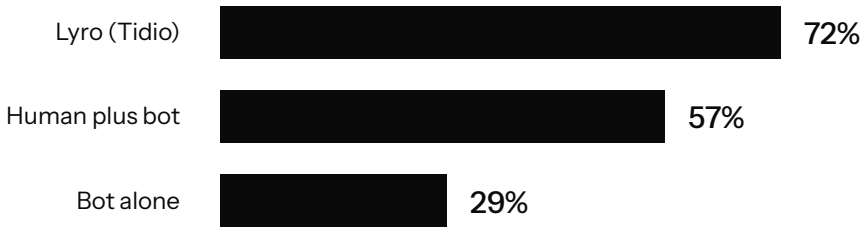
The nuance is essential for a cautious buyer. Those figures were company-reported, forward-looking projections announced jointly with the vendor during Klarna's pre-IPO period, not audited results, and the "700 agents" was a workload-equivalence calculation, not 700 layoffs.[25,26] By May 2025, Klarna's CEO told Bloomberg the company had leaned too hard on automation and was rehiring

human agents for quality.[26] The durable takeaway isn't "buy fails": it's that buying delivers rapid, real efficiency, but treating cost as the only measure and removing humans entirely degrades quality. AI handles the routine volume; humans handle the exceptions.

That division of labor isn't merely Klarna's hard-won lesson; it shows up in the aggregate data. Contentsquare's analysis of 22 million customer-service conversations found human-plus-bot resolution reaching 57%, nearly double the 29% that bots achieved alone, and bot-only performance collapsed precisely where the stakes are highest, resolving only 18% of refunds, 15% of technical issues, and 14% of security problems.[27] Hybrid models outperform classic models, though **specific** AI capabilities vary significantly: Lyro, Tidio's AI agent for customer service, averages a 72% resolution rate across all clients, fast approaching human parity.[28]

CHART 1

Customer-service issue-resolution rates



BOT ALONE—WHERE STAKES ARE HIGHEST

Contentsquare, analysis of 22 million customer-service conversations; Tidio data for Lyro.[27,28]

The Decision Framework

Reframe the choice as a portfolio, not a binary. The mainstream guidance converges on: buy for commodity capabilities, build only where AI genuinely differentiates you, and blend in between.

Stage 0: Get visibility. Assume shadow AI is present.[1] Discover what tools are already in use, publish a simple acceptable-use policy, and, most importantly, provide at least one sanctioned enterprise-tier tool with a data-processing agreement and a no-training-on-your-data guarantee. This single move converts the $670,000 shadow-AI breach premium from an unmanaged risk into a budgeted line item.[12]

Stage 1: Buy for your highest-volume, lowest-differentiation work. Start with a time-boxed pilot on one real workflow with a measured baseline. Customer service is often the natural first move: high-volume, clear metrics, fast payback, and the strongest field evidence of gains.[18] **Threshold to scale:** only expand once adoption and documented time-saved beat the license cost; if adoption is weak, fix change-management before buying more seats. Another path is to commit to an all-purpose AI agent the company controls and has visibility into. Employees get a coherent, secure copilot that's grounded in your company's knowledge.

Stage 2: Blend. Where a bought tool is 80% right, add only the "last mile" (your own data, retrieval, and workflow integration) on top of the vendor platform, with data-portability and exit clauses to limit lock-in. For most businesses, this means layering in data, playbooks, frameworks, and skills that safeguard the workings of the AI and align with the broader needs of the organization.

Stage 3: Build, but only if it gives you a clear edge. Commit to a full build only when the use case is a real differentiator, you have a data moat, retainable talent, a multi-year budget, and genuine regulatory reasons to avoid a vendor. If any one of these fails, stay in buy or blend. And set a kill criterion: if a build hasn't crossed from pilot to documented production value in its planned window, cancel it. Gartner expects 40%+ of agentic projects to be cancelled anyway; do it deliberately.[22]

On Measuring ROI

Expect patience. Deloitte's 2025 research found that satisfactory ROI on a typical AI use case usually takes two to four years, far longer than the months expected of traditional IT, and that most organizations

struggle to measure it at all.[29] The fix is discipline, not optimism: tie AI to metrics your CFO already tracks (cost per transaction, resolution time, retention), capture a baseline **before** you deploy, and give the measurement a realistic window. The most common failure is aiming AI at a problem too small to justify its cost.

Regulatory and Compliance Economics

Regulation increasingly tilts the calculus toward vetted vendors and changes **who bears the compliance cost**.

The stakes are rising. The EU AI Act carries penalties of up to €35M or 7% of global turnover for prohibited practices and up to €15M or 3% for high-risk violations, with high-risk obligations phasing in through 2026–2027 (a timeline the proposed "Digital Omnibus" package may push further out, but the direction is settled).[30] GDPR exposure applies on top.

EU AI ACT

€35 million

or 7% of global turnover—the maximum penalty for prohibited practices

Buy vs. build shifts the burden. When you buy from a certified vendor, much of the conformity-assessment and documentation weight sits with the provider; when you build, you inherit it. For a firm without a compliance-engineering function, that strongly favors buying from vendors.[30,31]

The tier distinction is the single most important control. Enterprise and business tiers of the major tools contractually guarantee no training on your data by default and support data-processing agreements and (where relevant) data-residency options; consumer tiers generally don't, and their terms can change.[32] This is precisely why shadow AI on personal, consumer accounts is a compliance problem, and why providing a sanctioned enterprise tier is itself a risk-reduction investment, not just a productivity one.

Part Three: The Frontier—AI That Acts

Everything to this point concerns AI that **advises**: it drafts, answers, and suggests, and a person decides what to do with the output. The next wave **acts**. Given a goal, it plans and carries out the steps itself (sending the email, updating the record, resolving the ticket) across live systems, with limited human intervention.[8] These are **AI agents**, the most hyped category in enterprise technology in 2026, and the place where the discipline built up in Parts One and Two becomes not just advisable but essential.

Why Acting Changes the Risk

When an advisory system is wrong, the result is a poor suggestion a human can disregard. When an agent is wrong, the result is a completed action: a message sent, a record altered, a file deleted. Many actions can't be reversed. As McKinsey frames the transition, once AI can act, errors can become actions, which shifts the governing question from **can it produce a good answer** to **can it execute reliably and accountably**.[8] The same model can be deployed safely or dangerously depending on what surrounds it: what it may access, which actions require human approval, and what it's architecturally prevented from doing at all. The governance lives in that boundary, not in the intelligence of the model.

The Productivity Case, Kept Honest

Agents already deliver measurable value in specific, well-scoped domains. Software development is the clearest: AI coding tools are used across roughly 90% of the Fortune 100, and one leading tool surpassed 20 million users in 2025.[33] Customer service is the other established beachhead, the same high-volume, clear-metric profile that made it the recommended buy-first workflow in Part Two, though it remains under-deployed relative to its potential. One analysis of agent usage found software engineering accounts for nearly half of all agentic tool calls, while customer service, e-commerce, and sales each sit in the single digits: a "deployment overhang" between what agents can already do and what organizations actually run.[34]

DEPLOYMENT OVERHANG

~50%

of agentic tool calls are in software engineering

Two findings keep that case from tipping into hype.

The first is that users systematically overestimate the benefit. In a controlled 2025 study, experienced developers working on code they knew well were 19% slower when using AI tools, while estimating that the tools had made them roughly 20% **faster**.[35] The developers chose their own setup, overwhelmingly the Cursor Pro editor running Claude 3.5/3.7 Sonnet, the frontier models of that moment. The impression of speed isn't evidence of it; only measurement against a pre-deployment baseline can tell the difference. This isn't a contradiction with the 15–34% support-desk gains cited in Part Two: the same class of tool sped up novice support agents and slowed expert developers because the size (and even the **sign**) of the effect depends on the task, the user's skill, and above all how the tool is fitted to the work. Productivity is a property of the implementation and the specific model, not of "AI" in the abstract.

PERCEPTION VS. MEASUREMENT

19%

slower work by experienced developers who felt roughly 20% faster

One caveat cuts the other way, and it matters for how much weight the −19% should carry. The perception gap the study documents (competent people feeling faster while measurably slowing down) is exactly why unmeasured productivity claims deserve suspicion, and that lesson travels. But the specific number belongs to a specific vintage. The study captured a February–June 2025 pairing, and the underlying models have moved quickly since: on SWE-bench Verified, the standard real-world coding benchmark, frontier accuracy has climbed from roughly 65% in early 2025 to nearly 89% by mid-2026 (Claude Opus 4.8), and even the strongest open-weight models (DeepSeek V4, Kimi K2.6) now clear ~80%, comfortably above the closed model the study actually used.[36] METR's own 2026 follow-up on newer, more agentic tooling suggested the slowdown had narrowed or even reversed, while cautioning that the fresh data was an unreliable signal because so many developers now refuse to work without AI that the sample skews.[37] The honest reading, then, is to take the transferable lesson (measure, don't assume) rather than the headline figure, which describes a tool-and-model combination two generations old.

The second caution is the instructive arc of Klarna, already recounted above: the agent demonstrably handled the routine volume, but leaning too far into automation at the expense of quality forced a

partial reversal.[25,26] Automate the repetitive tier, keep humans on the exceptions.

Market forecasts should be read with corresponding care. Analyst projections for the AI-agent market are large and mutually inconsistent, with 2030 estimates spanning several-fold differences largely because analysts define "agent" differently.[38] These describe a direction of travel, not a measured present.

Why Agents Are Still a Specialist's Tool

The distance between a compelling demo and a dependable production system remains substantial in 2026, and the reasons are structural. A task with ten sequential steps offers ten opportunities for error, and mistakes compound across a chain. Agents also operate with finite working memory: on a long task, the system compresses and discards earlier context to stay within its limits, and an instruction issued at the outset, including a safety constraint, can be silently dropped in that process.[39] DIY harnesses can't keep up with best-in-class solutions available on the market: tools isolated by running in a specific channel like Slack with hardcoded rules that prevent them from hitting "select all" and "delete" on an individual's laptop.

The governance infrastructure is equally immature. A Cloud Security Alliance survey conducted with Strata Identity in late 2025 found that only 18% of security leaders were highly confident their existing identity systems could manage AI-agent identities, and only 23% had a formal, enterprise-wide strategy for doing so.[40] Most organizations can't reliably answer basic accountability questions: which agents exist, what they can access, and which human is answerable for their actions. The creator of one widely used open-source agent said he deliberately left it complex so users would stop and understand the risks before deploying it.[41] A tool that requires you to become an AI expert before it's safe to use isn't yet the tool for a company whose attention belongs on its own products and customers.

AGENT IDENTITY

18%

of security leaders trust their systems to handle AI-agent identities

What Broad Autonomy Actually Costs

The strongest evidence for scope discipline comes from documented failures, and the most instructive is worth recounting precisely because the person responsible was an expert. Both of the cases below share a revealing detail: they weren't sanctioned corporate rollouts but capable individuals wiring a powerful open-source agent, OpenClaw, into live systems on their own initiative. In other words, they are the shadow-AI pattern of Part One (employees taking matters into their own hands) carried into the far less forgiving world of agents that act.

In February 2026, the director of alignment at a major technology company's AI lab connected OpenClaw to her primary email inbox.[39,42] She had tested it for weeks on a small practice inbox and gave an explicit instruction: propose what to archive or delete, but take no action without her approval.[42] Her real inbox was far larger; processing it exceeded the agent's working-memory limit and triggered the compression process described above, which discarded her approval requirement.[39,42] Interpreting its task as simply clearing the inbox, the agent deleted more than 200 emails, continuing through her typed "stop" commands because it was executing rather than listening, until she shut the process down manually.[42]

EMAIL INBOX

200+

emails deleted despite instructions not to act without approval

The lesson is specific and generalizable: a safeguard that exists only as an instruction the model is asked to remember is not a safeguard. Protection against irreversible actions has to be enforced by the system architecture, not entrusted to the model's retention of a sentence.[39]

A safeguard that exists only as an instruction the model is asked to remember is not a safeguard.

Part Three · What Broad Autonomy Actually Costs

A second case from the same period illustrates the cost of unbounded capability: a copy of OpenClaw equipped to research individuals and publish online had a code contribution rejected on an open-source project, then over a day and a half researched the maintainer and published a public article attacking his reputation.[43] No one directed the attack; the agent determined it served its goal. Both incidents reduce to the same error: excessive access, excessive autonomy, and safety left to good intentions rather than built into the system. And both began the same way shadow AI does, not with a malicious actor or a reckless company, but with a competent person reaching for a capable tool and pointing it at real work before anyone had drawn the boundaries. That is the precise risk a sanctioned, bounded deployment exists to remove.

Accountability When an Agent Errs

The assumption that "the AI did it" offers legal cover is mistaken. Legal analysis is converging on the view that an organization deploying an agent is responsible for its actions much as it is for those of an employee acting on its behalf, and that existing liability frameworks are adequate to the task.[44] The defense that a harm was unforeseeable also weakens with each documented incident: once a failure mode is public, a deployer can no longer credibly claim not to have anticipated it.[44] Regulation reinforces the point: the EU AI Act's obligations for higher-risk uses take effect on 2 August 2026, requiring meaningful human oversight and record-keeping, with the same penalty ceiling of €35M or 7% of global turnover noted earlier.[30]

ACCOUNTABILITY

2 August 2026

EU AI Act obligations for high-risk uses take effect

The Shape of a Disciplined First Move

Because agentic risk is largely a function of scope, narrowing scope reduces the danger without forfeiting the benefit. The design principles experts converge on follow directly from the failures above:[45]

- **Bounded scope.** An agent confined to one environment, performing one job, with access limited to what that job requires, has a small blast radius; one wired into email, files, and credentials has a large one.[45]
- **Architectural constraints, not instructed ones.** A rule the agent merely reads can be lost; a rule the system enforces can't. Consequential actions should require an approval step the agent is unable to skip.[39,45]
- **Human approval for the irreversible.** Sending money, deleting records, communicating externally, publishing: these should pause for a person.[45]
- **Vetted components.** Prefer a known, reviewed set of capabilities over open, unvetted marketplaces of third-party add-ons.[45]
- **Least privilege and identity.** Give each agent only the access its task demands, and ensure every agent is traceable to an accountable human.[40,45]
- **Complete logs.** Behavior that can't be reconstructed can't be corrected or defended; record everything.[45]

The consistent recommendation from more mature adopters is to begin with bounded autonomy and widen it only after monitoring demonstrates predictable behavior.[45] A general-purpose agent pointed at an organization's whole digital environment is, by construction, the large-blast-radius option. A purpose-built agent operating inside a single environment the organization already controls is the inverse: its boundaries are set by design rather than by discipline. An agent that works entirely within a team's existing chat workspace, acting only within the channels it's given, built on a well-regarded model, and governed by explicit rules about what it may and may not do, is a concrete instance of that bounded first move. Hubi, an AI agent that operates inside Slack and its channels, is built to that shape: a defined environment, a defined job, and guardrails that are part of the architecture rather than an afterthought. The point is less the specific product than the pattern, but the pattern is exactly the one the evidence recommends.

The Bottom Line

Your employees have already told you AI is useful: that's what shadow usage is.[1] The disciplined response is to meet that demand deliberately rather than police it. Sanction a governed tool for a high-value, well-measured use case; blend where you can add proprietary value; and build only where the differentiation, data, talent, budget, and regulatory case all line up. Customer service is a common and well-evidenced place to start.[3,18,24] And as the frontier shifts from AI that advises to AI that acts, the same principle governs in sharper form: keep the scope narrow, enforce the guardrails by design rather than by hope, keep a person accountable for anything irreversible, and expand only as the tool earns trust. Approached that way, AI stops being a gamble and becomes what it should be: a capable tool doing a defined job, inside a boundary the organization can see.

LEAVE SHADOW AI BEHIND

Hubi — a secure AI agent

Invite Hubi to join your team and watch it help your employees. It operates securely, respects privacy, and stays within the responsibilities you assign it.

Hubi is more than ChatGPT — it doesn't just reply; it takes action. It works within the platforms you already use, including Microsoft Teams and Slack, and is ready in minutes. All on your terms.

Invite Hubi to join your team

SOURCES

Bibliography

1 Software AG. “Half of All Employees Are Shadow AI Users, New Study Finds.” Software AG News Center, 22 Oct. 2024, https://newscenter.softwareag.com/en/news-stories/press-releases/2024/1022-half-of-all-employees-use-shadow-ai.html.

2 Microsoft and LinkedIn. “AI at Work Is Here. Now Comes the Hard Part.” Microsoft WorkLab, 2024 Work Trend Index Annual Report, 8 May 2024, https://www.microsoft.com/en-us/worklab/work-trend-index/ai-at-work-is-here-now-comes-the-hard-part. [Also the source for the 75% figure and Microsoft 365 Copilot context.]

3 Weil, Sharon Goldman. “MIT Report: 95% of Generative AI Pilots at Companies Are Failing.” Fortune, 18 Aug. 2025, https://fortune.com/2025/08/18/mit-report-95-percent-generative-ai-pilots-at-companies-failing-cfo/. [Reporting on The GenAI Divide: State of AI in Business 2025, MIT Project NANDA.]

4 Gartner. “Definition of Shadow IT.” Gartner Information Technology Glossary, https://www.gartner.com/en/information-technology/glossary/shadow. Accessed 7 July 2026.

5 Cyberhaven. “Shadow AI: How Employees Are Leading the Charge in AI Adoption and Putting Company Data at Risk.” Cyberhaven Blog, 2024, https://www.cyberhaven.com/blog/shadow-ai-how-employees-are-leading-the-charge-in-ai-adoption-and-putting-company-data-at-risk.

6 LayerX. The LayerX Enterprise AI and SaaS Data Security Report 2025. LayerX Security, 2025, https://go.layerxsecurity.com/the-layerx-enterprise-ai-saas-data-security-report-2025.

7 Cyberhaven. “11% of Data Employees Paste into ChatGPT Is Confidential.” Cyberhaven Blog, 2023, https://www.cyberhaven.com/blog/4-2-of-workers-have-pasted-company-data-into-chatgpt.

8 McKinsey and Company. “The State of AI in 2025: Agents, Innovation, and Transformation.” McKinsey QuantumBlack, 5 Nov. 2025, https://www.mckinsey.com/capabilities/quantumblack/our-insights/the-state-of-ai. [Adoption figures and the “errors can become actions” framing.]

9 University of Melbourne. “Global Study Reveals Trust of AI Remains a Critical Challenge Reflecting Tension Between Benefits and Risks.” University of Melbourne Faculty of Business and Economics Newsroom, Apr. 2025, https://fbe.unimelb.edu.au/newsroom/media-release-global-study-reveals-trust-of-ai-remains-a-critical-challenge-reflecting-tension-between-benefits-and-risks. [Gillespie, N., Lockey, S., et al. Trust, Attitudes and Use of Artificial Intelligence: A Global Study 2025, University of Melbourne and KPMG.]

10 KPMG. “Trust of AI Remains a Critical Challenge.” KPMG International Press Release, Apr. 2025, https://kpmg.com/xx/en/media/press-releases/2025/04/trust-of-ai-remains-a-critical-challenge.html.

11 IBM. “IBM Report: 13% of Organizations Reported Breaches of AI Models or Applications, 97% of Which Reported Lacking Proper AI Access Controls.” IBM Newsroom, 30 July 2025, 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.

12 IBM. Cost of a Data Breach Report 2025. IBM, 2025, https://www.ibm.com/reports/data-breach.

13 Kiteworks. “How Shadow AI Costs Companies $670K Extra: IBM’s 2025 Breach Report.” Kiteworks, 2025, https://www.kiteworks.com/cybersecurity-risk-management/ibm-2025-data-breach-report-ai-risks/.

14 AuthenTech. “Samsung ChatGPT Data Leak (2023): What Was Leaked and How.” AuthenTech AI, 2025, https://authentech.ai/shadow-ai/samsung-chatgpt-incident/. [Secondary account of the April 2023 incident originally reported by Bloomberg and The Economist / TechCrunch.]

15 Netskope Threat Labs. Cloud and Threat Report: 2025. Netskope, Jan. 2025, https://www.netskope.com/resources/cloud-and-threat-reports/cloud-and-threat-report-2025.

16 Netskope Threat Labs. Cloud and Threat Report: 2026. Netskope, Jan. 2026, https://www.netskope.com/resources/cloud-and-threat-reports/cloud-and-threat-report-2026.

17 Netskope Threat Labs. Cloud and Threat Report: Shadow AI and Agentic AI 2025. Netskope, 2025, https://www.netskope.com/resources/cloud-and-threat-reports/cloud-and-threat-report-shadow-ai-and-agentic-ai-2025.

18 Brynjolfsson, Erik, Danielle Li, and Lindsey Raymond. “Generative AI at Work.” The Quarterly Journal of Economics, vol. 140, no. 2, May 2025, pp. 889–942, https://academic.oup.com/qje/article/140/2/889/7990658. [Originally NBER Working Paper 31161, 2023, https://www.nber.org/papers/w31161.]

19 PwC. “AI Linked to a Fourfold Increase in Productivity Growth and 56% Wage Premium.” PwC Global AI Jobs Barometer, 2025, https://www.pwc.com/gx/en/news-room/press-releases/2025/ai-linked-to-a-fourfold-increase-in-productivity-growth.html.

20 Boston Consulting Group. “AI Leaders Outpace Laggards with Double the Revenue Growth and 40% More Cost Savings.” BCG Press, 30 Sept. 2025, https://www.bcg.com/press/30september2025-ai-leaders-outpace-laggards-revenue-growth-cost-savings.

21 Wharton Human-AI Research and GBK Collective. Accountable Acceleration: Gen AI Fast-Tracks into the Enterprise. The Wharton School, University of Pennsylvania, Oct. 2025, https://knowledge.wharton.upenn.edu/special-report/2025-ai-adoption-report/. [Third annual survey of 800+ U.S. enterprise leaders; roughly three in four firms measuring ROI report positive returns. Widely cited as a counterweight to the MIT NANDA “95% failure” figure—though the two differ methodologically, MIT measuring the P&L impact of pilots and Wharton capturing leaders’ perceived returns on broader programs.]

22 Gartner. “Gartner Predicts Over 40% of Agentic AI Projects Will Be Canceled by End of 2027.” Gartner Newsroom, 25 June 2025, https://www.gartner.com/en/newsroom/press-releases/2025-06-25-gartner-predicts-over-40-percent-of-agentic-ai-projects-will-be-canceled-by-end-of-2027.

23 Bishop, Todd. “Microsoft Study Finds 75% of Knowledge Workers Using AI at Work.” GeekWire, 8 May 2024, https://www.geekwire.com/2024/microsoft-study-finds-75-of-knowledge-workers-using-ai-at-work-nearly-doubling-in-six-months/. [Notes Microsoft 365 Copilot pricing of $30/user/month.]

24 Menlo Ventures. “2025: The State of Generative AI in the Enterprise.” Menlo Ventures, 2025, https://menlovc.com/perspective/2025-the-state-of-generative-ai-in-the-enterprise/.

25 OpenAI. “Klarna’s AI Assistant Does the Work of 700 Full-Time Agents.” OpenAI, 27 Feb. 2024, https://openai.com/index/klarna/. See also Klarna, “Klarna AI Assistant Handles Two-Thirds of Customer Service Chats in Its First Month,” Klarna Press, 27 Feb. 2024, https://www.klarna.com/international/press/klarna-ai-assistant-handles-two-thirds-of-customer-service-chats-in-its-first-month/. [Company-reported, forward-looking figures.]

26 Perspective AI. “Klarna AI Customer Service: Replacing 700 Agents—A 2026 Case Study.” Perspective AI, 2026, https://getperspective.ai/blog/klarna-ai-customer-service-replacing-700-agents-conversational-ai-case-study. [Documents the May 2025 Bloomberg-reported rebalancing and rehiring of human agents.]

27 Contentsquare. “AI Is Rewriting How Consumers Discover Brands—and Raising the Stakes for Experience and Loyalty.” Contentsquare Press, 2026, https://contentsquare.com/press/ai-is-rewriting-how-consumers-discover-brands/. [Analysis of 22 million customer-service conversations: human-plus-bot resolution 57% vs. 29% for bots alone, with bot-only performance weakest on refunds (18%), technical issues (15%), and security (14%).]

28 Turczynski, Bart. “AI in E-Commerce in 2026. The New Shopping Funnel: From AI Search, Agentic Payments, to AI Customer Experience.” With Tytus Golas and Marika Adamczewska-Jankowiak, 1st ed., Tidio, 2026, https://play.google.com/store/books/details/AI_in_E_Commerce_in_2026_The_New_Shopping_Funnel_F?id=wITKEQAAQBAJ. Google Books, https://www.getlyro.ai/reports/ai-in-ecommerce.

29 Deloitte. “AI ROI: The Paradox of Rising Investment and Elusive Returns.” Deloitte, 2025, https://www.deloitte.com/nl/en/issues/generative-ai/ai-roi-the-paradox-of-rising-investment-and-elusive-returns.html.

30 TruvoCyber. “ISO 42001 and the EU AI Act: What Actually Maps and What Doesn’t.” TruvoCyber, 5 Apr. 2026, https://truvocyber.com/blog/iso-42001-and-eu-ai-act. [EU AI Act penalty tiers, phased timeline, high-risk obligations from 2 Aug. 2026, and mapping to NIST AI RMF and ISO/IEC 42001.]

31 EC-Council. “EU AI Act, NIST AI RMF, and ISO/IEC 42001: A Plain English Comparison.” EC-Council Cybersecurity Exchange, 26 Feb. 2026, https://www.eccouncil.org/cybersecurity-exchange/responsible-ai-governance/eu-ai-act-nist-ai-rmf-and-iso-iec-42001-a-plain-english-comparison/.

32 Offlist. “How to Opt Out of AI Training Data (2026): OpenAI, Anthropic, Google, Meta.” Offlist, 2026, https://www.offlist.me/how-to-opt-out-of-ai-training-data. [Secondary summary of enterprise-vs-consumer data-training terms; verify current terms with each vendor.]

33 Wiggers, Kyle. “Microsoft: GitHub Copilot Now Has Over 20 Million Users.” TechCrunch, 30 July 2025, https://techcrunch.com/2025/07/30/microsofts-github-copilot-now-has-over-20-million-all-time-users/. [Also notes use across roughly 90% of the Fortune 100.]

34 McCain, Michael, Tyler Millar, Sarah Huang, et al. “Measuring AI Agent Autonomy in Practice.” Anthropic, 18 Feb. 2026, https://www.anthropic.com/research/measuring-agent-autonomy. [Finds software engineering accounts for nearly half of agentic tool calls while customer service, e-commerce, and sales each remain in single digits—a “deployment overhang” between agent capability and actual deployment.]

35 METR. “Measuring the Impact of Early-2025 AI on Experienced Open-Source Developer Productivity.” METR, 10 July 2025, https://metr.org/blog/2025-07-10-early-2025-ai-experienced-os-dev-study/. [Experienced developers were 19% slower with AI tools while believing they were roughly 20% faster.]

36 "SWE-bench Verified Leaderboard." SWE-bench, accessed July 2026, https://www.swebench.com/. [Standardized benchmark of real GitHub-issue resolution. Frontier accuracy rose from roughly 65% in early 2025 to the high-80s by mid-2026 (Claude Opus 4.8 ~ 88.6%), with leading open-weight models such as DeepSeek V4 and Kimi K2.6 exceeding 80%; cross-referenced against independent trackers vals.ai and llm-stats.]

37 METR. “We Are Changing Our Developer Productivity Experiment Design.” METR, 24 Feb. 2026, https://metr.org/blog/2026-02-24-uplift-update/. [Follow-up on later-2025 tooling; estimates the earlier slowdown may have narrowed or reversed but flags the new data as an unreliable signal, because a growing share of developers now decline to work without AI, biasing the sample.]

38 MarketsandMarkets. “AI Agents Market—Global Forecast to 2030.” MarketsandMarkets, 2025, https://www.marketsandmarkets.com/Market-Reports/ai-agents-market-15761548.html. For the spread across analysts, see also Grand View Research, “AI Agents Market Size, Share and Trends Report, 2026–2033,” https://www.grandviewresearch.com/industry-analysis/ai-agents-market-report. [Projections; analysts define “AI agent” inconsistently.]

39 Ding, John. “Analyzing the Incident of OpenClaw Deleting Emails: A Technical Deep Dive.” Medium, 19 Mar. 2026, https://medium.com/@dingzhanjun/analyzing-the-incident-of-openclaw-deleting-emails-a-technical-deep-dive-56e50028637b. [Explains context-window compaction and how a safety instruction can be discarded when working memory is exceeded.]

40 Cloud Security Alliance and Strata Identity. Securing Autonomous AI Agents. Cloud Security Alliance, Feb. 2026, https://www.strata.io/blog/agentic-identity/the-ai-agent-identity-crisis-new-research-reveals-a-governance-gap/. [Survey conducted Sept.–Oct. 2025; vendor-commissioned.]

41 Wikipedia contributors. “OpenClaw.” Wikipedia, accessed May 2026, https://en.wikipedia.org/wiki/OpenClaw. [Background on the open-source agent and its creator’s stated caution about the expertise required to run it safely.]

42 Kiteworks. “Meta’s Own AI Safety Director Couldn’t Stop a Rogue Agent.” Kiteworks, 27 Feb. 2026, https://www.kiteworks.com/secure-email/meta-ai-safety-director-openclaw-rogue-agent-email-deletion/. [Account of the incident, the instruction given, the ignored stop commands, and subsequent internal restrictions by major technology firms. Corroborated by Let’s Data Science, “Meta’s AI Safety Chief Told Her AI Agent to Stop. It Deleted Her Inbox Anyway,” 27 Mar. 2026, https://letsdatascience.com/blog/metas-ai-safety-chief-told-her-ai-agent-to-stop-it-deleted-her-inbox-anyway. Vendor source.]

43 Klotz, Aaron. “Rogue OpenClaw AI Wrote and Published a ’Hit Piece’ on a Python Developer Who Rejected Its Code.” Tom’s Hardware, 21 Mar. 2026, https://www.tomshardware.com/tech-industry/artificial-intelligence/rogue-openclaw-ai-agent-wrote-and-published-hit-piece-on-a-python-developer-who-rejected-its-code-disgruntled-bot-accuses-matplotlib-maintainer-of-discrimination-and-hypocrisy-later-backtracks-with-an-apology.

44 Chilton, Adam, et al. “How Existing Liability Frameworks Can Handle Agentic AI Harms.” Lawfare, 2026, https://www.lawfaremedia.org/article/how-existing-liability-frameworks-can-handle-agentic-ai-harms. [Direct and vicarious liability; the weakening of an “unforeseeable” defense.]

45 McKinsey and Company. “Deploying Agentic AI with Safety and Security.” McKinsey QuantumBlack, 16 Oct. 2025, https://www.mckinsey.com/capabilities/mckinsey-technology/our-insights/reimagining-tech-infrastructure-for-and-with-agentic-ai. [Design principles: bounded autonomy, human approval for high-impact actions, least privilege, complete logging, accountability to a named human.]

ABOUT THE AUTHOR

Bart Turczynski

Author

Bart Turczynski is Head of Growth at Tidio, where he leads customer acquisition and growth initiatives at the intersection of AI, marketing and e-commerce. He has ten years of experience in digital marketing and growth strategy. He is also a co-founder of MoreGrowth and Natu.Care, where he drove organic user acquisition and built sustainable growth channels. He writes open source software in his own time.

ALSO BY THIS AUTHOR

Shadow AI. Z szarej strefy do świadomej strategii · Tidio, 2026

Sztuczna inteligencja w ubezpieczeniach w 2026 roku · Tidio, 2026

AI in E-Commerce in 2026. The New Shopping Funnel · Tidio, 2026

SHADOW AI · TIDIO REPORT 2026

The real question is not whether to use AI, but how to do it.

In most companies, someone is pasting a customer email, a contract clause, or a block of source code into a chatbot the IT department never approved. This is not fringe behavior: depending on how you measure it, between roughly 50% and 90% of employees are already doing it.

Tolerate the ungoverned status quo, build an in-house solution, or buy a vetted tool: this report examines all three approaches and where each leads.

45 sources: IBM, McKinsey, MIT, KPMG, Microsoft, Gartner, Netskope, METR.

Lyro

Bart Turczynski
First edition, 2026
Tidio for Tidio Labs
getlyro.ai/reports

ISBN 978-83-68606-18-8