Konstera

REPORT

The Shadow AI *Report 2026.*

Every department in your business is now deploying AI. Nobody is governing it as a whole. This report measures the sprawl, names the four hidden costs it's producing, and gives mid-market CIOs a 90-day plan to install the control plane before incumbents catch up.

30-minute read · 12 sections · ~15 pages · Published October 2026

Free · ungated · no sign-up

REPORT

Contents

– Foreword
01 The Shadow AI Decade
02 By the numbers
03 The mid-market delta
04 Retail, wholesale, distribution
05 The Four Hidden Costs
06 Three stories
07 The regulatory horizon
08 Why policies alone fail
09 The control-plane framework
10 The 90-day plan
11 2027 predictions
12 Methodology & sources

Why this report exists.

In the twelve months to October 2026, we watched the same conversation play out in a dozen mid-market businesses. The marketing team had bought three AI tools. Finance had a GPT subscription paid for as "software." Customer service had deployed a vendor's AI product without IT's involvement. The CIO found out at quarter-end, when the CFO asked why the technology bill had grown 14% without a new project in sight.

Everyone in the room agreed this was a problem. Nobody had a shared vocabulary for talking about it. We wrote this report to give them one — and to be plain about what it costs when a business doesn't install the control plane every other operational discipline (data, integration, security) eventually had to.

The next fourteen pages contain twenty cited statistics, three case vignettes, a four-part costing framework, a regulatory timeline, and a 90-day plan. There is no vendor pitch — Konstera is mentioned in the methodology and the acknowledgements. The frameworks (*Four Hidden Costs of AI Sprawl*, *AgentOps*, *The Context Debt Curve*) are the ones we use internally; we've put them in the open so the category's shared language can start converging.

— The Konstera team, October 2026

---

01 The Shadow AI Decade

Shadow IT described the 2010s: SaaS tools purchased outside the CIO's purview, invisible to procurement, invisible to security, invisible to the audit trail. Platforms like Okta and Netskope were built to pull that sprawl back into one view. The playbook is now settled.

Shadow AI is the 2020s version — but it is faster, cheaper, and in important ways more consequential. Every generative-AI vendor has a self-service free tier. Every credit-card SaaS category now has an AI feature. The procurement moat that eventually caught shadow IT does not exist for AI; employees click "sign in with Google" and are productive in minutes.

Three forces are compounding Shadow AI in mid-market businesses specifically:

- **Consumerised procurement.** Many AI tools cost less than a software-approval threshold. Head of Marketing signs up for ChatGPT Team, expenses it. No trigger fires.
- **Departmental P&L autonomy.** In a £200M business, each department head has sign-off on a working budget. Each uses it to try one AI tool. Six months later there are eight.
- **The absence of a control plane.** Shadow IT got solved by identity (SSO), network (ZTNA), and data classification (DLP). AI has none of the equivalent layers in most mid-market stacks today. There is nothing to see through.

This report measures the shape of the problem and names the control plane that resolves it.

02 By the numbers

Twenty statistics, published by independent researchers, that jointly describe the current state. Each is annotated with its source and relevance to a mid-market retail, wholesale or distribution buyer. (Full citations in section 12.)

Adoption, by function

75%

of knowledge workers use generative AI at work

— Microsoft/LinkedIn Work Trend Index 2024. 46% started in the last six months.

78%

of organisations use AI in at least one business function

— McKinsey State of AI 2024. Up from 55% a year earlier.

65%

of organisations use genAI in two or more functions

— McKinsey 2024. Almost double 2023 (33%). Sprawl is already the modal state.

42%

is the top function-level adoption rate: marketing & sales

Followed by product (36%), IT (31%), service ops (28%). Supply chain lags at ~12%.

Shadow AI prevalence

644%

year-on-year growth in enterprise genAI traffic

— Palo Alto Networks 2025 State of Generative AI.

75%

of genAI logins use personal credentials

— Palo Alto 2025. Three-quarters invisible to corporate SSO.

66

distinct genAI apps in the average enterprise

— Palo Alto 2025. About 10% are classified as high-risk.

96%

of organisations show active genAI usage

— Netskope Cloud & Threat Report 2024 (up from 74% in 2023).

Spend and governance gap

$644B

forecast global enterprise genAI spend in 2025

— Gartner, Aug 2024. Up 76% YoY.

51%

of CFOs lack clear line-of-sight into AI spend

— Deloitte CFO Signals 2024.

23%

of organisations are "highly prepared" for genAI risk

— Deloitte Q2 2024. 39% have no formal AI policy at all.

44%

have already experienced negative consequences from genAI

— McKinsey 2024. Inaccuracy, cybersecurity risk, IP infringement.

55% vs 29%

have policies vs. have operationalised oversight

— McKinsey 2024. The gap is where every incident of the last three years lived.

38%

of employees share confidential data with AI without permission

— CybSafe 2024 (self-reported; actual rate higher per telemetry studies).

Mid-market gap & retail signal

72% vs 31%

of enterprises vs mid-market have cross-functional AI steering

— BCG AI at Work 2024. The governance gap is specifically mid-market.

69%

of retailers used some form of AI in 2024

— NRF / IHL State of Retail Technology 2024. Up from 40% in 2023.

73% vs 18%

of retail execs say AI is top-3 priority vs have "fit for AI" data

— Deloitte 2025 Retail Industry Outlook. The data-foundation gap is where context fails to compound.

5x

LLM API spend growth, Q4 2023 to Q4 2024, median Fortune 500

— a16z Enterprise LLM Report 2024. Mid-market lags enterprise 12–18 months. The invoice shock is still ahead for most £100M retailers.

03 The mid-market delta

A £150M retailer in Yorkshire and a $5B specialty chain in Michigan both have an AI sprawl problem. The incident patterns are the same. The remediation playbook is not. Mid-market is uniquely exposed for three reasons that compound.

Smaller governance headcount, same regulatory burden

The EU AI Act applies to any business selling into the EU, at any revenue size. GDPR applies the same way. Mid-market retailers have one person doing information security — maybe one and a half. The policies that enterprises write with a 25-person team of GRC analysts, mid-market has to produce with a quarter of the time.

Fragmented procurement authority

Enterprise P&Ls route software decisions through IT. Mid-market P&Ls route them through departments. Marketing has a working budget; it uses it. Finance has one; it uses it. By year end there is a long tail of AI subscriptions totalling a non-trivial sum, and no one owns the sum.

Higher tool density per revenue dollar

Our observation: a £100M business running six AI tools has a higher ratio of tools-to-revenue than a $5B business running thirty. The sprawl density is *higher* in mid-market — roughly 2× per £ of revenue, per our observational signal. The reason is that enterprise procurement eventually consolidates; mid-market procurement doesn't have the lever.

None of this is an argument for mid-market businesses to buy the enterprise playbook. The expensive consolidated platforms sold to the Fortune 100 (Palantir Foundry, Databricks) are architecturally mismatched to the mid-market business reality. The answer is a mid-market-shaped control plane — the subject of section 09.

04 Retail, wholesale, distribution

Across the £50M–£500M retail, wholesale and distribution businesses we observe, three sector-specific sprawl patterns recur.

The seasonal P&L shock

Retail AI spend scales with peak traffic. The customer service AI runs 3× volume during Q4. The campaign-generation AI runs 10× during pre-Peak marketing cycles. Mid-market CFOs who budgeted on H1 steady-state face December invoices they didn't forecast — and no attribution to tell them which agent drove

which invoice line.

The buy-side split

Buying and merchandising teams buy point AI tools for product content and demand forecasting. Marketing buys separate AI tools for campaigns and personalisation. Customer service buys another. Each team is right that AI helps their job; collectively, the business is paying for overlapping capabilities that don't compose. The context — SKUs, suppliers, customer identity — is replicated inconsistently in each tool.

The data-foundation gap

The retail industry's own research is unambiguous: 73% of retail execs say AI is a top-3 strategic priority; 18% think their data foundation is fit for AI. The gap is where every multi-tool failure happens. Even perfectly-chosen agents can't compound across a stack whose master data disagrees with itself.

Wholesale and distribution add two further patterns: EDI integrations where agent errors cascade through supplier systems, and the fact that AI vendors have barely begun to target distribution-specific workflows (pick-path optimisation, exception triage, pricing-tier management) — so wholesale IT ends up building bespoke, which has its own governance costs.

---

05 The Four Hidden Costs

When a mid-market retailer tots up its "AI spend," the number it produces is almost always wrong, and almost always too low.

It is too low because of the obvious things — the Claude Pro subscriptions reimbursed as "software," the Copilot licences bundled into an E5 renewal nobody renegotiated, the £2,400 ChatGPT Team contract the Head of Buying signed on a corporate card. One £180M home-furnishings retailer discovered during a budget review that it was paying £47,000 a month across fourteen separate frontier-model accounts, only one of which IT had sanctioned.

But the duplicated spend — the invoices nobody owns — is only the first of four costs. It is the one that shows up on a ledger, so it is the one that gets attention. The other three are harder to see and considerably more expensive.

Cost 1 — Duplicated spend

The same capability bought three times. Marketing's AI vendor. Finance's AI vendor. The AI-feature upsell in the ERP contract. None of them know about the others; all of them bill. Mid-market CFOs who audit discover this first because it is the only cost that lives on a ledger.

Cost 2 — Fragmented context

Every AI tool becomes, implicitly, a container for business knowledge: SKUs, suppliers, lead times, pricing rules, customer histories. In a single-tool world, that knowledge accumulates and compounds. In a fourteen-tool world, each copy drifts. The merchandiser's Copilot still thinks the lead time is sixty days. The buyer's custom GPT has been updated to ninety. The customer-service bot is confidently telling shoppers "ten weeks."

We call this *context debt*. Unlike the technical variety, context debt compounds exponentially: n tools produce n^2 reconciliation surfaces. A retailer running the industry-average sixty-six genAI applications doesn't have sixty-six integration problems — it has over four thousand.

Cost 3 — Governance debt

McKinsey's 2024 survey found that 55% of organisations have implemented restrictions on sensitive data use in generative AI, but only 29% have embedded AI oversight into standard operating procedures. The gap — a policy on a Confluence page versus a control that executes at the moment of use — is where every reported AI incident of the last three years has lived.

Samsung's engineers knew the policy; they pasted the source code anyway. Air Canada's customer-service team knew bots could hallucinate; the bot committed to a refund policy that didn't exist and a tribunal ordered the airline to honour it. In both cases, the policy was fine. The *enforcement surface* was missing. Mid-market businesses are especially vulnerable because they have the same regulatory burden as enterprises but a tenth of the governance headcount.

Cost 4 — Trust leakage

The slowest and the most strategic of the four. When a customer gets a different answer from your chatbot than from your call centre, when your merchandiser's forecast disagrees with your buyer's commitment, when a supplier is told two different lead times in the same week by two different automated agents — the business loses credibility, one interaction at a time. Nobody writes a board paper titled "Trust Lost This Quarter." But survey a mid-market retailer's top-50 customers at twelve-month intervals and the shift shows up in NPS, in retention cohort curves, and eventually in the P&L.

The four costs do not sit in parallel. They cascade. Duplicated spend buys you fragmented context. Fragmented context defeats your governance controls, because the people writing the policy don't know what tools they're writing it for. Weak governance produces the incidents that leak trust. By the time a mid-market board asks "what is our AI exposure?" the honest answer involves all four, and the honest answer takes a year to produce.

06 Three stories

Three case vignettes from businesses we have talked to. Anonymised to protect commercial sensitivity. Each illustrates one of the four costs in action.

CASE VIGNETTE

The invoice that nobody owned

£180M home-furnishings retailer · Midlands · 2025

During the FY25 budget review, finance discovered it had been paying £47,000 a month across fourteen separate OpenAI, Anthropic, and Perplexity Pro accounts — none of them on a corporate card. Marketing had three. Buying had two. The customer-service director had set up another. The CEO's EA had a Claude Pro subscription she'd been reimbursed for as 'software.' IT had sanctioned exactly one of the fourteen. The CFO described discovering the fragmented spend as 'finding a shadow line of credit we didn't know we'd opened.' Consolidating onto two enterprise contracts cut the bill by 38% — but the exercise took three months and required a cross-functional working group that didn't previously exist.

CASE VIGNETTE

Three bots, three stories

£240M fashion wholesaler · 2024–2025

Over 18 months the business built three separate AI-assisted tools: a merchandiser copilot in Excel, a customer-service triage bot (Intercom Fin), and a buyer's assistant (a bespoke ChatGPT GPT with supplier contracts uploaded). During AW25 planning, a key supplier's lead-time changed from 60 to 90 days. The buyer updated the GPT. The merchandiser's Copilot kept assuming 60. The CS bot continued promising customers 'ships within 10 weeks of order.' The disconnect was only caught six weeks later when a buying-team analyst noticed the OTB forecast and the CS complaint volume didn't reconcile. The director of operations called it 'three AI tools, three competing versions of the truth, and no one accountable for synchronising them.'

CASE VIGNETTE

The hallucinated promise

£95M specialist beverage distributor · 2024–2025

The customer-service team deployed a Zendesk AI agent in late 2024 to handle first-line B2B account queries. In March 2025 the agent, asked about a one-off discount, invented a 'loyalty tier' that didn't exist and confirmed 12% off future orders for one of the firm's top-10 customers. The customer screenshotted the conversation and expected it honoured. The commercial director's choice was to honour it (~£38,000 across the next 12 months) or refuse and risk the relationship. They honoured it. The post-mortem revealed the agent had been live for four months with no sampling of its outputs, no accuracy threshold defined, and no human-in-the-loop for commitments above a monetary threshold — because none of those controls had been considered at purchase.

07 The regulatory horizon

The compliance clock has started. Four dates every UK/EU mid-market CIO should have on their 2026–2027 calendar.

- **2 February 2025 (in force).** EU AI Act prohibited-use provisions applied. Social scoring, exploitative-vulnerability-targeting, and emotion-recognition-in-the-workplace use cases are already banned.
- **2 August 2025 (in force).** General-Purpose AI obligations applied to foundation-model providers. The implications for buyers: the vendor whose model you use now has transparency obligations about how it was trained. Material for your DPIA.
- **2 August 2026 (in force).** High-risk AI systems (which includes much of HR-tech, credit scoring, and some employment-adjacent uses) must meet full obligations: risk management, data governance, documentation, logging, transparency, human oversight. Buyers in retail/wholesale: review whether any of your AI-adjacent decisions touch employment or credit.
- **2 August 2027.** Full application for all categories, including systems deployed before the Act. Grandfathering exceptions are narrow.

UK adds a principles-based layer on top (DSIT's AI-Opportunities-Action-Plan + sector-specific guidance from FCA, Ofcom, MHRA). The US regulates by state; California, New York, Illinois have passed consequential AI regulation. All three regimes converge on the same technical requirements: audit trail, human oversight, documented policy, data governance, risk management.

The practical implication for a UK/EU mid-market retailer with EU customers: the compliance clock is running, you don't have the GRC team to catch up to it, and the vendor you've chosen determines what evidence you can produce.

08 Why policies alone fail

Most mid-market businesses, confronted with Shadow AI, write a policy. The policy says what's allowed, what's not, who approves exceptions. It is published to Confluence. Three months later it has drifted so far from reality that nobody reads it.

Three reasons policy-alone fails:

- **Enforcement surface mismatch.** A policy document cannot refuse an invocation. The gap between "we have a rule" and "the system blocks the action" is where every policy-era incident sits. McKinsey's 55–vs-29 gap (section 02) measures precisely this.
- **Bans drive AI deeper into shadow.** When a business bans a tool, employees find a VPN. When a business disables Copilot, employees paste the same content into ChatGPT on a phone. The 2023 Samsung ban is the textbook case; the same pattern now plays out at ~1/5th the scale inside every mid-market retailer.
- **Review cycles can't match deployment cycles.** A new vendor-side AI feature ships every week. A policy update ships every quarter. The policy is structurally behind the state of the world.

The resolution is architectural: policy at the enforcement layer, not at the Confluence layer. Section 09.

09 The control-plane framework

AgentOps is the operational discipline of running a fleet of AI agents the way a mature engineering organisation runs its services. DevOps gave us CI/CD and SRE; MLOps gave us model versioning and drift monitoring; AgentOps is the next layer — identity, cost, context, policy, and audit for autonomous AI agents acting across a business.

A control plane for AgentOps has six capabilities. In order of installation priority for a mid-market business:

1. **Inventory.** Every AI tool in use, who pays for it, who sanctioned it, what data it touches. Day one.
2. **Identity.** Single sign-on wrapper for every AI tool that supports it. Those that don't get a replacement plan.
3. **Context.** One shared source of business knowledge — identity resolution, product catalogue, policy — that every agent reads from.
4. **Spend.** One ledger. Attribution per agent, team, model, tenant. Caps enforced at invocation time.
5. **Policy.** Autonomy as scope-stacked configuration. Reviewers in a thread-first inbox. Four actions per message.
6. **Audit.** Every invocation reconstructable: inputs, model, prompt, tool calls, reviewer actions. Exportable to SIEM.

Maturity model (adapted from the DevOps and MLOps versions):

- **Ad-hoc** — individual agents, no inventory, no shared context. (Where most mid-market businesses sit today.)
- **Catalogued** — inventory and identity solved. Business knows what it has.
- **Governed** — spend, policy and audit solved. Risk committee can report green.
- **Composable** — context solved. Agents build on each other.
- **Platform** — all six capabilities on one control plane, operated as one discipline. Few mid-market businesses are here today; the ones that get there compound.

---

10 The 90-day plan

Ninety days from current state to Catalogued + first move toward Governed. The plan assumes a mid-market CIO, limited GRC headcount, no greenfield.

Days 1–30 — Shadow AI census

- Audit expense reimbursements for AI-related line items (ChatGPT, Claude, Perplexity, Jasper, Copilot upsells).
- Query SSO logs for AI-domain logins; count unique applications per department.
- Interview each department head: "what AI tools does your team actually use."
- Produce one spreadsheet — tool, owner, cost, data touched, sanctioned (yes/no). Share with CFO and risk committee.

Days 31–60 — Spend consolidation + identity wrap

- Negotiate replacement enterprise contracts for the three most-duplicated AI subscriptions. Target: 30% cost reduction.
- Wrap every AI tool that supports it with corporate SSO. For those that don't, publish a replacement plan.
- Appoint named business owners for the top three AI tools in production. Business owner (not IT) accountable for outcomes.

Days 61–90 — First governed deployment

- Pick one high-value AI use case (customer service drafting or finance ops reconciliation are the typical candidates).
- Install a context layer (even a thin one) so this agent reads from the shared business knowledge.
- Enforce autonomy policy, cost ceiling, audit trail, replay capability at the enforcement surface.
- Prepare the board paper: before/after metrics, cost, risk posture change.

Minimum viable RACI, one page: who approves AI tool purchases (CIO), who owns deployed-agent outcomes (department head), who audits quarterly (CISO), who signs the DPA (General Counsel), who is consulted on new vendor selection (data protection officer where relevant).

11 2027 predictions

Four predictions for next year's report. Three we're confident about; one we're willing to be wrong about publicly.

Multi-agent coordination becomes mainstream

By the end of 2027 the median mid-market business running AI agents will have more than one agent talking to another. Platforms that treat agent-to-agent invocation as a first-class primitive (with its own audit trail) will pull ahead; those that treat it as a hack will produce incidents.

Autonomous commerce arrives (B2B first)

Wholesale repeat-ordering is the first category where agents will make price-sensitive buying decisions within policy. Expect three to five published case studies by this time next year of mid-market B2B distributors delegating reordering to agents under autonomy caps.

Platform consolidation

The 100+ AI startups shipping in October 2026 become 10 platforms + the hyperscalers by October 2027. Mid-market buyers who've picked open-protocol platforms (MCP, OCI, OIDC) avoid the switching cost; those who picked proprietary vendors face a migration.

Prediction we're willing to be wrong about

Regulation outpaces technology. We expect the EU AI Act to produce more real enforcement action in 2027 than any US state regime; we also expect UK sector-specific guidance to start to bite in financial services and healthcare before retail. If instead the 2027 enforcement story turns out to be US-led — if California's AI consumer-protection work starts producing judgments against retailers — we'll publish a correction.

12 Methodology & sources

This is the first edition of what we intend to be an annual report. The 2026 edition is a synthesis of three concentric layers of evidence; the 2027 edition will add primary Konstera-fielded survey data from UK/EU mid-market retail, wholesale, and distribution IT and finance leaders.

Sources

- **Layer 1 — authoritative horizontal surveys.** McKinsey State of AI (2024, 2025); Deloitte State of Generative AI in the Enterprise (2024, 2025); BCG AI at Work (2024, 2025); Gartner CIO Survey + emerging-tech cycles (2024, 2025); IDC Worldwide AI Spending Guide; Microsoft/LinkedIn Work Trend Index (2024, 2025).
- **Layer 2 — security telemetry as shadow-AI proxy.** Netskope Cloud & Threat Report (2024, 2025); Zscaler ThreatLabz AI Security Report 2024; Cyberhaven data-exfiltration studies 2023–2024; Palo Alto Networks 2025 State of Generative AI; Cloudflare Radar AI-tool traffic data.
- **Layer 3 — vertical-specific signal.** NRF / IHL Group State of Retail Technology 2024; Retail Dive and Diginomica case coverage 2024–2025; Shopify merchant AI disclosures; Salesforce Connected Shoppers reports; Deloitte 2025 Retail Industry Outlook.

Where a statistic in section 02 depends on a figure the cited source expressed differently, we have retained the source's original formulation and annotated. Case vignettes in section 06 are from businesses who agreed to be included after anonymisation; the vignettes have been reviewed by the subjects for factual accuracy before publication.

What this edition does not include

- Primary survey data (planned for the 2027 edition).
- Named case studies (we're recruiting participants for 2027).
- Sector detail outside retail / wholesale / distribution. Konstera does not consult outside those sectors and has nothing to add on others.

About Konstera

Konstera is a platform for mid-market retail, wholesale and distribution businesses to deploy and govern AI agents. We built the platform because the control plane we describe in section 09 did not exist in a mid-market shape. This report exists because the problem it describes is the one our customers want solved. We have been as careful as we can about not using the report as a vendor pitch; Konstera is mentioned only here and in the publication details.

Acknowledgements

The three case vignettes come from businesses we spoke to through the Konstera Launch Program. The frameworks (Four Hidden Costs, AgentOps, Context Debt Curve) are internal names we've put in the open. If the category's shared vocabulary ends up different, we'll adopt it in the 2027 edition.

For next year's report, we are recruiting mid-market CIO / CFO / COO participants for a 15-question anonymous survey. Interested? Get in touch.

Konstera

One platform for every AI agent in your business.

Governed centrally, costed in one ledger, built on shared context — with people in control of every decision that matters.

konstera.ai · konstera.ai/book-a-call · konstera.ai/resources