Konstera

GUIDE

The AI Agent Guide *for CXOs.*

A framework for the exec who's being asked to approve AI budgets, govern tooling they didn't pick, and defend spend they can't see. What to buy, what to build, how to govern, how to budget, how to measure.

25-minute read 10 sections Konstera Edition — 2026

For CFOs, CIOs, CTOs, COOs & CMOs

GUIDE

Contents

01 Why this guide exists
02 What AI agents actually are
03 Where agents land first
04 The buy / build / extend matrix
05 Governance before deployment
06 Budgeting for agents
07 Measuring success
08 Vendor evaluation checklist
09 The 90-day pilot playbook
10 What's next (2026–2027)
FAQ Common questions

01 Why this guide exists

You're being asked, in some order, to approve AI budgets, sign vendor contracts, govern tooling you didn't pick, and explain to the board why marketing's AI invoice doubled this quarter. The vendors in the room pitch confident numbers. The analysts in the room pitch bigger confident numbers. Nobody offers you a framework for deciding what's true.

This guide is that framework. It's written for CFOs, CIOs, CTOs, COOs and CMOs at mid-market retail, wholesale and distribution businesses — businesses that are too big to wing it and too small to run a dedicated AI centre of excellence. It assumes you don't have time to become a prompt engineer, and it doesn't ask you to.

What it does ask: thirty minutes of reading, honest answers to the questions in the vendor checklist and governance sections, and a willingness to be opinionated. The opposite of an opinionated AI strategy is a sprawling one — and sprawl is the actual risk.

02 What AI agents actually are

An AI agent is **not** a chatbot. It's not a Copilot sidebar. It's not a prompt pasted into ChatGPT. Those are useful; they're not what we mean. An agent, in the sense that matters for your 2027 planning, has three properties:

AUTONOMOUS

Takes an input, reasons about it, picks actions, executes — all without a human writing each prompt. Policy decides when humans must intervene.

SCOPED

Lives inside a defined job. Not a general-purpose assistant — a specific role (customer-service drafter, invoice classifier, forecast adjuster).

TOOL-USING

Calls your systems. Reads from CRM, writes to ERP, triggers Klaviyo, posts to Zendesk. The agent is the brain; your systems are the hands.

The reason this definition matters is budget. A chatbot replaces a FAQ page — it's a tactical UI improvement. An agent replaces an operational workflow — it's a line-item in operating expense. These are different purchases. If a vendor calls their chatbot an agent, they're either confused or hoping you are.

03 Where agents land first

Agents pay back fastest in departments that already have (a) a high volume of structured decisions, and (b) a clear hand-off point when the agent is unsure. These two conditions sort the departments more reliably than any vendor's roadmap.

Our observed order of payback for mid-market retail, wholesale, and distribution businesses:

DEPT	FIRST AGENT	PAYBACK
Customer service	First-reply drafter + returns triage	4–8 weeks. High volume, structured decisions, clear reviewer sign-off.
Finance ops	PSP reconciliation + invoice classification	6–12 weeks. Repetitive, auditable, policy-constrained.
Merchandising	PIM content + localisation	8–12 weeks. Content work scales with catalogue size.
Supply chain	Demand forecasting + exception triage	12–20 weeks. Data-heavy; needs clean history.
Marketing	Campaign copy + segment-aware personalisation	6–10 weeks. Big content burden; measurable uplift.
Warehouse ops	Pick-path optimisation + exception handling	10–16 weeks. Needs WMS integration maturity.

What NOT to automate first

Three categories reliably destroy value if agents go first: high-stakes legal communications, hiring and HR decisions, and anything regulated enough that a wrong answer triggers a notification obligation. These benefit from AI eventually — just not in month one of your platform rollout.

04 The buy / build / extend matrix

For every candidate use case, you're making a buy / build / buy-and-extend decision. The right choice depends on two variables: how much the agent's workflow differs from an off-the-shelf product, and how strategic the output is to your competitive position.

HIGH FIT, LOW STRATEGIC — **Buy**

Procurement AI for commoditised functions. Agents that look identical to competitors' agents. If a vendor's hello-world does what you need, don't build it.

LOW FIT, HIGH STRATEGIC — **Build**

Proprietary workflows, differentiated data, agents that encode your business judgment. Writing your pricing logic into someone else's SaaS is giving away the moat.

HIGH FIT, HIGH STRATEGIC — **Buy and extend**

Reference agent as a foundation, customise the pieces that matter. Konstera's model: ship the hello-world, let you own the differentiation.

LOW FIT, LOW STRATEGIC — **Not yet**

Interesting but not valuable. Revisit in 12 months — either the vendor landscape fills in, or you've learnt enough to know why not.

05 Governance before deployment

Agents in production without governance are the single biggest risk category we see. Not model accuracy, not prompt injection — ungoverned deployment. The sequence matters: policy first, deployment second. In practice this means five commitments before any agent handles a real workflow.

1. An autonomy policy for every agent

For each agent, define three levels: what runs autonomously, what requires reviewer sign-off, what requires named-approver approval. The policy is not static — it tightens or loosens based on confidence, amount, counterparty, time of day. Insist this is a configuration not a code change.

2. Audit trail that spans vendors

If each AI tool keeps its own log, you have thirteen separate audit stories to tell when a regulator asks. Platform-level audit — every agent, every decision, every reviewer — removes that risk.

3. Cost ceilings per agent, per team

A monthly budget cap enforced in the platform, not promised in an email. The CFO shouldn't find out about overspend in month four — the agent should refuse the invocation at the ceiling and page the owning team.

4. Data residency aligned to the business

EU-based businesses with UK operations have specific residency needs. If your AI platform can't pin tenant data to a region, you have a GDPR problem waiting for the right incident.

5. Replay from day one

When an agent makes a wrong decision, you need to reconstruct exactly what the agent saw, which model answered, what prompt was used, and what the reviewer did. "We trusted the vendor's SLA" is not an answer post-incident.

06 Budgeting for agents

AI budgets land in three cost buckets. Count each one; don't let any vendor pretend their line item is the whole picture.

Platform cost (predictable)

The AI control plane itself — Konstera, Palantir, or a cobbled internal stack. In the mid-market, modern platforms are annual-license, flat-fee. Expect somewhere between £5,000 and £150,000 per year depending on scale. Predictable; defensible at board.

LLM cost (volatile)

Anthropic / OpenAI / Google invoices. BYOK (bring your own keys) is the right posture here — you contract with the LLM vendor directly, see their invoice, control it. Managed-LLM offerings bury the margin and surprise you. For mid-market operations expect £1,000 to £30,000 per month, scaling with invocation volume.

People cost (invisible)

Reviewer time, integration engineering, data preparation, ongoing prompt and policy tuning. Rule of thumb: in year one, people cost will match platform + LLM combined. The reviewer inbox is a real job — treat it as one.

Hidden budget lines

Three CFO-killers nobody warns you about: compute cost for embeddings and vector search (grows fast), connector licence fees (per-system; iPaaS platforms love this model), and observability (you'll want OpenTelemetry traces into your existing APM, that's another SKU).

---

07 Measuring success

Most AI metrics are vanity. "Model accuracy", "invocation count", "tokens per month" — none of those tell you whether the agent is working for the business. Four KPIs matter:

Reviewer approval rate

The percentage of agent outputs the reviewer approves without editing. Above 85% means the agent's drafts are good enough that reviewers mostly ship them. Below 60% means your reviewers are still doing most of the work and the agent isn't earning its keep.

Time-to-decision

How long between input and final human-approved output. If a customer-service agent cuts first-reply time from two hours to twenty minutes, that's a measurable business outcome. If it saves two seconds, it's a toy.

Cost per outcome

Not cost per invocation. Cost per *outcome* — resolved ticket, approved invoice, scheduled campaign. Includes LLM cost, platform cost, reviewer-time cost. Divide by the number of actual business outcomes. Benchmark against pre-agent baseline.

Deflection or capacity shift

Agents don't replace people in year one — they shift the mix of work. Measure the capacity unlocked: reviewer hours redirected to higher-judgment work, previously-impossible volume now achievable, response SLAs achievable at the current headcount.

When to kill an agent

Three signals: reviewer approval rate drops below 50% for a month and doesn't recover, cost-per-outcome inverts vs. baseline, business owner stops using the output in their workflow. Any one of the three — pull the plug and redeploy the capacity.

---

08 Vendor evaluation checklist

Twelve questions that separate a credible AI platform vendor from a well-marketed one. Ask them in writing; keep the answers.

01 Is it a platform or a framework?

A framework is a library you ship with. A platform runs your agents. If the vendor can't tell you which, they're not either.

02 Can we use any LLM, or just the ones you resell?

BYOK (bring your own key) is the right posture. Vendors who insist on their own LLM wrapper are capturing margin and locking you in.

03 What's your multi-tenancy posture?

Shared infrastructure with per-tenant row-level security is production-grade. Shared infrastructure without is a leak waiting to happen.

04 Is human-in-the-loop a product or a feature?

A reviewer UX built as a core product reflects weeks of thought. One built as an afterthought is a Zapier dashboard with buttons.

05 **What's your open-source commitment?**

Apache-2.0 SDK + source-available platform means low vendor lock-in. Pure proprietary means switching costs grow with every agent.

06 **How is cost attributed and capped?**

Per-agent, per-team, per-model cost attribution is table stakes. Enforced ceilings (platform refuses invocations above budget) is the differentiator.

07 **What audit and replay capabilities exist?**

Every invocation reconstructable — model, prompt, tool calls, reviewer actions. If replay isn't possible, incident response isn't either.

08 **What's your compliance roadmap?**

Dates, not aspirations. SOC 2 Type II target date, GDPR posture today, HIPAA roadmap if relevant. Vagueness here is a procurement red flag.

09 **Can we self-host?**

Even if you don't, the answer tells you about the vendor's product discipline. Vendors who can't self-host often can't survive your departure.

10 **Who owns the agents we build on your platform?**

You should. Full stop. Contracts that give the vendor rights to your agents, prompts, or data are not negotiable in your favour post-signature.

11 **What's your pricing model?**

Per-tenant or per-agent is defensible. Per-seat is wrong for infrastructure. Per-token is a bet on your invocation growth being flat.

12 **Who runs the reviewer inbox training?**

If the vendor's answer is 'you figure it out', the reviewer UX isn't ready for your users. Onboarding support is a product-maturity signal.

09 The 90-day pilot playbook

Twelve weeks is enough to prove or disprove an AI platform investment. Longer than that and you're in product-trial purgatory. Shorter than that and you haven't actually validated the operational workflow. Here's the structure we recommend.

Weeks 1-2 — Scoping

- Pick two agents from the first-payback table. Not one — two tests the platform's multi-agent story.
- Business owner for each agent is identified by name. Not "the AI team" — the person who owns the outcome.
- Success criteria signed by both sides, in writing, before any code.

Weeks 3-6 — Build

- Agents built against real data (ideally your production estate; otherwise a time-boxed snapshot).
- Reviewer onboarding by week five — reviewers using the inbox for real decisions.
- Weekly 30-minute checkpoints with the business owner.

Weeks 7-10 — Measure

- Daily KPI dashboard tracking the four metrics from Section 07.
- Two iteration cycles on autonomy policy and prompts.
- Reviewer feedback captured systematically — not through anecdotes.

Weeks 11-12 — Decision

- Compare measured results against the week-one success criteria. No moving the goalposts.
- Prepare the one-page exec review: outcomes, cost, risks, scaling plan.
- Go, no-go, or extend — all three are legitimate answers.

10 What's next (2026-2027)

Four shifts worth watching between now and the end of 2027, each with a "what to do about it" attached.

Multi-agent coordination becomes standard

Today's agents handle single workflows. The next generation handles chains: a demand-forecast agent that feeds a buying agent that feeds a supplier-negotiation agent. **What to do:** insist your platform supports agent-to-agent invocation as a first-class primitive, not a hack.

Autonomous commerce (B2B first)

Agents will make price-sensitive buying and selling decisions within policy. Retail wholesale will see this before DTC — B2B repeat-ordering is more amenable to policy. **What to do:** start treating "autonomous commercial decisions" as a 2027 governance question, not a 2026 one.

Platform consolidation

The vendor list will shrink. The current 100+ AI startups become 10 platforms + the hyperscalers. Mid-market buyers who've picked a BSL / Apache-licensed platform avoid the switching cost; mid-market buyers who've picked a proprietary vendor face a migration. **What to do:** favour open protocols (MCP, OCI containers) at the vendor-selection stage.

Regulation catches up

EU AI Act obligations start biting for high-risk use cases in 2026. UK follows with vertical-specific guidance. The US regulates by state. All three regimes require audit, replay, and documented policy. **What to do:** treat the compliance roadmap of your AI platform as a hard filter — not a "we'll work on it" promise.

FAQ Common questions we hear from CXOs.

Direct answers to the eight questions we hear most often from CFOs, CIOs, CTOs, COOs and CMOs evaluating AI agent platforms.

What is the difference between an AI agent and a chatbot? ×

A chatbot replaces a FAQ page — it's a tactical UI improvement. An AI agent replaces an operational workflow. Agents are autonomous (take inputs, pick actions, execute without a human writing each prompt), scoped (live inside a defined job, not general-purpose), and tool-using (call CRM, ERP, Zendesk, Klaviyo). If a vendor calls their chatbot an agent, they're either confused or hoping you are.

How much should a mid-market retailer budget for AI agents? +

What should a CFO ask before approving an AI platform contract? +

How does a CIO govern AI agents across departments? +

Where do AI agents pay back fastest for a mid-market retailer? +

Should I buy or build AI agents? +

What KPIs matter for AI agent deployments? +

When should a CXO kill an AI agent in production? +

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