AIMONGER
WHITE PAPER

AI Discovery

Multi-Agent Systems Boards Can Actually Trust

Agentic AI is entering enterprise software just as many agentic projects are forecast to be cancelled for cost, unclear value, or weak controls. The board question is not whether the global agent fashion is real. It is which workflows justify autonomy, which should remain single-loop, and which require isolation rules, cost envelopes, and human gates before a crew is funded.

David Saliba . 2026-07-16 . aimonger.com/whitepapers/multi-agent-systems-board-trust/

The board question has changed

Six months of vendor noise can make multi-agent systems sound inevitable. In the boardroom, the question is sharper: can we make agents work together safely, at predictable cost, without the infrastructure quietly catching fire?

What you see globally right now is agent fashion moving faster than operating discipline. US and Asian platforms are packaging agents into suites, startups are selling crews for every workflow, and internal teams are under pressure to show an “agentic” roadmap. Gartner predicts that 40 per cent of enterprise applications will include task-specific AI agents by the end of 2026, up from less than 5 per cent in 2025 (press release, 26 August 2025). That is a product-surface forecast: agents become a common feature in enterprise software. Separately, Gartner predicts that more than 40 per cent of agentic AI projects will be canceled by the end of 2027 because of escalating costs, unclear business value, or inadequate risk controls. Reuters reported that cancellation forecast on 25 June 2025.

Both can be true at once. Agent features proliferate. Many internal programmes still die.

McKinsey’s State of AI 2025 survey adds the value context: AI use is widespread, enterprise EBIT impact is not. Multi-agent programmes that cannot state a workflow metric will join the cancelled cohort, not because the idea is impossible, but because cost and operating risk were never designed.

This paper is a trust and capital-allocation briefing for leadership teams. It is not a framework catalogue.

> **Agentic AI, plainly:** systems that can plan multi-step work and call tools under policy without a human click at every step, not a chatbot with a persona switch. **In practice:** an agent may retrieve a policy, draft a customer reply, and queue it for human send, but must not email the client without a named approver.

> **Multi-agent system, in short:** two or more agents with defined roles, tool limits, and handoff contracts orchestrated to complete a workflow. **Worked case:** a research agent and a drafting agent may split work on a tender response, only after a single-agent loop proved insufficient on cost and quality data.

Agentwashing versus agents

> **Agentwashing means:** rebranding an assistant or chatbot as an "agent" without material tool use, isolation, or accountability, marketing language ahead of operating design. **On Monday:** refuse to fund "agentic transformation" until the vendor shows tool permissions, budget stops, and audit logs on a real workflow.

Gartner has publicly warned about "agentwashing", rebranding assistants and chatbots as agents without material agency. Boards should use a simple test:

Label	Behaviour	Board treatment
Assistant	Waits for each human step; limited tools	Useful; do not call it agentic transformation
Agent	Can plan multi-step work with tools under policy	Fund only with isolation, eval, and cost controls
Multi-agent system	Multiple agents coordinate with contracts and orchestration	Fund only when single-agent loops are proven insufficient

If the vendor cannot explain what the system may do without a human in the loop, and what it must never do, you are buying a label.

Why multi-agent systems fail in production

1. Trust that lives only in a prompt

> **Human gate, put simply:** a designed pause where a named role must approve, edit, or send before an irreversible or consequential action executes. **Board reading:** outbound regulator correspondence is always propose-and-wait; internal read-only retrieval may auto-run.

Prompts are not access control. If an agent can call tools that move money, change records, send external messages, or exfiltrate documents, "please be careful" is not a control.

Trust must live in:

- **Role tool allowlists.** Tool allowlists per role cap what each agent may invoke, so goal-completion pressure cannot widen privileges through prompt tricks alone. An allowlist is the fixed inventory of calls an agent is permitted to make, and without it the model can attempt any action it can phrase.
- **Environment separation.** Environment separation between read-only and write paths prevents a drafting agent from silently gaining record mutation because a demo needed "just one quick update." Separation is how the firm keeps a read-only agent read-only even under delivery pressure.
- **Authenticated tool calls.** Authentication and authorisation on every tool call create an audit trail finance and security can read after the fact, not a trust assumption buried in system prompt prose. An authenticated call is a recorded decision; an unauthenticated one is a hope.
- **Human-readable audit logs.** Audit logs that a human can read after the fact turn post-incident review from archaeology into a reconstructable timeline of prompts, tools, and overrides. A log no human can read is a log no human uses until after the incident.

When goal-completion pressure rises, models optimise for the goal. Infrastructure assumptions that were never made explicit become the incident report.

2. Topology cost nobody owns

> **Topology (plain English):** how agents are wired, single loop, specialist pair, orchestrator plus workers, and what that wiring does to tokens, latency, and failure blast radius. **Operating case:** an orchestrator that debates for fifteen minutes to draft one email is a superlinear cost centre; finance must see P50 and P95 cost per successful case before scale.

Orchestrator-worker designs can multiply tokens, latency, and failure points. A crew that chats with itself for fifteen minutes to draft an email is not intelligence. It is a cost centre with a narrative.

Board rule: every multi-agent proposal includes a cost model:

- **Expected cost per case.** Expected tokens and retrieval calls per successful case must be modelled at planned volume, so finance sees unit economics before coordination overhead is waved away as engineering detail. A cost model at planned volume is the only honest way to know whether the crew is affordable.
- **P95 latency budget.** P95 latency belongs in the proposal because agent chatter that clears quality on average may still miss SLA on the slow tail that operations remembers. Average latency flatters a crew that fails on the cases that take longest.
- **Human review minutes.** Human review minutes per case must be budgeted explicitly, since propose-and-wait gates that are understaffed become rubber stamps that fail exactly when risk is highest. Unbudgeted review minutes are how a gate becomes a bottleneck or a rubber stamp.
- **Monthly ceiling and alerts.** Monthly ceiling and alert threshold with hard stop must exist before scale, treating Gartner's cancellation forecast as a risk hypothesis rather than someone else's statistic. A hard stop is what converts a budget into a control.

If finance cannot see the envelope, do not scale the crew.

3. Shared state without contracts

Agents that share an unbounded scratchpad eventually collide: overwritten plans, leaked secrets, contradictory actions. Production systems need explicit contracts:

- **Read scope per agent.** Define what each agent may read so corpus slices stay minimal and cross-role leakage becomes a design violation rather than a surprise incident. Read scope is how the firm keeps an agent out of documents it has no reason to see.
- **Write scope per agent.** Define what each agent may write with default read-only posture, because write privileges should require higher role or human approval not casual demo configuration. A read-only default is the posture that prevents most accidental record damage.
- **Tool call rights.** Specify who may call which tools under which credentials, preventing "temporary" tool grants from becoming permanent paths to money movement or external send. Tool rights are the boundary between an agent that helps and an agent that acts on the firm's behalf.
- **Human escalation path.** Document how conflicts escalate to a human with named on-call coverage, so orchestrator debates that deadlock do not stall customer-facing workflows indefinitely. A named escalation path is how a deadlock becomes a decision instead of a stall.

4. No evaluation for the workflow

Unit demos ("the researcher agent found a page") do not prove end-to-end correctness. Multi-agent systems need workflow-level evaluation: did the final artefact meet the proposition tests, with provenance?

See the companion AIMonger paper on enterprise knowledge system trust for holdouts, proposition coverage, and judge calibration. Those controls apply harder when multiple agents can amplify a wrong intermediate step.

5. No kill criterion

Permanent pilots are a tax. Gartner's cancellation figure is a **forecast**, not an observed failure rate, useful as a risk hypothesis for programmes that cannot show value. Pre-write the kill criterion: metric, date, owner.

Isolation checklist (minimum before production)

> **Isolation, briefly:** hard boundaries on what each agent may read, write, spend, and call, so one compromised or runaway step cannot widen privilege or exhaust budget. **Operating case:** Agent A's corpus slice excludes customer PII; write tools require a separate role; per-run spend hard-stops at EUR 50.

Use this as a go/no-go gate.

1. **Role isolation.** Each agent has a written role and tool allowlist reviewed before production, so "helpful" prompt edits cannot silently expand what the system may do. A written role is the artefact the reviewer compares against when something looks off.
2. **Data isolation.** Agents see only the corpus slices required for the role, preventing cross-domain leakage when orchestrators pass context between specialists. Minimal slices mean a leak only exposes what the agent needed, not the whole corpus.
3. **Write isolation.** Default read-only posture applies until write privileges are explicitly granted with human approval or higher role, reducing irreversible record mutation from drafting agents. Read-only-by-default is the single most effective control against accidental writes.
4. **Secret isolation.** Credentials never appear in shared scratchpads or user-visible traces, because multi-agent chatter is a common path for accidental secret exfiltration. A secret in a scratchpad is a secret one prompt away from leaving the firm.
5. **Network isolation.** Egress allowlists govern tools that call the open internet, so one agent cannot become an uncontrolled export path for corpus content. An egress allowlist is how the firm keeps a helpful agent from becoming an export channel.
6. **Tenant isolation.** Shared indexes do not cross trust boundaries without a design review, since retrieval mistakes across tenants are trust failures no accuracy score fixes. A cross-tenant retrieval miss is a breach dressed as helpfulness.
7. **Budget isolation.** Per-run and per-day spend caps hard-stop runaway loops, treating nested agent debate as a FinOps risk not an engineering curiosity. A hard stop is what prevents a stuck loop from consuming a month of FinOps spend overnight.
8. **Audit isolation.** Immutable logs capture tool calls, prompt versions, and human overrides, giving risk committees reconstructable evidence after incidents. Immutable logs are the difference between an incident you can investigate and one you can only guess at.
9. **Failure isolation.** One agent crash does not grant another agent broader tools, preventing cascade privilege expansion under partial failure conditions. A crash that widens privileges turns a partial failure into a full incident.

10. **Human gate on consequence.** Consequential actions pause for a named role before execution, so autonomy stops where accountability and reversibility require a person. The gate is where the firm decides which proposed actions become real ones.

If fewer than eight are green, the system is a laboratory, not an operating dependency.

Topology decision tree

Start with one agent loop when:

- **Linear workflow shape.** The workflow is mostly linear with few branches, so coordination overhead would dominate any quality gain from adding specialist agents. A linear workflow does not need a crew; it needs a well-placed single loop.
- **Small tool surface.** Tool count is small enough to allowlist completely, making isolation and audit tractable for a mid-market team without a dedicated agent platform. A small surface is one a mid-market team can actually control.
- **Tight latency budget.** Latency budget is tight for customer or operator-facing steps, ruling out orchestrator chatter that trades minutes of token spend for marginal drafting polish. A tight budget is the constraint that makes single loops the responsible choice.
- **Early evaluation discipline.** The organisation is still learning evaluation discipline, and a single loop teaches holdouts and provenance before topology complexity obscures failure modes. Learning evaluation on a single loop is hard enough without orchestrator noise on top.

Add a second specialist agent when:

- **Separable skill boundary.** A clear separable skill exists such as retrieval specialist versus drafting specialist, with interfaces that can be written on one page and tested independently. A separable boundary is the precondition for a split that actually helps.
- **One-page handoff contract.** The handoff contract fits on one page with explicit inputs, outputs, and refusal rules, so debugging does not require replaying an opaque orchestrator debate. A contract that fits on one page is a contract the team can actually maintain.
- **Measured split benefit.** You can measure whether the split improves quality or cost against a frozen single-loop baseline, not merely whether the demo looks more “agentic.” A measured benefit is the only honest justification for the added complexity.

Add an orchestrator crew only when:

- **Real parallel specialisation.** Parallel specialised work is real rather than theatrical, meaning tasks can run concurrently without constant reconciliation that burns tokens. Real parallelism pays for itself; theatrical parallelism burns tokens to look busy.
- **Designed conflict resolution.** Conflict resolution and escalation are designed with named human paths, so orchestrator disagreements do not stall production waiting for an ad hoc meeting. Designed resolution is how a crew avoids stalling on its own debate.
- **Business case after overhead.** Cost model still clears the business case after coordination overhead is included at P50 and P95, not optimistically at happy-path volume only. A case that only clears at happy path is a case that fails in production.

Refuse multi-agent when:

- **No workflow metric.** The proposal says “agents” without a workflow metric and baseline, which is fashion funding unless absorption and trust gates are already green elsewhere. A proposal without a metric is a proposal without a definition of success.
- **Vendor isolation opacity.** Vendors cannot explain isolation, tool permissions, or budget stops in detail, signalling the control plane is slideware not production infrastructure. Opacity here means the controls do not exist yet.

- **Single-loop sufficiency.** The same outcome is achievable with deterministic automation plus one assisted drafting step, making crews an expensive substitute for absorption discipline. Sufficiency is the test that keeps the firm from buying complexity it does not need.

Gartner's application forecast, agents becoming common in enterprise apps by end-2026, will pressure teams to adopt crew language. The decision tree above is how leadership resists fashion without refusing useful technology.

that restraint is an advantage. A well-measured single loop in procurement, claims, discovery, or customer operations can beat a fashionable crew that consumes budget and teaches the organisation nothing.

Human gates are not a failure of automation

In regulated EU contexts, human oversight is often an operating and reputational requirement, not only a taste preference. Sector practice, customer expectations, works council realities, and EU rules can all matter depending on the workflow. Multi-agent systems that can send messages, alter records, or influence customer outcomes need a named human gate.

Design gates as product behaviour:

- **Auto-run actions.** Define which actions auto-run under logged, bounded conditions so low-risk read and draft steps keep tempo without hiding irreversible moves inside "helpful" defaults. Auto-run is the category that keeps the system fast on the work that does not need a human.
- **Propose-and-wait actions.** Define which actions propose-and-wait for a named approver before execution, especially for external send, financial transfer, and record mutation paths. Propose-and-wait is the category that keeps the firm in control of the actions that cannot be undone.
- **Override on-call.** Name who is on-call for overrides when gates backlog or edge cases spike, because a gate with no staffed owner becomes decoration the first busy week. An unstaffed gate is a gate that will be bypassed exactly when it matters.
- **Override-to-fixture path.** Document how overrides become evaluation fixtures within weeks, so human corrections compound into institutional learning instead of one-off heroics. An override that becomes a fixture is learning; one that stays a heroics is a recurring cost.

A gate with no owner is decoration.

Cost and cancellation: run the programme like finance

Translate Gartner's cancellation drivers into monthly steering questions:

Cancellation driver	Steering question
Escalating costs	Are we inside the token/tool budget? What is cost per successful case trend?
Unclear business value	Which workflow metric moved: cycle time, error rate, capacity hours?
Inadequate risk controls	How many isolation checklist items are red? Any critical incidents?

McKinsey's finding that only about 6 per cent of organisations are AI high performers with material EBIT impact is a reminder: sophistication of architecture does not equal value. A boring single-agent loop that removes twenty hours a week beats an elegant multi-agent graph that produces slides.

Ninety-day trust programme

Days 1-30: Single loop, real workflow

- **Measurable pain workflow.** Pick one workflow with measurable pain such as document assembly, exception triage, or discovery prep, so ninety-day reporting can show movement finance recognises. A measurable pain workflow is what makes the ninety-day report meaningful instead of anecdotal.

- **Gated single loop.** Ship a single-agent or assisted loop with provenance and a human gate before any crew language appears in steering slides or vendor contracts. A gated single loop is the honest first step that crews must earn their way past.
- **Proposition evaluation set.** Build an evaluation set with proposition coverage and holdout integrity, treating unit demos as insufficient proof for multi-step autonomy. Proposition coverage is what keeps the eval set honest about which claims it actually tests.
- **Cost per successful case.** Establish cost per successful case at expected volume including human review minutes, giving finance a baseline before topology experiments multiply tokens. A baseline cost is the anchor that lets the board tell improvement from drift.

Days 31-60: Harden isolation

- **Isolation checklist to green.** Run the isolation checklist to green with documented evidence, because fewer than eight greens means the system remains laboratory equipment not an operating dependency. Documented greens are what the risk committee can review, not just claims.
- **Shadow live evaluation.** Add shadow evaluation on live traffic samples where available, catching permission and temporal failures that rehearsed fixtures miss. Live samples expose the failure modes the team did not think to rehearse.
- **Operator failure-mode training.** Train operators on failure modes and override paths, so the first production surprise does not invent training curriculum under incident pressure. Training before the incident is what keeps the incident small.
- **Second-agent evidence decision.** Decide whether a second agent is justified by cost and quality data against the frozen single loop, not by vendor roadmap fashion or internal ambition. A data-justified second agent is the only kind worth the added complexity.

Days 61-90: Expand or stop

- **Adjacent expansion on control plane.** If metrics move and isolation holds, expand to an adjacent workflow reusing the same control plane rather than funding a new crew from scratch. Reusing the control plane is how the firm compounds its investment instead of restarting it.
- **Cancel narrow restart.** If metrics do not move, cancel or restart with narrower scope instead of adding agents to compensate for a weak workflow choice or missing absorption. Cancelling narrow is discipline; adding agents to a weak workflow is doubling down on a bad bet.
- **Board report bundle.** Report to the board with cost, quality, risk, and training status in one bundle, connecting agent spend to workflow outcomes not architecture diagrams alone. A bundle that connects spend to outcomes is what lets the board judge the programme.

Training is part of the control plane

Multi-agent systems fail socially when operators do not understand what the system is allowed to do. Workshops and team training should cover:

- **Per-agent touch boundaries.** What each agent may touch in systems of record and corpora, so operators know when a run has left its intended slice before damage occurs. Touch boundaries are what let an operator spot a runaway run early.
- **Provenance reading.** How to read provenance on drafts and tool traces, enabling supervisors to verify sources before approving consequential sends or record changes. Provenance reading is the skill that turns a gate into a real review rather than a rubber stamp.
- **Override discipline.** When to override versus when to stop the run entirely, preventing heroic prompt edits that widen privilege instead of filing a controlled failure into the eval set. Override discipline keeps a correction from becoming a privilege expansion.

- **Failure filing path.** How to file a failure into the evaluation set with reproduction steps, so corrections become fixtures within weeks rather than hallway anecdotes. A filed failure is an asset; an anecdote is a recurring cost.

Tools without trained operators become either shelfware or incident generators. For invitation-quality delivery, training is not optional polish.

Across global demos, the agent acts as if the organisation is frictionless. Inside real firms, operators need to know when to stop the system, when to trust it, and who carries the decision.

Incident classes boards should pre-write

Do not wait for the first production incident to invent language. Pre-classify:

Class	Example	Immediate action
Cost runaway	Nested agent chatter exhausts budget	Hard stop; postmortem on topology
Privilege creep	Agent gains write tool "temporarily"	Revoke; audit all tool grants
Cross-domain leakage	Agent A sees Agent B's corpus slice	Isolate indexes; notify security
Silent wrongness	Confident final artefact, failed proposition	Disable auto-run; human-only mode
External send error	Message sent without gate	Kill egress; customer notification plan

Each class needs an owner and a communication path. Multi-agent systems without incident classes are unfinished.

Single-loop excellence before crew complexity

A useful maturity ladder:

1. **Assisted drafting.** Human always sends on external and consequential paths, using models for speed while accountability stays visible to customers and regulators. This stage keeps the firm safe while it learns the rest of the controls.
2. **Single agent with tools.** Single agent with allowlisted, logged, gated tools runs production workloads once isolation checklist items are green and cost per case is baselined. This stage is where most workflows should stop and stay profitable.
3. **Specialist pair.** Two roles operate under a one-page contract with explicit handoffs, justified only when measured split beats the frozen single loop on cost or quality. A pair earns its place with data, not aspiration.
4. **Orchestrated crew.** Parallel work runs with conflict rules and human escalation paths, funded only when coordination overhead still clears the business case at P95 volume. A crew is a commitment the firm should make only with cost evidence in hand.
5. **Cross-system agents.** Cross-system agents come only after stages one through four are boring in operations, not when a vendor slide declares "agentic transformation" ready. Skipping to cross-system is how programmes hit the cancellation drivers Gartner names.

Skip-a-level programmes are where Gartner's cancellation drivers concentrate: cost surprises, unclear value, weak controls.

Vendor due diligence for agent platforms

1. Show the tool permission model in detail.
2. Demonstrate a hard budget stop, not a warning email.
3. Provide an exportable audit trail.

4. Explain how you prevent agentwashing in your own packaging.
5. Share reference architectures for human gates.
6. Disclose subprocessors and data residency.
7. Provide a cost calculator for multi-step runs.
8. Allow independent evaluation harnesses.
9. Document failure modes from real deployments (anonymised).
10. State support expectations when an agent takes a harmful action.

If the sales process cannot leave slideware, the control plane is not ready.

Organising the team

Recommended minimum roles for a production agentic workflow:

- **Business owner.** Business owner holds the workflow metric and kill criterion, signing residual risk acceptance when gates and refusal rates are tuned for consequence. The business owner is who decides the programme is worth the risk it carries.
- **Technical owner.** Technical owner owns isolation, evaluation, and releases with rollback authority, certifying that each promotion matches the versioned evidence packet. The technical owner is who the board calls when a release regresses.
- **Risk partner.** Risk partner runs security and legal checkpoints on tool permissions, egress, and data handling before autonomy expands beyond read-only paths. The risk partner is who stops the programme from expanding autonomy before the controls catch up.
- **Operator leads.** Operator leads own training quality and override feedback, ensuring gates are staffed and failures become fixtures instead of silent workarounds. Operator leads are who keep the workflow running when the project team leaves.

Avoid a pure "AI centre of excellence" that owns everything and therefore owns nothing in the line organisation. Centres can set standards; lines must own outcomes.

Operating constraints under EU rules

- **Customer and reputation risk.** Dense professional markets remember visible failures, so start with bounded measurable workflows that can become referenceable wins rather than broad autonomous crews. A bounded win is referenceable; a broad failure is a trust tax.
- **Works council consultation.** In some jurisdictions monitoring and automated decision practices require staff consultation before scale, making surprise autonomy deployments a legal and cultural risk. Consultation before deployment is how the firm avoids a works council complaint that halts the programme.
- **Sector rule constraints.** Finance, health, gaming, and professional services may constrain autonomous actions harder than horizontal IT policy, requiring sector-specific gate design not generic "AI policy" slides. Sector rules are where generic policy fails and sector-specific design is required.
- **Cross-border data residency.** Multi-agent tool calls can accidentally move data across borders, so map egress and subprocessors before production traffic not after the first residency question from a customer. Residency mapping before traffic is how the firm avoids a cross-border incident.

- **EU AI Act obligations.** Documentation, transparency, and oversight expectations may apply depending on role and risk classification, pairing vendor claims with your own inventory and human oversight records. The Act is use-specific, so the firm's obligations depend on what the system actually does.

These are not reasons to freeze. They are reasons to prefer gated single loops with clean audit trails over theatrical crews.

Linking the four AI Discovery papers

Read as a set:

1. **Absorption capacity.** Decide which workflows deserve intelligence and owners before funding crews, because agents without absorption produce cancellation statistics with better branding. Absorption is the precondition that makes agents worth funding.
2. **Knowledge trust.** Measure discovery systems with holdouts and provenance before they feed agents, since wrong retrieval amplified across roles becomes expensive confident wrongness. Trust is what stops a crew from amplifying a single wrong step.
3. **Durable advantage.** Invest in feedback, retention, and judgement above the API, so temporary model leads compound into assets rather than evaporating at the next price cut. Durable advantage is what keeps the firm's position when the model changes.
4. **Multi-agent trust.** Fund crews only when single loops are proven insufficient and isolation, cost envelopes, and kill criteria are already green on the control plane. Multi-agent trust is the last gate, not the first announcement.

Capital should flow through that sequence. Crews that skip absorption and trust become cancellation statistics with better branding.

Appendix: multi-agent business case (required fields)

No crew funding without:

1. **Workflow metric and baseline.** Workflow metric and baseline dated and signed, so the board can judge whether the crew moved cycle time, error rate, or capacity hours rather than token charts alone. A signed baseline is the anchor that makes later movement meaningful.
2. **Single-loop insufficiency evidence.** Evidence why a single-agent loop is insufficient on cost or quality data, not taste or vendor keynote language about "agentic transformation." Evidence is what separates a justified crew from a fashionable one.
3. **Role and tool matrix.** Role and tool matrix reviewed by risk, showing allowlists, write paths, and credential scope before autonomy expands beyond read-only operations. A matrix reviewed by risk is how the firm keeps tool grants from creeping.
4. **Isolation checklist self-score.** Isolation checklist self-score with fewer than two reds or documented compensating controls, treating eight-of-ten greens as laboratory status not production dependency. A self-score is the honest internal read on whether the system is ready.
5. **Volume cost model.** Cost model at expected volume with P50 and P95 per successful case, including human review minutes and retrieval calls finance can reconcile monthly. A volume model is what lets finance sign off without guessing.
6. **Human gate staffing plan.** Human gate staffing plan with named on-call coverage for propose-and-wait paths, because unstaffed gates become rubber stamps under load. A staffed gate is a real control; an unstaffed one is a line item.

7. **Proposition evaluation plan.** Evaluation plan with proposition coverage and holdout integrity, proving end-to-end workflow correctness not unit demos of individual agents. End-to-end evaluation is what proves the crew works as a crew.
8. **Kill date and metric.** Kill date and kill metric pre-written with owner signature, converting Gartner's cancellation forecast into an explicit programme contract rather than a vague worry. A kill date is how the firm keeps the programme from becoming immortal.
9. **Operator training plan.** Operator training plan covering touch boundaries, provenance, overrides, and failure filing before go-live traffic is accepted. Trained operators are what keep the crew running after the project team leaves.
10. **Incident class table.** Incident class table with owners and communication paths, so the first cost runaway or external send error does not invent response roles under pressure. A pre-written table is how the firm responds in hours instead of days.

Business cases missing item 2 are fashion. Business cases missing item 8 are immortality projects.

Appendix: red-team script (quarterly)

- **Prompt privilege escalation.** Attempt privilege escalation via prompt on each agent quarterly, verifying tool grants cannot widen through conversational tricks or shared scratchpad contamination. This test is how the firm proves the allowlist holds under pressure.
- **Cross-role data access.** Attempt cross-role data access between corpus slices, ensuring retrieval filters and tenant boundaries hold when orchestrators pass context between specialists. The test catches the leakage path that accuracy scores never show.
- **Recursive budget exhaustion.** Attempt budget exhaustion via recursive delegation, confirming per-run and per-day hard stops fire before a nested crew consumes the monthly FinOps envelope. This test is how the firm proves the hard stop actually stops.
- **Ungated external send.** Attempt external send without gate on every path that could reach customers or regulators, treating a single ungated egress as a promotion blocker. One ungated path is one incident waiting to happen.
- **Superseded policy injection.** Inject superseded policy documents and test temporal behaviour, because agents that always answer with "latest" fail historical and compliance questions silently. Temporal injection is how the firm catches the failure that hides behind a current, authentic citation.
- **Mid-run agent disable.** Disable one agent mid-run and observe failure mode, verifying failure isolation does not grant surviving agents broader tools or orphaned write privileges. A mid-run disable is how the firm proves a crash does not widen privileges.
- **Audit log completeness.** Verify audit log completeness after each attempt, giving risk committees reconstructable evidence that matches what operators saw in the UI traces. Log completeness is what turns a red-team into evidence the board can rely on.

Record results for the risk committee. Multi-agent trust without adversarial testing is optimism.

Cost topology examples (illustrative)

Design	Typical cost behaviour	Prefer when
Single agent, few tools	Linear with steps	Most mid-market workflows
Draft agent + critic	~2x generation cost, often worth it for quality	High-stakes drafting
Orchestrator + N workers	Superlinear; coordination tax	True parallel specialised work
Unbounded debate crew	Pathological	Never in production

Require engineering to place the proposed design on this table before procurement.

From board curiosity to controlled capacity

Boards are right to be curious about agents. Curiosity becomes negligence when curiosity is funded without isolation, cost envelopes, and kill criteria. The winning posture in 2026-2027, given Gartner's adoption and cancellation forecasts, is selective aggression: move fast on single loops that clear trust gates; move slowly on crews until single loops are boring.

That posture matches AIMonger's delivery style: measurable change, human capacity around the system, and no theatre.

Evidence base: deployment intent is ahead of operating maturity

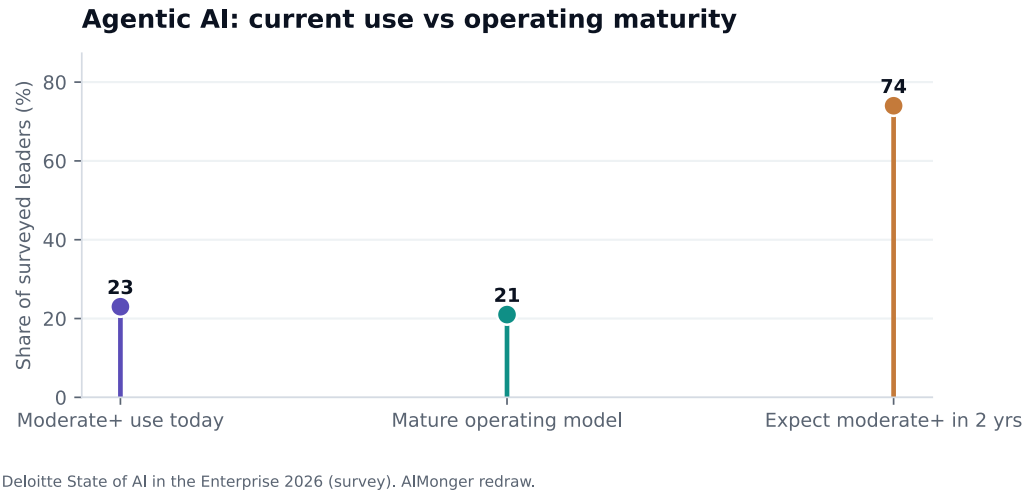

Figure 1. Current moderate+ agentic use vs mature operating models vs two-year expectations. Plain read: intent outruns discipline; most firms are not ready to scale crews. **On Monday:** a Portuguese bank funds single-loop exception triage before any orchestrator roadmap. Source: Deloitte State of AI in the Enterprise 2026 (survey, 3,235 leaders). AIMonger redraw.

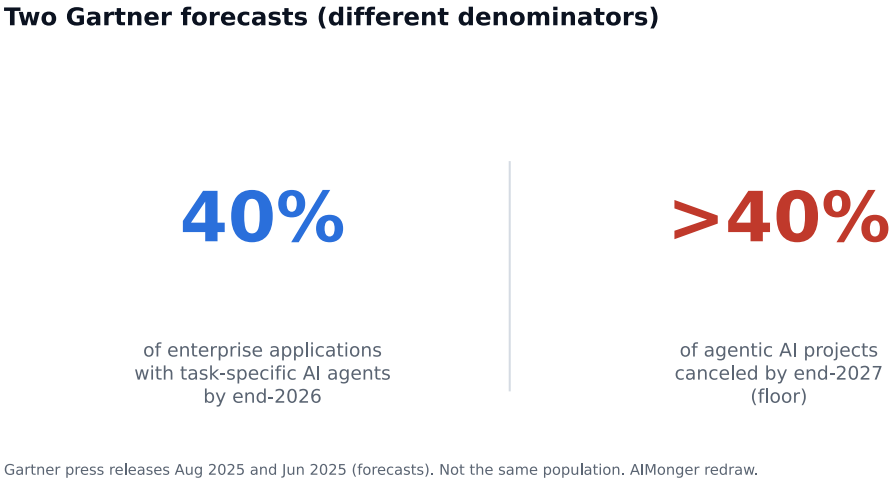

Figure 2. Two separate Gartner forecasts - not the same denominator. Left: share of enterprise applications with task-specific agents by end-2026. Right: floor share of agentic projects forecast canceled by end-2027. Both can be true: agents become product features while many internal programmes still die on cost and controls. **On the P&L:** treat the cancellation forecast as a risk hypothesis; require cost envelope and kill criteria upfront. Source: Gartner press releases, 26 Aug 2025 (apps with task agents) and 25 Jun 2025 (project cancellation forecast). See References. AIMonger redraw.

Deloitte's 2026 State of AI in the Enterprise surveyed 3,235 business and IT leaders in 24 countries during August-September 2025. It found 23% reporting at least moderate agentic AI use, while 74% expected at least moderate use within two years. Only 21% reported a mature operating model for autonomous agents. Because respondents were directly involved in AI initiatives, these rates describe an engaged enterprise sample, not all organisations.

Gartner's public June 2025 forecast that more than 40% of agentic AI projects would be cancelled by end-2027 is a forecast, not an observed cancellation rate. Its stated drivers, escalating cost, unclear value, inadequate risk controls, are useful risk hypotheses for investment committees.

OWASP's Top 10 for Agentic Applications 2026 was developed with more than 100 industry experts, researchers, and practitioners. It identifies goal hijacking, tool misuse, identity/privilege abuse, supply-chain compromise, unexpected code execution, memory poisoning, insecure inter-agent communication, cascading failures, human-agent trust exploitation, and rogue agents. This is security-practice consensus, not incidence-frequency data.

Research-to-control mapping

Research signal	Required control
21% mature autonomous-agent operating model (Deloitte respondent sample)	Operating maturity gate before scaling autonomy
Agent project cancellation forecast (Gartner)	Cost envelope, value metric, kill criterion
Tool misuse / identity abuse (OWASP)	Least agency, task-bound credentials, parameter validation
Cascading failures / insecure inter-agent communication (OWASP)	Blast-radius isolation, authenticated messages, circuit breakers
Human-agent trust exploitation (OWASP)	Confirmation for sensitive actions and calibrated operator training

Counter-position: controls can remove the speed advantage

A control that requires manual approval for every low-impact read action can make an agent economically pointless. Apply controls to action reversibility and consequence. Read-only, bounded, auditable work can move faster. Irreversible external actions require stronger gates. "Human in the loop" is not one design; it is a portfolio of approval, review, sampling, and exception patterns.

Board risk appetite statement (template)

The organisation permits agent autonomy only where actions are reversible or bounded, credentials are task-scoped, full action traces are retained, cost ceilings are enforced, and a named owner can stop the system. External communication, financial transfer, destructive writes, regulated decisions, and privilege changes require explicit approval unless the board has approved a documented exception.

Research addendum: multi-agent gains are conditional

Public research does **not** support "more agents are better."

Evidence	Result	Board reading
Google multi-agent study (financial analysis vs sequential planning)	Large gains on parallelizable analysis; **declines** of roughly 39-70% across multi-agent topologies on sequential planning in the reported setting	Fit topology to task structure
Anthropic multi-agent research evaluation	About **90%** improvement on an internal breadth-first research evaluation; about **15x** tokens versus chat interactions	Parallel research can pay; token tax is real; dependency-heavy work often poor fit
Equal-budget 2026 preprint	Single agents matched or beat multi-agent systems on multi-hop reasoning when reasoning-token budgets were equal	Normalise compute before declaring crew superiority
Error amplification measurements	Independent multi-agent setups amplified errors more than centralized orchestration in the reported benchmarks	Orchestration and isolation are risk controls
τ-bench tool-agent results	Pass rates far below "enterprise ready" aspirations in retail/airline settings for function-calling agents in the published 2024 GPT-4o-era results (historical benchmark era; re-run on current models before funding)	Tool reliability needs evaluation, not demos

Never present Gartner's "130 real agentic vendors" estimate as a verified census; method and list are not publicly reproducible.

Closing position

Multi-agent systems can be real operating capacity. They can also be an expensive way to fail.

Gartner says agents are entering the enterprise application fabric, and that a large share of agentic projects will be cancelled by end-2027 without cost discipline, clear value, and risk controls. Boards that fund isolation, topology restraint, human gates, evaluation, and operator training will still have useful operating capacity when the fashion wave moves on.

Trust is infrastructure. If it lives only in a prompt, you do not have it.

References

1. Gartner, "Gartner Predicts 40% of Enterprise Apps Will Feature Task-Specific AI Agents by 2026, Up from Less Than 5% in 2025," press release, 26 August 2025. https://www.gartner.com/en/newsroom/press-releases/2025-08-26-gartner-predicts-40-percent-of-enterprise-apps-will-feature-task-specific-ai-agents-by-2026-up-from-less-than-5-percent-in-2025
2. Gartner, "Gartner Predicts Over 40% of Agentic AI Projects Will Be Canceled by End of 2027," press release, 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
3. Reuters, "Over 40% of agentic AI projects will be scrapped by 2027, Gartner says," 25 June 2025. https://www.reuters.com/business/over-40-agentic-ai-projects-will-be-scrapped-by-2027-gartner-says-2025-06-25/
4. Deloitte AI Institute, "The State of AI in the Enterprise: The Untapped Edge" (2026 edition); survey of 3,235 leaders, Aug-Sep 2025. https://www.deloitte.com/us/en/what-we-do/capabilities/applied-artificial-intelligence/content/state-of-ai-in-the-enterprise.html
5. OWASP GenAI Security Project, "OWASP Top 10 for Agentic Applications for 2026," 9 December 2025. https://genai.owasp.org/resource/owasp-top-10-for-agentic-applications-for-2026/
6. McKinsey & Company / QuantumBlack, "The State of AI: Global Survey 2025." https://www.mckinsey.com/capabilities/quantumblack/our-insights/the-state-of-ai
7. Stanford Institute for Human-Centered Artificial Intelligence, "The 2026 AI Index Report," Economy chapter. https://hai.stanford.edu/ai-index/2026-ai-index-report/economy
8. Regulation (EU) 2024/1689 of the European Parliament and of the Council (EU AI Act). https://eur-lex.europa.eu/eli/reg/2024/1689/oj
9. Eurostat, "20% of EU enterprises use AI technologies," 11 December 2025 (20.0% in 2025; 13.5% in 2024). https://ec.europa.eu/eurostat/en/web/products-eurostat-news/w/ddn-20251211-2
10. Anthropic, "How we built our multi-agent research system" (engineering notes on token cost and orchestration). https://www.anthropic.com/engineering/built-multi-agent-research-system
11. Yao et al., "τ-bench: A Benchmark for Tool-Agent-User Interaction in Real-World Domains," arXiv:2406.12045 (2024). https://arxiv.org/abs/2406.12045

Published by AIMonger . https://aimonger.com . Malta . Europe