AI Agents for the Mid-Market Back Office: Controls before Automation

A Board Framework for Authority, Evidence, Human Approval and Safe Execution

Authored by Chennakeshav Adya, Independent Researcher

August 2026

Abstract

Artificial-intelligence agents can assemble information, select tools, make intermediate decisions and initiate actions across finance, procurement, customer operations, human resources and compliance. This expands the economic opportunity beyond task assistance. It also changes the control problem. A conventional workflow follows a defined route. An agent can choose a route, call an external service, retain contextual memory and continue until it believes an objective has been met. The company therefore needs to govern authority, evidence and safe failure as carefully as it governs model quality.

This paper develops a board framework for controlled agent adoption in the mid-market back office. It distinguishes deterministic automation, AI assistance and bounded agency; inventories processes and decisions; defines an authority envelope; separates identities and credentials; controls data, tools and memory; places deterministic checks around probabilistic reasoning; and designs human approval around consequence and reversibility. It also establishes testing, release, monitoring, incident and third-party controls.

The evidence base spans multiple jurisdictions and operating contexts. The United States National Institute of Standards and Technology has launched an AI Agent Standards Initiative focused on security, identity and interoperability, while its AI Risk Management Framework and generative-AI profile organise risk work through govern, map, measure and manage. The Dubai International Financial Centre's Regulation 10 addresses personal-data processing through autonomous and semi-autonomous systems. The United Kingdom National Cyber Security Centre places secure design, development, deployment and operation across the system lifecycle. European Union transparency obligations began applying on 2 August 2026, alongside a revised implementation timeline for high-risk rules. OECD principles, Singapore's Model AI Governance Framework and ISO/IEC 42001 reinforce accountability, human oversight, traceability and continuous improvement.

Six original figures and six implementation tables support a 120-day adoption sequence. Any thresholds, scoring bands, performance ranges or operating examples are illustrative management assumptions. They are neither legal conclusions nor predictions for a specific company. Management should validate each use case against current process evidence and obtain qualified legal, data-protection, employment, cyber-security, audit, accounting and regulatory advice before production deployment.

JEL Classification: D23, G30, L21, M15, O32

Keywords: AI agents, back-office automation, artificial intelligence governance, internal control, human oversight, audit trail, identity and access management, mid-market, operational resilience, enterprise value

1. Define the board decision as delegated authority

An AI agent should be treated as a delegated operating capability. The board decision concerns which objectives the system may pursue, which information it may use, which tools it may call, which commitments it may create and which actions remain subject to approval. A licence to use a model or a successful demonstration does not answer those questions.

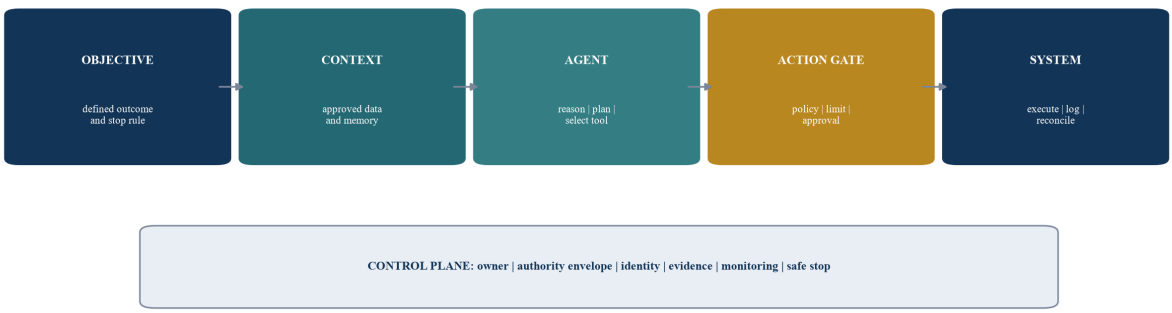
The economic case can be attractive. Back offices contain repeated searches, reconciliations, document handling, exception triage, follow-ups and status updates. An agent can connect these steps across systems and time. It can prepare a supplier-onboarding pack, trace a disputed invoice, assemble a cash forecast or coordinate a customer-service case. Value arises when the company removes delay, rework and fragmented ownership while preserving reliable evidence.

The control exposure also expands. The agent may select the wrong record, follow an instruction embedded in an untrusted document, use a credential beyond the intended purpose, disclose personal data, create a duplicate payment, apply an outdated policy or continue after the business context has changed. The consequence can travel through several systems before a human sees it.

The board mandate should therefore define four boundaries. The outcome boundary states what the agent may achieve. The resource boundary states which data, applications, identities, budget and time it may use. The action boundary states which acts it may execute and which require approval. The failure boundary states how the system stops, contains harm, preserves evidence and returns work to a person.

The first policy decision should identify prohibited actions. Examples may include releasing cash, changing bank details, terminating employment, accepting contractual liability, filing regulatory returns, overriding segregation of duties or communicating a final adverse decision to an individual. The exact list depends on the company's legal obligations, operating model and risk appetite.

Figure 1. The controlled agent operating architecture



The objective, identity, tools and output remain inside explicit policy, approval and evidence boundaries.

2. Ground the programme in current cross-market evidence

Official frameworks converge on lifecycle governance, accountability, testing and the capacity to intervene. NIST's AI Risk Management Framework organises work through govern, map, measure and manage. Its generative-AI profile identifies risks that can be novel or amplified by generative systems and provides actions across the lifecycle. In February 2026, NIST launched the AI Agent Standards Initiative with specific pillars for agent security, identity, authorisation, interoperability and evaluation. The initiative confirms that enterprise agent controls remain an active standards-development area.

The DIFC amended its Data Protection Regulations in 2023 to include Regulation 10 on processing personal data through autonomous and semi-autonomous systems. The regulation and related guidance place personal-data use, accountability and affected individuals within an outcomes-based control environment. A mid-market company operating in or through the DIFC should map the regulation to each agent that processes personal data and seek current advice on scope and obligations.

The UK NCSC's international secure-AI guidance covers secure design, development, deployment, operation and maintenance. It calls for threat modelling, supply-chain due diligence, protected infrastructure, logging, monitoring, incident management and responsible release. These controls matter where a back-office agent relies on external models, connectors, libraries or application-programming interfaces.

The European Commission states that transparency obligations under Article 50 of the AI Act started to apply on 2 August 2026. The Commission's 2026 implementation materials also reflect revised dates for high-risk rules. A company should confirm whether it acts as provider, deployer, importer or another operator and assess each use case against the current official text and applicable sector law.

Financial-services evidence offers a useful control signal. The joint Bank of England and Financial Conduct Authority 2024 survey reported that 75 percent of responding firms were using AI, 33 percent of use cases involved third parties, 55 percent involved some automated decision-making and only 2 percent were fully autonomous. It also reported that cyber security was the greatest perceived risk and that complete understanding of deployed AI was limited. Survey evidence does not determine the right design for an individual company; it does show the importance of accountable ownership, third-party visibility and calibrated autonomy.

Table 1. Official evidence and control implications

Authority	Current signal	Control implication	Management boundary
NIST AI RMF and agent initiative	lifecycle risk management; identity, authorisation and security work continues	maintain an agent register, authority envelope and evaluation record	voluntary guidance must be mapped to the use case
DIFC Regulation 10	autonomous and semi-autonomous personal-data processing is expressly addressed	complete data, individual-impact and accountability analysis	territorial and processing scope require legal review
UK NCSC	secure design through secure operation; supply-chain and incident controls	threat-model tools, connectors, credentials and failure paths	guidance complements applicable law and cyber duties
European Commission	Article 50 transparency duties apply from 2 August 2026; timelines vary by rule	classify role, system and interaction; retain current compliance map	the official text and amendments control
FCA and Bank of England	third-party use, partial understanding and cyber risk are material themes	evidence vendor dependencies, accountable owners and monitoring	survey findings describe respondents, not every firm
OECD, Singapore and ISO	human oversight, traceability, robustness and continuous improvement recur	integrate AI into enterprise governance and internal control	voluntary frameworks require proportionate adoption

Applicability, definitions and obligations require current company-specific confirmation.

3. Separate automation, assistance and agency

The word automation hides three different operating models. Deterministic automation follows encoded rules: match a purchase order, validate a field, route a document or calculate a threshold. AI assistance produces a draft, classification, summary or recommendation for a person. Bounded agency selects and sequences actions within a defined objective and can continue across multiple steps.

These models should not share one approval standard. Deterministic automation can often be tested against known inputs and outputs. AI assistance requires controls over grounding, review and use. Agency requires all of those controls plus authority, identity, tool access, memory, termination, exception handling and evidence of each intermediate action.

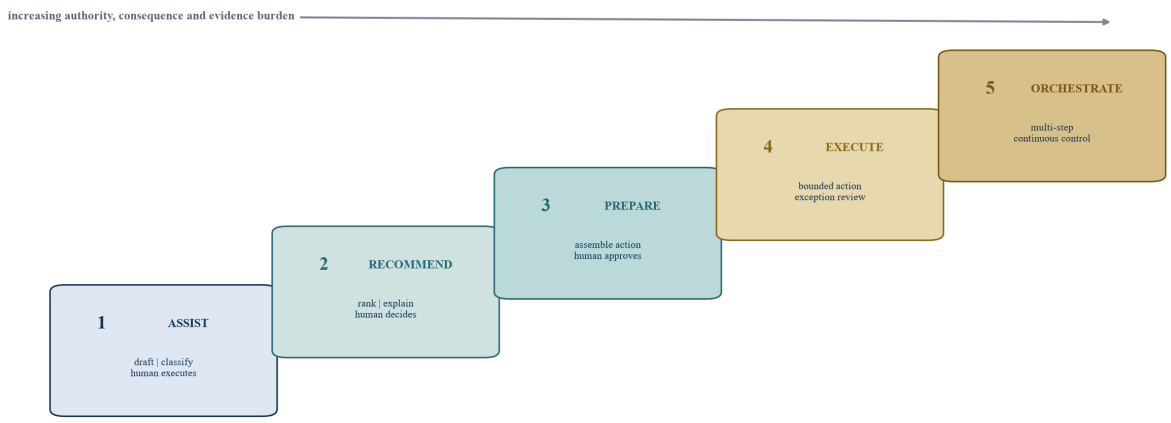
Autonomy should be decomposed rather than described with a single label. An agent may have freedom to plan while holding no execution rights. It may execute low-consequence internal updates while requiring approval before external communication. It may operate only on a named case, within a defined value limit, during a short time window and with a capped number of tool calls.

The company should choose the lowest level of agency that can produce the desired economic outcome. Many processes improve substantially through better data, deterministic controls and AI-assisted preparation. Full unattended execution is suitable only where the process is well understood, the action is reversible, the authority can be tightly bounded, the evidence is complete and the failure can be contained.

An autonomy ladder makes progression observable. A use case advances only after it meets defined evidence gates at its current level. Production incidents, material process changes, new tools, model changes or weak monitoring should trigger reassessment or a return to a lower level.

Figure 2. The autonomy ladder

USE THE LOWEST AUTONOMY LEVEL THAT DELIVERS THE ECONOMIC OUTCOME



Authority expands only after evidence, reversibility and containment have been demonstrated at the preceding level.

4. Build a process and decision inventory

Agent adoption should begin with the process, decision and evidence trail. A catalogue of appealing demonstrations can miss the operating causes of cost and delay. The company should inventory the back office by value stream: order to cash, procure to pay, record to report, hire to retire, case to resolution, regulatory reporting and management information.

Each process should be decomposed into triggers, inputs, decisions, actions, systems, owners, handoffs, controls, exceptions and outputs. Management should measure volume, cycle time, error, rework, backlog, cost, monetary exposure, data sensitivity and service consequence. The inventory should identify where judgment is genuinely required and where missing data or inconsistent policy creates the appearance of judgment.

The decision inventory is especially important. A decision may approve a payment, release a customer order, classify a complaint, change a supplier record, recognise a journal, select a candidate or communicate a conclusion. The record should identify who is affected, the evidence required, the relevant policy, the permitted discretion, the appeal or correction path and the legal or control consequence.

Prioritisation should combine economic value and control feasibility. High-volume, low-consequence, evidence-rich work can be an attractive starting point. High-consequence, ambiguous or poorly evidenced work should first receive process repair and stronger data controls. A company should avoid automating a broken exception process whose root cause remains unresolved.

The inventory becomes the governing source for use-case approval. It connects the claimed value to a measurable baseline and prevents a technology team from granting authority without the

process owner, risk owner, data owner and control owner agreeing on the design.

Table 2. Process and decision inventory

Inventory field	Core question	Evidence	Decision use
trigger and objective	what starts the work and what outcome closes it?	source event, service standard and completion record	define scope and stop rule
input and decision	which facts and policy determine the next step?	documents, master data, policy and approval record	test grounding and permitted discretion
action and consequence	what can change in the business or affect a person?	system transaction, communication and value exposure	set authority and approval level
exception and failure	where does normal processing break?	error codes, backlog, disputes and incidents	design escalation and safe failure
control and owner	who remains accountable and which control must operate?	control matrix, segregation and role description	assign governance and evidence owner
baseline and value	what cost, delay, error or cash exposure can improve?	volume, time, quality, cash and service measures	build the business case and scorecard

Measures are selected for the operating context and should be based on reconciled evidence.

5. Define the authority envelope

The authority envelope translates policy into machine-enforceable limits. It should state the objective, permitted data, permitted tools, actions, values, counterparties, jurisdictions, operating hours, duration, call count, approval points and stop conditions. A natural-language instruction alone is insufficient because it may be interpreted differently as context changes.

Authority should be narrow by default. An invoice agent might read approved invoices and purchase orders, request missing internal evidence, prepare a payment proposal and update a case record. It should hold no standing right to change supplier bank details or release payment. A collections agent might draft a follow-up based on approved facts and terms while a human approves external communication above a defined sensitivity or customer tier.

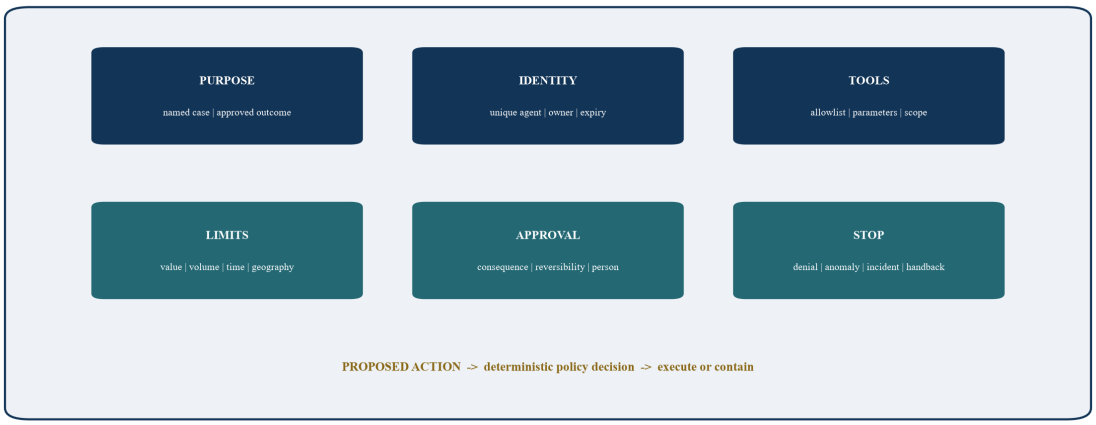
The envelope should be attached to a named business owner and risk tier. Temporary authority should expire. Value and volume limits should be cumulative as well as transactional so that many small actions cannot bypass a daily or case-level cap. Geographic, entity and currency limits should match the legal and operating perimeter.

Policy enforcement should occur outside the reasoning model. A deterministic service should validate each proposed tool call against identity, purpose, permissions, value, case status and approval. The model can explain why it proposes an action; the policy engine decides whether the action may proceed.

Every denial should create structured evidence. Repeated denials can reveal a poor prompt, weak process design, emerging misuse or an authority envelope that needs review. Envelope changes should follow version control, testing and approval, with the previous version retained.

Figure 3. The authority envelope

AUTHORITY IS EXPLICIT, LIMITED, EXPIRING AND MACHINE-ENFORCED



Each proposed action must pass cumulative limits, policy and approval before execution.

6. Give every agent a separate identity

An agent should not operate through a shared employee account or an unrestricted service credential. It needs a unique machine identity connected to a named owner, a defined purpose, a risk tier, a version and an expiry. This supports least privilege, attribution, credential rotation, revocation and reliable audit.

Permissions should be task-specific. Read access, proposal rights and execution rights are distinct. The agent should receive short-lived credentials at the moment of need after the policy service validates the case and approval state. Secrets should remain in a managed vault and should never be placed in prompts, memory, logs or documents.

The identity architecture should preserve segregation of duties. An agent that creates a supplier should not approve the supplier or release a payment. An agent that prepares a journal should not post and approve the same journal. A human approval does not cure a conflict when the approver lacks independence, evidence or adequate time.

Agent-to-agent calls require the same discipline. The receiving agent should authenticate the caller, validate the delegated scope and refuse authority that cannot be traced to the originating principal. Delegation chains should have a maximum depth and should preserve the original owner, case and cumulative limits.

Emergency revocation must be practical. The company should be able to disable the identity, revoke active tokens, stop queued actions and isolate the case without waiting for the model provider. Access recertification should occur after material process or role changes and on a periodic schedule proportionate to risk.

7. Control data, prompts, tools and memory

The agent's effective operating environment includes more than the model. It includes system instructions, user instructions, retrieved documents, memory, tool descriptions, application responses and third-party content. Any of these can be inaccurate, malicious, stale or outside the authorised purpose.

Data access should follow classification, purpose and minimum necessity. The company should map personal data, confidential commercial information, payment data, employment data and legally privileged material. Retrieval should enforce entity, case and user boundaries. Sensitive fields can be masked until the task and approval require them.

Untrusted content should be treated as data rather than instruction. A message in an invoice attachment that asks the agent to ignore policy or send information elsewhere should have no authority. The architecture should separate trusted control instructions from retrieved content, sanitise tool outputs where appropriate and test prompt-injection scenarios.

Tools require allowlists and parameter controls. A generic database query, command shell or unrestricted browser can give the agent a wider operating surface than the business purpose requires. Purpose-built tools with typed inputs, narrow actions, deterministic validation and idempotency provide a stronger boundary.

Memory should have a defined purpose, retention period and deletion path. The company should distinguish temporary case state, approved reference knowledge and learned preferences. Persistent memory can carry outdated facts, unauthorised personal data or one case's instructions into another. Memory writes should be explicit, attributable and reviewable.

Table 3. Control register for the agent operating surface

Surface	Principal risk	Preventive control	Detective or recovery control
system and user instructions	ambiguous or conflicting authority	signed, versioned control prompt and precedence rules	instruction trace and denied-action review
retrieved documents	prompt injection, stale facts or wrong entity	trusted source classes, case filters and content isolation	provenance log and adversarial-content tests
tools and connectors	excessive capability or unsafe parameters	narrow typed tools, allowlist, policy validation and idempotency	tool-call telemetry, reconciliation and revocation
credentials and secrets	disclosure, reuse or privilege escalation	vault, short-lived token and least privilege	secret scanning, rotation and access alerts
working memory	cross-case leakage or outdated context	case isolation, retention rule and explicit write permission	memory audit, expiry and deletion evidence
external models and services	data transfer, change or concentration exposure	contract, configuration, data minimisation and fallback	provider monitoring, version tests and exit plan

Technical implementation should be validated by the company's security, privacy and control owners.

8. Put deterministic controls around probabilistic work

An agent can help interpret messy information while deterministic controls should govern facts that must be exact. Bank account formats, tax identifiers, invoice arithmetic, currency, duplicate detection, approval limits, ledger periods, sanctions status and segregation rules should be validated through controlled data and code.

The workflow can use the model to locate a likely invoice number, explain a discrepancy or propose a category. A deterministic service should then verify the identifier against the source system, calculate the amount, test the rule and return a pass, fail or exception. The agent should not convert a low-confidence interpretation into an executed transaction.

Reconciliation is central. The company should compare requested, approved, executed and recorded actions. Each transaction needs an idempotency key so a retry cannot create a duplicate. Batch totals, control accounts and exception queues should reconcile to the source and destination systems. A successful tool response is evidence of a call, not proof of the intended accounting or operating outcome.

Confidence should not serve as a substitute for authority. A highly confident model can still be wrong, and a low-confidence output may be useful when clearly routed to a person. Decision rules should use consequence, evidence completeness, policy and reversibility alongside any model score.

The design should support graceful degradation. If the model, connector or data source is unavailable, the process should pause, revert to a simpler rule or hand work to a person according to a documented service plan. The company should know which operational commitments depend on the agent and how long they can tolerate interruption.

9. Design human approval around consequence and reversibility

Human involvement creates control only when the reviewer receives the right evidence, understands the decision, holds authority and has time to challenge it. A queue of hundreds of routine approvals can become ceremonial. The design should concentrate human judgment on material, unusual or irreversible actions.

Approval levels should reflect consequence. Low-value internal classification can be reviewed by sample. A customer communication may require approval when it changes contractual position or concerns a vulnerable individual. A payment, bank-detail change, journal posting, employment action or regulatory submission can require dual control or specialist review.

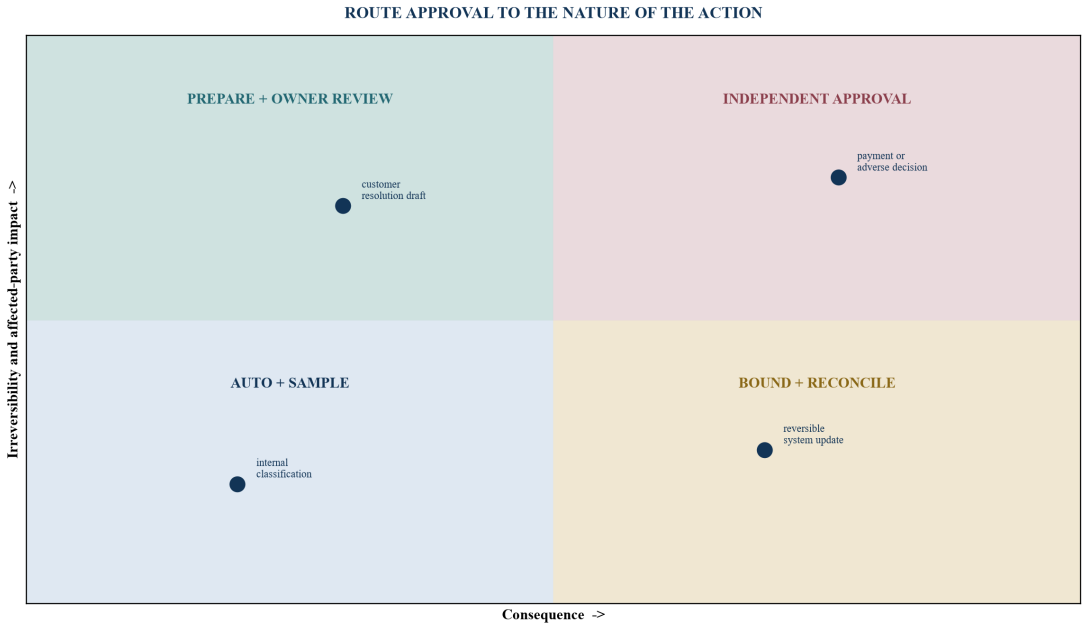
The approval pack should show the proposed action, purpose, source evidence, policy rule, material assumptions, exceptions, value, affected party, alternatives and downstream consequence. The reviewer should be able to approve, amend, reject or return the case. A rejection reason should feed process improvement without automatically training the system.

Reversibility affects the gate. Updating an internal task status is easier to reverse than sending funds, disclosing data or issuing a final decision. Time sensitivity also matters. The system should escalate before a deadline rather than pressure a reviewer into a weak decision at the last moment.

Human performance should be monitored. Repeated rapid approvals, high override rates, frequent post-approval corrections or large reviewer variation can indicate poor evidence, inadequate

training or automation bias. Accountability stays with the assigned owner even when an agent prepared the work.

Figure 4. Consequence-based approval routing



Higher consequence and lower reversibility require stronger evidence and independent approval.

10. Govern multi-agent and third-party dependencies

An agent rarely operates alone. It can depend on a model provider, cloud platform, orchestration layer, vector database, connector, identity service and business application. A multi-agent design adds delegated tasks and communications among systems with different owners and capabilities.

The dependency register should identify provider, service, data location, model or version, criticality, contractual rights, sub-processors, retention, incident notification, service level, audit evidence, change process and exit path. Concentration should be visible where several critical workflows rely on one provider or credential plane.

The company should test provider changes. Model behaviour, tool schemas, rate limits, safety settings and data terms can change. A release should move through a controlled environment with regression tests before production. Emergency provider changes should trigger a defined containment and fallback path.

Agent-to-agent messages should use authenticated, typed and bounded protocols. The receiving system should treat free-form content as untrusted until validated. Each task should carry purpose, originating principal, delegated scope, expiry and evidence requirements. A child agent should not obtain more authority than its parent.

Contracts should address data use, confidentiality, security, intellectual property, audit, incident, change, service continuity, portability and termination. Legal advice is required for the relevant jurisdiction and operating model. Technical safeguards remain necessary because a contractual remedy occurs after an incident and may not restore data or customer trust.

11. Preserve financial and procurement control

Finance and procurement offer valuable agent use cases because they contain repeated evidence assembly and exception resolution. They also contain direct cash, reporting, fraud and supplier risks. The agent architecture should extend the existing internal-control environment rather than create a parallel route around it.

In procure to pay, agents can check onboarding completeness, compare invoice and purchase-order evidence, identify duplicates, prepare exception narratives and route approvals. Bank-detail creation or change should remain subject to independent verification outside the communication channel used for the request. Payment proposals should be reconciled to approved invoices and released through established treasury authority.

In record to report, agents can assemble reconciliations, trace variances, prepare journal support and draft commentary. Period status, account ownership, evidence and approval should be machine-validated. Material or unusual journals should receive independent review. The system should preserve the exact source, transformation and decision trail used for each proposed entry.

In order to cash, agents can identify documentation gaps, classify disputes, prepare collection actions and coordinate internal owners. External communication should use approved facts and authority. Credit-limit changes, concessions, write-offs and legally sensitive escalation should follow existing approvals.

The design should avoid control dilution through volume. An agent can produce more proposals than reviewers can assess. Capacity limits, risk-based routing and queue ageing should be part of the release decision. Financial benefit should be measured through cycle time, rework, error, cash timing and capacity released, with service and control indicators beside them.

Table 4. Finance and procurement control matrix

Use case	Agent contribution	Deterministic gate	Human or independent control
supplier onboarding	assemble evidence and identify gaps	required fields, duplicate, tax and account checks	owner approval; independent bank-detail verification
invoice exception	compare documents and explain mismatch	arithmetic, currency, duplicate and tolerance rules	exception owner approves resolution
payment proposal	group approved due items and forecast cash	approval status, value, due date and idempotency	treasury release under existing mandate
reconciliation	match records and prepare break analysis	control total, period and source validation	account owner certifies unresolved items
journal preparation	assemble support and proposed entry	period, account, balance and segregation rules	authorised reviewer approves and posts
collections case	locate evidence and prepare next action	customer, amount, due date and communication rule	owner approves sensitive or binding action

The matrix complements the company's existing authority, accounting and treasury policies.

12. Protect customers, employees and affected individuals

Back-office agents can affect people even when the use case appears administrative. A customer may receive a collection escalation. A supplier may be blocked. An employee may be ranked, investigated or denied a benefit. A complaint may be classified in a way that changes its treatment. The company should assess the individual and relationship consequence of each decision.

The assessment should identify the affected group, data used, purpose, legal basis, potential harm, fairness concern, transparency duty, human review and challenge path. Personal data should be limited to the authorised purpose. Special-category or sensitive data requires heightened analysis and safeguards under applicable law.

Communications should accurately describe the role of automation where required and provide a practical route to a person. Affected individuals should not be trapped in a loop where one agent reviews another agent's decision without meaningful independent intervention. Complaints and appeals should preserve the original evidence, decision path and versions used.

Employment use cases require particular care. Candidate screening, performance analysis, absence management, disciplinary action and termination can carry significant legal and human consequences. The company should obtain current employment and data-protection advice, involve qualified HR owners and apply independent review.

Fairness should be tested in context. Aggregate accuracy can conceal uneven error or burden across groups, geographies, languages or customer segments. The company should define relevant outcome measures, examine error patterns and monitor changes after deployment. A fairness review should also consider whether the process itself is appropriate for agent use.

13. Design for cyber security and safe failure

The threat model should cover the complete agent system. Attackers can manipulate prompts, documents, tool outputs, memory, connectors, credentials and dependencies. They can seek data, money, persistence, disruption or influence over business decisions. Ordinary configuration error can produce similar consequences without malicious intent.

Secure design begins with narrow tools, isolated environments, least privilege, input and output controls, protected secrets and explicit trust boundaries. Development should use controlled repositories, reviewed dependencies, reproducible builds and testing. Deployment should protect infrastructure, configuration and model access. Operation should monitor behaviour, vulnerabilities, provider changes and incidents.

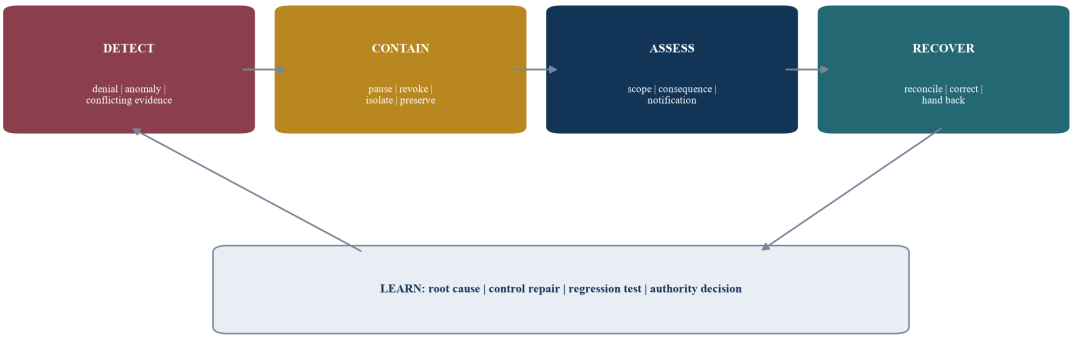
The safe-failure design should state what happens when evidence conflicts, a tool returns an unexpected structure, a limit is reached, a credential fails, a model changes, a dependency is unavailable or the agent behaves outside its tested distribution. The preferred response is usually pause, preserve state, contain pending actions and route the case to a person.

Kill capability must be independent of the agent. Operators should be able to disable an identity, connector, use case, model or entire orchestration layer. Queued work and partial transactions need a recovery procedure. Business continuity should cover the manual or simplified process needed during interruption.

Incident response should preserve prompts, retrieved evidence, tool calls, approvals, model and configuration versions, timestamps and executed transactions. The team should determine scope, contain active authority, notify relevant owners and advisers, reconcile outcomes, correct affected records and decide whether regulatory or contractual notification is required.

Figure 5. The exception and containment loop

A SAFE AGENT STOPS, CONTAINS AND PRESERVES THE EVIDENCE NEEDED TO RECOVER



Safe failure preserves evidence, contains authority and returns unresolved work to an accountable owner.

14. Make every material action reconstructable

Auditability requires more than a transcript. The evidence record should reconstruct who or what initiated the case, which authority applied, which data and document versions were retrieved, which model and configuration were used, which intermediate decisions occurred, which tools were called, which policy checks operated, who approved the action and what the destination system recorded.

Logs should be structured, time-synchronised, access-controlled and protected from unauthorised alteration. They should avoid unnecessary secrets and personal data. Retention should align with legal, regulatory, contractual and operational requirements. The company should be able to join identity, agent, application and transaction logs through a common case and action identifier.

Provenance should distinguish source facts from model-generated interpretation. A reviewer should be able to open the controlling invoice, contract, policy or system record. If the agent combined several sources, the evidence should show which statement came from which source and whether the source was current at the decision time.

Monitoring should address outcomes as well as technical activity. The company should reconcile executed transactions, customer communications, ledger entries, access changes and exception dispositions. Sampling should cover routine cases, while material actions can receive complete review.

Access to the evidence itself is sensitive. Audit logs can contain business processes, security details and personal data. Role-based access, purpose limits and independent monitoring are required. The company should test whether it can reconstruct a case before relying on the

evidence design in production.

15. Test the system before production authority

Testing should begin with the process and control design. The team should confirm that the objective is clear, the source data is reliable, the policy can be encoded, the authority is narrow and the failure path is practical. A model benchmark cannot compensate for an ambiguous process or missing evidence.

The test set should include normal cases, rare but important cases, incomplete records, conflicting evidence, stale documents, duplicate requests, adversarial instructions, unauthorised tool calls, value-limit breaches, unavailable dependencies, delayed approvals and attempts to cross entity or case boundaries. The team should test multi-step accumulation, because a sequence of individually plausible actions can create a material exposure.

Evaluation should measure task outcome, factual grounding, policy compliance, tool correctness, duplicate prevention, escalation quality, evidence completeness, latency and cost. Security tests should cover prompt injection, data leakage, privilege escalation, dependency compromise and denial of service. Privacy and fairness tests should reflect the affected population and legal context.

Release should progress through offline evaluation, sandbox, shadow mode, human-approved production and bounded unattended operation. Each stage should have entry and exit criteria. Shadow mode compares recommendations with actual process outcomes without granting execution rights. Limited production should cap counterparties, values, users, volume and time.

Regression testing is required after changes to model, prompt, data source, tool, policy, connector or material business process. A release record should show the approved version and evidence. The company should retain the ability to roll back or remove authority when results deteriorate.

Table 5. Test and release gates

Gate	Required evidence	Production permission	Failure response
design review	process, decision, data, threat and authority maps	build in isolated environment	repair process or narrow the use case
offline evaluation	representative and adversarial test results	no live authority	correct model, data, tool or policy design
sandbox	end-to-end tool and control evidence	synthetic or non-production actions	contain and retest
shadow mode	comparison with real cases and owner decisions	read-only production context	resolve outcome and evidence gaps
human-approved production	stable metrics, approval quality and reconciliation	bounded proposals and approved actions	reduce scope or return to shadow mode
bounded unattended operation	sustained control, monitoring and safe-failure evidence	reversible actions within strict limits	revoke authority and activate recovery

Thresholds and sample sizes should reflect use-case risk, volume and applicable assurance requirements.

16. Establish an accountable operating model

The operating model should assign business accountability before technical delivery. The business process owner owns the outcome and operating policy. The risk owner approves the risk treatment. The data owner governs source quality and permitted use. Security owns the threat model and access architecture. Technology owns the reliable platform. Internal audit or an independent assurance function evaluates design and operation according to its mandate.

An AI governance forum can approve use cases, risk tiers, authority envelopes, releases and material changes. The forum should include business, technology, security, privacy, legal, compliance, finance, HR or other functions according to the use case. The board or relevant committee should approve risk appetite and receive a concise portfolio view.

The agent register should contain owner, purpose, process, affected parties, data, model, tools, identity, authority, risk tier, approvals, test evidence, release, monitoring, incidents, dependencies and retirement status. The register should distinguish experiments from production systems while still applying suitable controls to experiments that use sensitive data or live tools.

Change management should address the workforce. Employees need training on agent capability, limitations, escalation, evidence review and prohibited use. Role design should identify which work is removed, which judgment remains and which new control responsibilities arise. Capacity released should be deliberately redeployed rather than assumed.

Independent challenge should focus on evidence. Owners should demonstrate the control and outcome rather than rely on policy statements. Where skills are limited, the company can use external specialists for architecture, cyber security, privacy, legal, audit or process design while retaining internal accountability.

17. Execute through a 120-day sequence

The first 30 days establish truth and boundaries. Management selects a narrow value stream, maps the process and decisions, reconciles the baseline, classifies data, identifies affected parties, assigns owners, completes the threat model and drafts the authority envelope. A manual or AI-assisted prototype can test whether the proposed value exists before any production execution right is granted.

Days 31 to 60 build the controlled path. The team creates the unique agent identity, narrow tools, policy gate, case evidence model, approval pack, logging, safe-stop procedure and test set. Offline and adversarial evaluation should resolve defects in instruction handling, source grounding, tool parameters and exception routing.

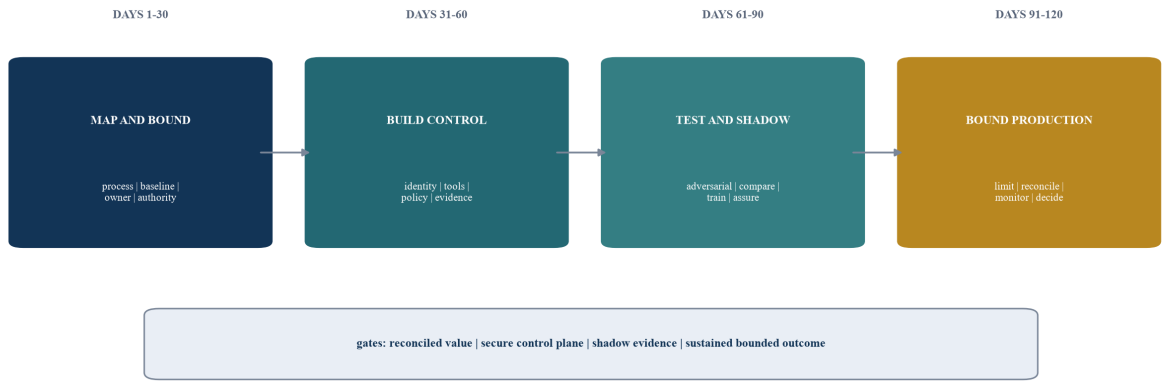
Days 61 to 90 move through sandbox and shadow mode. The team compares agent proposals with real process outcomes, measures baseline and control performance, tests continuity and incident procedures, completes third-party due diligence and trains reviewers. Production scope remains read-only or human-approved until exit criteria are met.

Days 91 to 120 use bounded production. The company limits value, volume, entities, counterparties, actions and time. It reconciles every executed action, monitors exceptions and reviewer behaviour, and verifies the claimed operating benefit. The governance forum decides whether to retain, expand, narrow or withdraw authority.

Expansion should reuse the control plane while repeating use-case analysis. A common identity, policy, evidence and monitoring architecture can lower marginal implementation cost. Each new process still needs its own decision, data, consequence and legal assessment.

Figure 6. The 120-day controlled adoption path

PRODUCTION AUTHORITY FOLLOWS EVIDENCE, TESTING AND ACCOUNTABLE OWNERSHIP



Authority advances through evidence gates from process truth to bounded production.

18. Measure control, operating value and enterprise value

The scorecard should distinguish technical activity from business outcome. Tool calls, tokens and model latency can support capacity planning. They do not prove that the process improved. The board needs evidence of cycle time, backlog, error, rework, cash timing, service quality, control exceptions, incidents and capacity released.

Control measures should include unauthorised-action denials, escalation rate, evidence completeness, duplicate prevention, reconciliation breaks, approval overrides, post-action corrections, security events, privacy events, provider changes and recovery time. A low denial rate is useful only when the authority and monitoring are effective.

Value should be measured against a reconciled baseline. Labour capacity can be estimated from observed volume and handling time, then validated through actual role and workload changes. Cash effects should reconcile to transactions and bank or ledger evidence. Avoided loss should be stated cautiously and supported by observed incidents or a documented risk model.

Enterprise value can improve through durable cost capacity, faster cash conversion, better customer or supplier service, stronger controls, more scalable operations and lower operational risk. Transaction evidence should show that the benefit persists, depends on controlled infrastructure and can survive key-person or provider changes. Buyers and lenders may discount a benefit that relies on an opaque system or weak evidence.

Commercial delivery should connect the diagnostic to execution. External advisers can support process mapping, business-case design, control architecture, implementation governance, vendor selection, financial modelling and board reporting. Specialist legal, data-protection, cyber-security, audit, employment, accounting and regulatory work should be performed by

qualified advisers. Success should be evidenced through accepted deliverables, signed mandates, collected fees and sustained client outcomes.

Table 6. Board scorecard for controlled agent adoption

Dimension	Primary evidence	Board measure	Escalation trigger
operating outcome	process and transaction record	cycle time, backlog, error, rework and service	outcome deteriorates outside tolerance
authority and policy	identity, policy and tool-call logs	denied, approved and executed actions by risk tier	action occurs outside the envelope
evidence and approval	case file and reviewer decision	completeness, override, correction and ageing	material action lacks reconstructable evidence
security and privacy	alerts, incidents and access record	events, exposure, containment and recovery time	suspected data, credential or integrity compromise
financial value	baseline, ledger, cash and capacity evidence	recurring benefit, cash timing and cost to run	claimed benefit cannot be reconciled
resilience and dependency	provider, continuity and change records	concentration, change failures and fallback readiness	critical service lacks viable containment or fallback

Targets are illustrative management choices and should be approved for each process and risk tier.

Implementation conclusion

AI agents can improve the mid-market back office when authority is engineered with the same care as the workflow. The practical sequence begins with a reconciled process and decision inventory, selects the lowest useful autonomy level, defines an explicit authority envelope and gives every agent a separate identity. Deterministic policy, validation and reconciliation surround probabilistic interpretation.

Human approval is concentrated on consequence and reversibility. Data, tools, prompts, memory and providers remain inside defined trust boundaries. Testing progresses from offline evidence through sandbox, shadow mode and bounded production. Logging makes material actions reconstructable, while independent kill and recovery mechanisms contain failure.

A 120-day programme can establish the control plane and demonstrate operating value in one narrow process. Expansion follows the same gates for each additional use case. This provides the board with evidence on both productivity and risk and creates a more credible foundation for scalable, transaction-ready enterprise value.

References

[1] National Institute of Standards and Technology. AI Risk Management Framework, including AI RMF 1.0, Playbook and Generative Artificial Intelligence Profile. https://www.nist.gov/itl/ai-risk-management-framework

[2] National Institute of Standards and Technology. Artificial Intelligence Risk Management Framework: Generative Artificial Intelligence Profile, NIST AI 600-1, July 2024. https://nvlpubs.nist.gov/nistpubs/ai/NIST.AI.600-1.pdf

[3] National Institute of Standards and Technology. AI Agent Standards Initiative, created 17 February 2026 and updated 20 April 2026. https://www.nist.gov/artificial-intelligence/ai-agent-standards-initiative

[4] Dubai International Financial Centre. Regulation 10 on processing personal data through autonomous and semi-autonomous systems. https://www.difc.com/business/registrars-and-commissioners/commissioner-of-data-protection/regulation-10

[5] Dubai International Financial Centre. Data Protection Guidance for Compliance, including Processing Personal Data Through Autonomous and Semi-Autonomous Systems. https://www.difc.com/business/registrars-and-commissioners/commissioner-of-data-protection/guidance

[6] UK National Cyber Security Centre. Guidelines for Secure AI System Development, 27 November 2023. https://www.ncsc.gov.uk/collection/guidelines-secure-ai-system-development

[7] European Commission. AI Act regulatory framework and application timeline, updated July 2026. https://digital-strategy.ec.europa.eu/en/policies/regulatory-framework-ai

[8] European Commission. Transparency obligations for providers and deployers of certain AI systems, 20 July 2026. https://digital-strategy.ec.europa.eu/en/news/commission-publishes-guidelines-transparency-obligations-providers-and-deployers-certain-ai-systems

[9] Financial Conduct Authority and Bank of England. Artificial intelligence in UK financial services: 2024 survey findings. https://www.fca.org.uk/publications/research-notes/ai-uk-financial-services

[10] OECD. OECD AI Principles, adopted 2019 and updated 2024. https://www.oecd.org/en/topics/ai-principles.html

[11] Singapore Personal Data Protection Commission. Singapore's Approach to AI Governance and Model AI Governance Framework. https://www.pdpc.gov.sg/help-and-resources/2020/01/model-ai-governance-framework

[12] International Organization for Standardization. ISO/IEC 42001:2023, Artificial intelligence management systems. https://www.iso.org/standard/42001

[13] Cybersecurity and Infrastructure Security Agency. CISA and UK NCSC Unveil Joint Guidelines for Secure AI System Development, 26 November 2023. https://www.cisa.gov/news-events/alerts/2023/11/26/cisa-and-uk-ncsc-unveil-joint-guidelines-secure-ai-system-development

About the Author

Chennakeshav Adya is an Independent Researcher and Managing Partner at Matchpoint Partners. His work focuses on corporate finance, capital strategy, transaction execution and the operating systems that connect financial evidence, board decisions and measurable enterprise performance.