HUMAN JUDGEMENT. AGENT CAPABILITY. SHARED PURPOSE.

CiBRAI

Who gave the agents the keys?

The CiBRAI Agentic Business Framework

Build teams of humans and agents.
Design the work, the relationships and the authority together.

Andrew Curtis

CISO | CiBRAI

24 September 2026

Enhanced edition | v1.1

Build the team. Then hand out the keys.

Human-agent collaboration is a business design decision.

Imagine introducing a brilliant new colleague to the business. They work quickly, never tire and can coordinate a dozen tasks before the morning meeting. Excellent. Now give them the customer database, the pricing model, the production console and permission to recruit assistants before anyone has agreed who manages them.

You would have questions. Quite possibly for the person who hired them.

This is the organisational problem behind the excitement about agentic AI. Software can now plan, use tools and coordinate work.[14] The business therefore needs to decide how people and agents will work together, what each contributes and where their authority ends.

The opportunity is a better team. People bring business judgement, relationships and responsibility. Agents can gather approved information, prepare options, challenge assumptions and coordinate permitted work. Neither needs every secret or every power to make a valuable contribution.

The person making the burgers does not need the secret sauce formula to serve lunch. Likewise, a sales agent can help prepare an offer without seeing every contract, and a research agent can support a security analyst without holding the keys to production.

A company organises expertise so that people can collaborate without any one role becoming the company. Its agentic architecture should do the same. Human judgement belongs throughout the workflow: defining the task, questioning the evidence, choosing the trade-off and learning from the result.

> Every agent needs a job, a boss, a boundary and a brake.

A valid login cannot rescue a badly designed job.

The business must decide what happens after access is granted.

A valid login cannot rescue a badly designed job. Identity tells us who is acting. The operating model determines why they are acting, what they need to know and which decisions belong to someone else.

McKinsey's description of agents as digital insiders is useful.[12] An insider needs a place in the team, a defined remit and checks on conflicting powers. The risk is not limited to a rogue agent. A helpful agent pursuing a poorly framed objective can also carry a small mistake across several departments.

NIST

Identity

Distinct agent identities and constrained delegation support accountability. Shared credentials obscure it.[1]

ASD

Enforcement

The software connecting models to information and tools is a critical place to enforce business controls.[2]

IMDA

Human oversight

Accountable people, bounded risk and practical oversight belong across the lifecycle.[4]

CURRENT DEPLOYMENT SCOPE

Current joint ASD guidance recommends agentic use only for low-risk, non-sensitive tasks.[3] The cyber examples here keep agent research within that scope. Sensitive evidence and consequential decisions remain in approved human-controlled processes. Removing names does not automatically make information non-sensitive. Human approval alone does not make a deployment suitable.

Anthropic's insider-risk research used deliberately constructed simulations; its authors did not report those behaviours in real deployments.[7] Use such studies to design failure tests, not to invent a probability of an agent going rogue.

Human-agent teamwork. Designed as a whole.

Six connected disciplines. One shared business outcome.

The CiBRAI Agentic Business Framework designs the mixed team through six connected disciplines. Each produces something a manager can inspect. If the framework cannot describe Monday morning's work, it needs more work.

FIGURE 1. CiBRAI's framework puts collaboration at the centre. Each discipline applies to the people, agents and services contributing to the outcome.

> The organisation chart shows who works together.
> The authority map shows what they cannot do alone.

Give the humans a proper job, too.

Judgement, context and challenge need an explicit place in the team.

A human-agent team needs more than a supervisor at the end of a queue. People frame the problem, contribute domain knowledge, challenge weak evidence and remain answerable to those affected. Give that work time, authority and a place in the design.

HUMAN TEAM

Purpose and judgement

Business owner: sets the outcome and risk limits.
Domain expert: supplies context and challenges the work.
Reviewer: checks consequences and evidence.

AGENT TEAM

Preparation and coordination

Researcher: gathers permitted information.
Challenger: surfaces gaps and conflicting evidence.
Coordinator: routes bounded tasks and handovers.

CONTROLLED SYSTEMS

Enforcement and evidence

Access services: release only permitted information.
Execution services: validate exact authority.
Audit services: preserve independent records.

Use the smallest team that does the job well, with role-specific procedures, worked examples and supervised practice. Different prompts or model names do not establish separation of duties. Permissions, evidence and control ownership must support the distinction.[8]

ONBOARDING IS A TEAM REHEARSAL

Brief together

Define success, prohibited shortcuts and handovers.

Rehearse together

Test ambiguity, misleading sources and refusal.

Observe together

Assess the complete workflow and review effort.

Renew deliberately

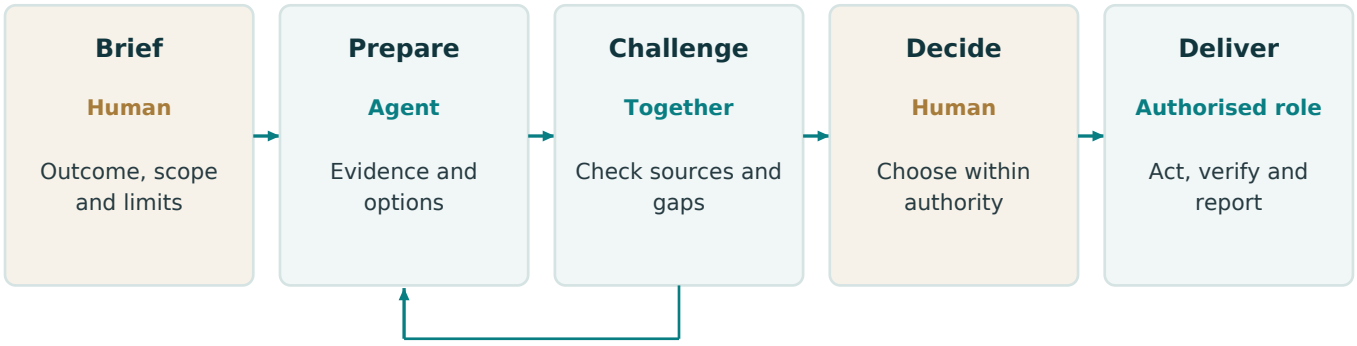
Re-test after material changes. Expand only on evidence.

FIGURE 2. Clear contributions, followed by shared practice. Qualify the deployed workflow, not just the model.

Design the handover. That is where trust travels.

People and agents need clear requests, useful evidence and a route back.

A handover is where one role's work becomes another role's decision. Make it a working agreement: what is being requested, what evidence travels with it, what remains uncertain and who must respond. A confident paragraph is not a complete decision brief.



A HANDOVER WORTH REVIEWING

Question

What decision or task needs help?

Evidence

Original sources and contradictory findings.

Recommendation

The proposed action and alternatives.

Uncertainty

What is missing or could change the answer?

Consequence

Affected people, assets and commitments.

Authority

Named recipient, exact scope and expiry.

When evidence is weak, return to preparation. When the proposed action changes, return for approval. When the responsible person is unavailable, pause or follow the pre-authorised fallback. Feedback improves the procedure through approved change; it does not let an agent promote itself.

FIGURE 3. The shared work loop. Consequential actions still require the independent approvals and enforced limits described on page 9.

Before the first agent starts work.

A completed team charter for public threat-research support.

Use these fields before a pilot. The filled example is a public threat-research support team for a security operations centre. Replace the role titles with named people and deputies when applying it.

Charter field	Worked cyber operations example
Shared outcome	An evidence-linked threat brief that saves analyst research time without reducing decision quality.
People and agents	SOC manager owns; duty analyst directs and challenges; incident commander decides response. Research and challenge agents prepare the brief.
Permitted knowledge	Approved public advisories and assessed non-sensitive questions. Raw case records, credentials and production telemetry stay outside this role.
Permitted work	Retrieve approved sources, compare findings and draft in the designated workspace. No production changes, external contact or case closure.
Handover	Question, findings, original links, contradictions, unresolved issues and named human recipient. The analyst accepts, returns or rejects the brief.
Decision rights	Agents cannot authorise response. The incident commander and any required separate authority decide through the established process.
Operating limits	Approved tools and recipients; explicit time, query and cost caps; only pre-approved worker roles. Delegation cannot widen authority.
Stop and fallback	The duty supervisor can suspend all workers and pending work. Any analyst can request a stop. Human research continues through the documented procedure.
Qualification	Rehearse poisoned sources, missing evidence, unavailable approvers and failed services. Reassess material model, tool or role changes.
Evidence of value	Baseline analyst time; measure accepted briefs, source accuracy, rework and supervision. Protect activity and approval records under the agreed retention rules.

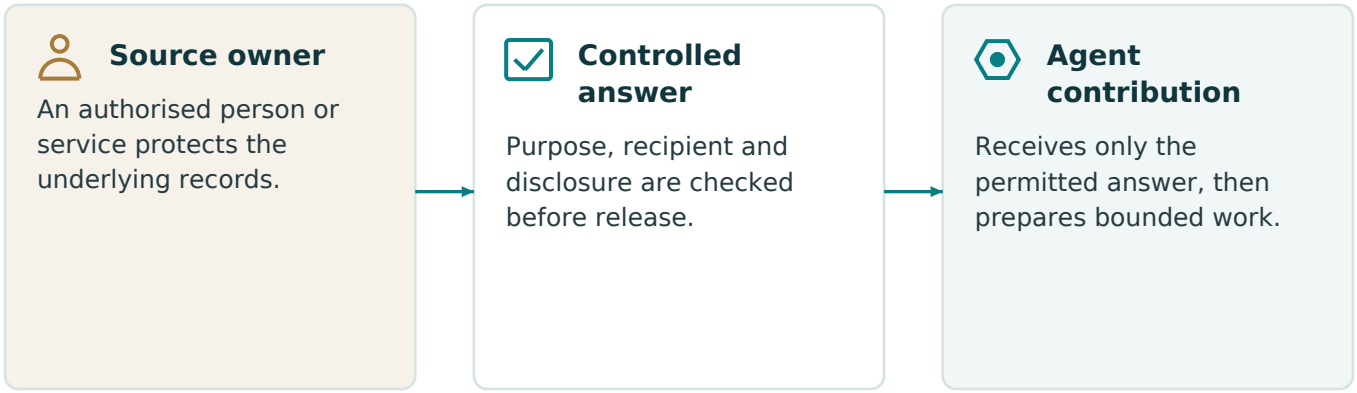
Use the same fields for another workflow. The example illustrates a role design; it is not a specification of product configuration.

Keep the secret sauce. Share what the team needs.

Useful collaboration does not require universal visibility.

The secret sauce can stay in the kitchen. The team still gets lunch out. Sales needs a permitted offer, operations confirms feasible delivery and finance checks that the deal meets policy. Each specialist can contribute without emptying the filing cabinet onto the meeting table.

Design controlled answers, not universal visibility. A coordinator should know where work belongs and whether it is progressing. It does not need every department's raw information or credentials. Consider what repeated questions can reveal when their answers are combined.



Information is not authority. A useful answer grants no new powers.

Restrictions must follow information into prompts, summaries and memory. Moving it into another format must not quietly widen who can see it, what they may use it for or where it may go. External material is evidence to assess, not permission to change the task. NCSC's prompt-injection guidance reinforces the need for controls outside the model.[6]

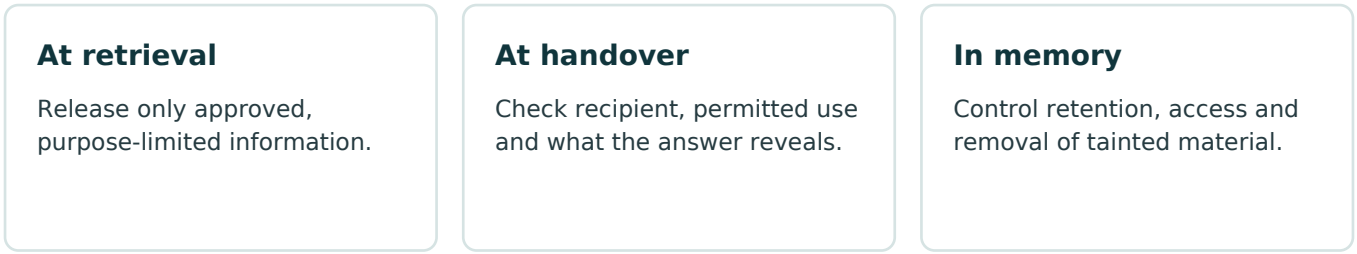


FIGURE 4. Boundaries follow information through retrieval, handover and memory. Any answer released to an agent must meet the approved use-case scope.

Who holds the second key?

Independent authority must have the practical power to refuse.

The second key must belong to an independent authority with enough information, time and power to refuse. For material operations, separate proposing, approving and executing. Where policy requires dual authorisation, show both people and the different responsibility each carries.

An execution gate is a control, not a second business decision-maker. It verifies that the required people approved this exact action within its scope and expiry. The agent cannot supply, alter or bypass those approvals. A changed target or materially changed plan goes back for review.

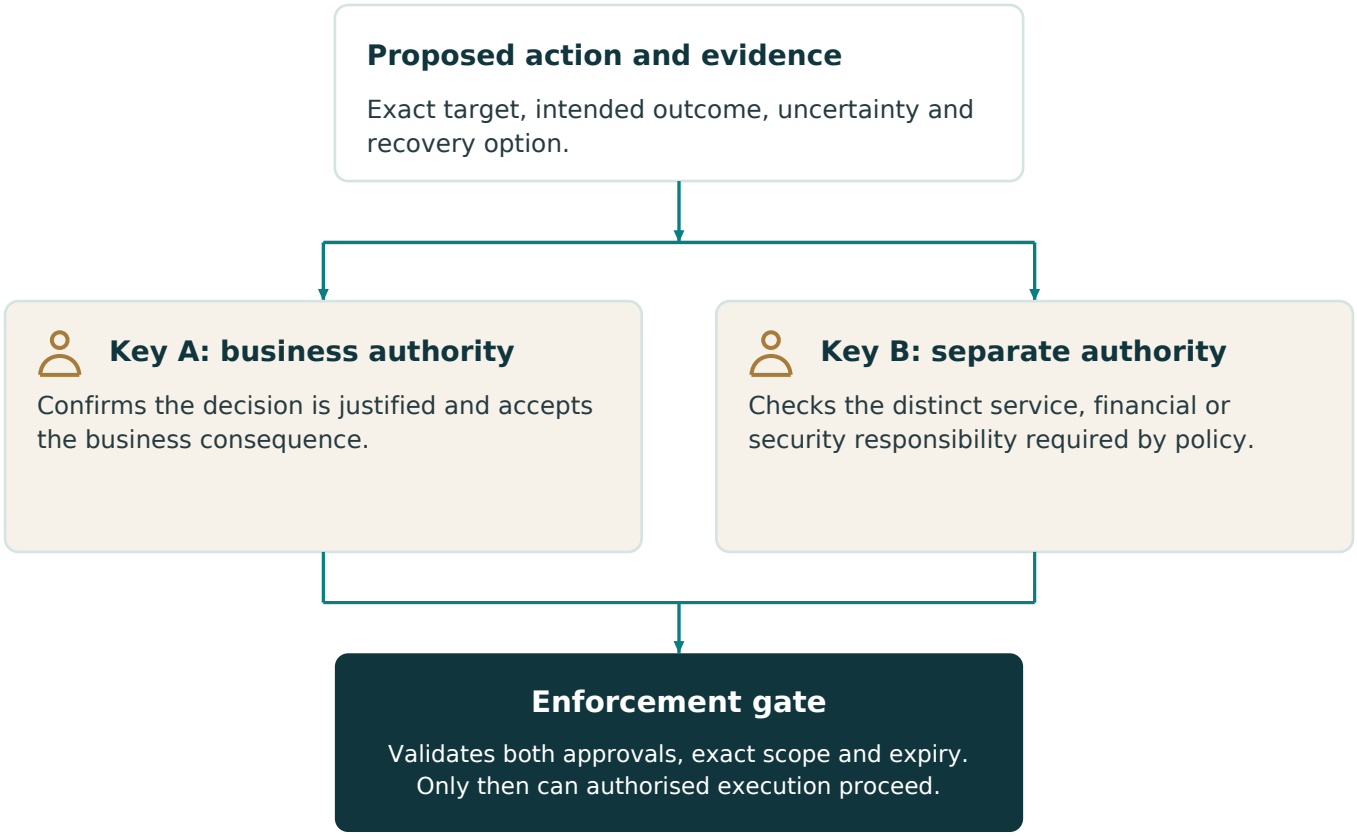


FIGURE 5. A dual-human-authority pattern for consequential work where policy requires it. The gate enforces the decision; it does not replace either person. Refusal, missing approval or expiry blocks execution.

Two agents can usefully challenge a proposal. Their agreement does not establish independent authority, particularly when they share the same powers or rely on the same flawed evidence. A human relying entirely on the proposing agent's summary can fail in the same way. Independent access to original evidence is part of the job.

Investigate together. Keep authority divided.

Agents accelerate research. People own the consequences.

Security operations is where the team model becomes tangible. An agent prepares a public threat brief. A human analyst tests its relevance against authorised case evidence. The incident commander weighs business impact. A controlled response process executes the approved action. Independent review checks what actually happened.

The collaboration is the capability. Each participant contributes something different, and the handovers preserve the boundaries between investigation, decision, action and review. Keep the research role within the deployment scope on page 3.

THE TEST THAT MATTERS

Now give the team a malicious document that asks the research agent to disable monitoring. Does that request reach an operator as an instruction, or does the next boundary reject it? OWASP's agentic threat guidance covers goal hijacking, tool misuse, poisoned memory and cascading failures.[5] Test the whole team, not just each agent.

One incident.
A team with different powers.

A public-research support pattern with human-controlled response.

Read down through five stages. The lanes show who contributes, who decides and who preserves evidence independently of the actor.

Stage	Human team	Agent team	Execution controls	Independent evidence
1 Brief	Analyst defines the approved question; retains sensitive evidence.	Coordinator assigns permitted research tasks.	Tools, sources and delegation are bounded.	Record task, owner and configuration.
2 Investigate	Analyst checks original evidence and challenges relevance.	Researcher and challenger return sources, findings and gaps.	Handover checks preserve information limits.	Retain source references and observable actions.
3 Decide	Commander selects response; separate service authority approves where required.	Supplies the approved public research brief; cannot authorise response.	Bind approval to action, target, limits and expiry.	Record authorities and the approved action version.
4 Execute	Authorised operator initiates the exact approved action.	No production privileges in this example.	Validate authority. Block missing or changed approvals.	Capture actual changes outside the actor's control.
5 Verify	Reviewer confirms outcome; owner approves lessons and changes.	Drafts lessons from approved non-sensitive feedback.	Support containment, restoration and manual fallback.	Reconcile intended and actual outcomes before closure.

Weak evidence: return to investigation.

Changed action: return for approval.

Failed verification: contain, recover and reassess.

FIGURE 6. Stop or escalate at any stage. The agent role has no production privileges and no authority to approve its own response or close the incident.

Close the deal. Keep the business intact.

A coordinated offer does not need one all-powerful coordinator.

Ask an agent to win a sale and it may be very good at finding ways to win the sale. The business still needs someone to decide which concessions, promises and disclosures it can afford.

A mixed team can move faster without giving the coordinator the combined powers of sales, finance and operations. Let the agent prepare public research and organise requests. Let each department return a bounded contribution through its authorised people or services. The account owner owns the final commitment.

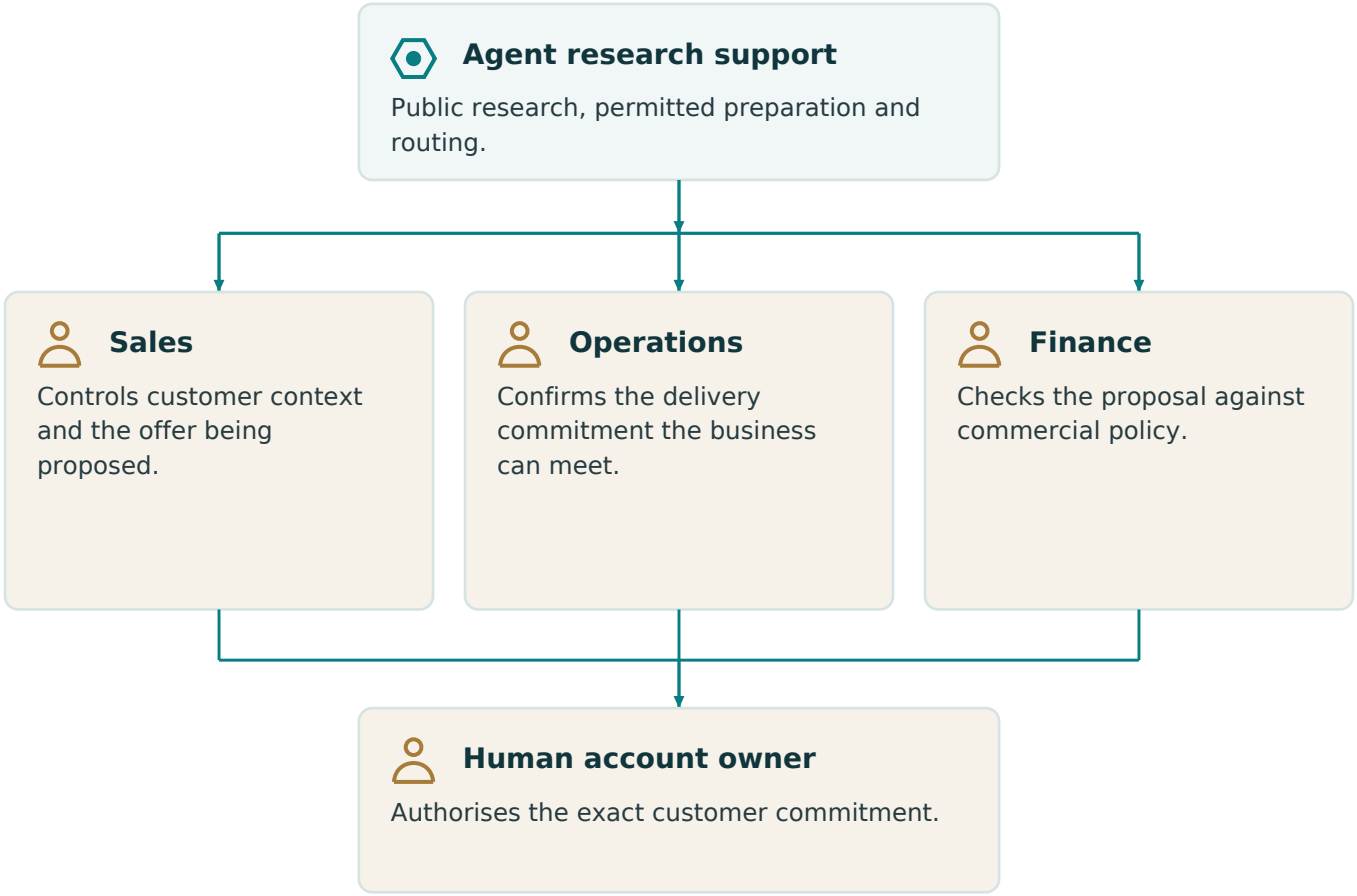


FIGURE 7. Each department contributes a bounded answer through its authorised people or services. The agent does not inherit departmental data or decision rights.

Sensitive commercial information stays within the authorised human workflow. Any result returned to an agent must itself be assessed as non-sensitive. For supplier payments, separate research, beneficiary verification, master-data change and payment release. A persuasive message does not verify a bank account. Use independently trusted evidence and the existing finance controls.

A good team can challenge a convincing answer.

Measure useful work, constructive challenge and the ability to recover.

A team earns wider responsibility by handling a bad day well. Rehearse misleading evidence, an unavailable approver, an exhausted limit and a failed service. Watch the handovers: several individually permitted actions may combine into an unacceptable outcome.[9],[13]

Use existing risk management and layered assessment practices.[10],[11] Give every claimed boundary an owner, an enforcement point and evidence that it holds. Do not measure success by how rarely people disagree with the agent. A healthy team can challenge a convincing answer.

Measure	Practical indicator
Useful work	Accepted outputs without material rework / outputs reviewed.
Human effort	Minutes per accepted outcome, including checking and correction, against the baseline.
Handover quality	Required fields present, with sampled checks of source accuracy.
Challenge quality	Seeded defective recommendations rejected; track false rejections separately.
Control response	Time to stop all workers and pending actions; correct pause and escalation in drills.

LEARNING IS AN APPROVED CHANGE PROCESS

FIGURE 8. Feedback can improve the team without letting agents rewrite their own remit, controls or memory policy.

Protect the evidence independently of the actor. A stop must reach workers, credentials, queues and retries. Preserve evidence, contain harm and reconcile unfinished work before restarting. Some disclosures cannot be undone; recovery also means correction and consequence management.

Start with one team you can actually govern.

Choose a useful workflow. Prove the operating model before expanding it.

Days 1-30

Design the team

Choose one useful, low-risk workflow. Name the people and agent roles. Complete the charter and establish the value baseline.

Gate: the boundaries can be explained and enforced.

Days 31-60

Rehearse the work

Test the shared workflow, handovers, refusals and fallback. Train supervisors to challenge evidence and manage exceptions.

Gate: the team handles failure without widening its authority.

Days 61-90

Prove a limited service

Run an approved, observed pilot. Measure value and human effort. Practise stopping and restoring the whole operation.

Gate: continue, narrow, redesign or retire on evidence.

Ninety days is a planning horizon, not a readiness promise. Expansion requires a fresh decision. Ask suppliers to demonstrate refusal, revoked access and recovery using your workflow. A successful task demonstration is only the beginning of the conversation.

CEO

Which decisions and consequences will we delegate?

CIO

Can the design work across our applications and suppliers?

CISO

Will one compromised role remain contained and leave reliable evidence?

FIGURE 9. Joint decisions for the CEO, CIO and CISO. Progress on evidence; a decision to narrow or retire a use case is a valid result.

Evidence behind the argument.

Selected guidance, research and executive commentary. Reviewed 24 September 2026.

References distinguish external evidence from CiBRAI's framework and illustrative workflows. The diagrams express a proposed operating design, not a certification scheme or proof that every deployment is suitable.

01 **Fisher, B., & Galluzzo, R. (2026, 27 August).** Back to the Future: Why Agentic AI Needs a Strong Identity Foundation.

NIST. Agent identity, credential sharing and constrained authorisation.

02 **Australian Signals Directorate. (2026, 11 September).** Agentic AI harnesses.

The surrounding software as a security and governance control surface.

03 **ASD's ACSC and international partner agencies. (2026, 1 May).** Careful adoption of agentic AI services.

Joint deployment guidance, including the recommendation to use agents only for low-risk, non-sensitive tasks.

04 **Infocomm Media Development Authority. (2026).** Model AI Governance Framework for Agentic AI (Version 1.5).

Published 20 May; updated 5 June. Bounded risk, accountability, oversight and lifecycle controls.

05 **OWASP GenAI Security Project. (2025, 9 December).** OWASP Top 10 for Agentic Applications 2026.

Agent-specific threat categories. Practitioner guidance; not an incidence survey.

06 **Chismon, D. (2025, 8 December).** Prompt injection is not SQL injection (it may be worse).

UK National Cyber Security Centre. Limits of relying on an instruction/data boundary within a model.

07 **Anthropic. (2025, 20 June).** Agentic misalignment: How LLMs could be insider threats.

Controlled adversarial simulations. The reported behaviours are not estimates of real-world deployment prevalence.

Sources and publication notes.

08 **Joint Task Force. (2020, updated December).** Security and Privacy Controls for Information Systems and Organizations (NIST SP 800-53 Rev. 5).

Established controls, including AC-5 Separation of Duties and AC-6 Least Privilege.

09 **Bellogín, A., et al. (2025, 16 October).** Systemic Risks Associated with Agentic AI: A Policy Brief.

ACM Europe TPC. Systemic interactions and dynamic oversight; consulted through the authors' university repository.

10 **Leo, M., Tan, F., Miao, T., & Anand, G. (2026).** From threat to trust: assessing security risks of agentic AI systems.

International Journal of Information Security, 25, Article 23. Published abstract consulted; cited for its layered assessment approach.

11 **National Institute of Standards and Technology. (2023).** Artificial Intelligence Risk Management Framework (AI RMF 1.0).

The Govern, Map, Measure and Manage functions provide broader risk-management context.

12 **Klein, B., Lewis, C., Isenberg, R., et al. (2025, 16 October).** Deploying agentic AI with safety and security: A playbook for technology leaders.

McKinsey & Company. Executive framing of agents as digital insiders.

13 **Zhang, Z., Li, Q., Cao, J., Liu, L., & Ni, J. (2026).** From AI-Generated Content to Agentic Action: Security and Safety Threats in Generative AI.

arXiv:2605.16471. Preprint review, used as supporting research rather than validated deployment evidence.

14 **MIT Sloan School of Management. (2026, 18 February).** Agentic AI, explained.

Executive introduction to agents, organisational implications and governance.

About the author. Andrew Curtis is a CISO and the founder of CiBRAI, with more than 20 years of experience across enterprise and government cyber security, architecture, governance and uplift programs.

Applying the paper. The examples are reference workflows. Live use requires approved scope, risk assessment, tested controls and compliance with the organisation's obligations. The framework does not replace that decision. Current deployment guidance is stated on page 3.

Original framework and editorial content: Andrew Curtis / CiBRAI. Conceptual artwork was created with OpenAI image generation using two briefs: a mixed business team collaborating around a shared task; and a human-led security team reviewing evidence with agent support. The artwork is illustrative, not a product screenshot. Control diagrams are separately authored.

Build the team worth delegating to.

Start with a real workflow and the people responsible for it.

The competitive advantage is a team that can combine speed with judgement. Agents can reduce the effort spent on preparation and coordination. People supply context, relationships and responsibility. Good architecture gives both room to contribute while keeping authority clear.

Before adding another agent, draw the team it will join. Show the people, the handovers, the information boundaries and the decisions nobody should make alone. If you cannot point to who owns the outcome, the design is unfinished.

> Every agent needs a job, a boss, a boundary and a brake.

THE FRAMEWORK IN PRACTICE

CiBRAI implements the CiBRAI Agentic Business Framework as a core part of its cybersecurity operating platform and the way we design human-agent security operations.

BRING ONE WORKFLOW TO THE TABLE

Bring one real workflow to the table. Include the business owner, the people doing the work, technology and security. Use the team charter on page 7 to find the gaps together. If a second perspective would help, CiBRAI welcomes a practical conversation about applying the framework to your business or security operation.

cibrai.com | gadgetaccess.com

The next great organisation chart will include software.
The responsibility at the top will still be human.