

Which Work Can AI Own?



AI Peer Group, Session Five

A leadership test for putting agents into everyday operations.

Rob Williams

September 29, 2026

13:00 to 14:00 EDT, Zoom

Rob Williams

Technology leader
AI advisor

- 30 years in technology, with CPO, CTO and CAIO experience
- AI consulting across Insurance, Agriculture, VC, Government, Entertainment and SaaS
- Hands-on experience running a full-time multi-agent AI harness

Where are you using agents?

An agent can take actions through tools, rather than only answer a question.

How is your organisation using AI agents today?

A. In everyday operations

B. In pilots or experiments

C. Planning, but not yet using them

D. Not using them, or unsure

Choose one answer on your phone. Include agents used by individual teams.

GPT-6 adds more options for everyday work

Sol and Luna join Astra. The launch covers ChatGPT Work, Codex and the API, not yet Chat.

Astra

Full capability

OpenAI positions it for its most demanding work.

Sol

Everyday depth

A lower-cost option for complex work.

Luna

Speed and cost

A lower-cost option for routine work.

Leadership question: which model delivers an acceptable result at the lowest total workflow cost?

[OpenAI, Introducing GPT-6 Sol and Luna, 22 Sep 2026](#)

Opus 5.5 lowers the cost of complex work

Claude Opus 5.5, released September 22, 2026

- Standard API pricing per million tokens: \$4 input, \$20 output and \$0.20 for cache reads.
- Anthropic estimates 40% lower typical workload cost than Opus 5 at default settings.
- AutomationBench: 40.0% versus 26.9% for Opus 5, in Zapier tests cited by Anthropic. This measures business workflows.

The green light still needs evidence from your workflow, with an owner and clear limits on what the agent may change.

[Anthropic, Claude Opus 5.5, 22 Sep 2026; vendor pricing and cited benchmark results](#)

Sonnet 5.5 strengthens everyday agents

Claude Sonnet 5.5, released September 28, 2026

- Standard API pricing per million tokens: \$2 input, \$10 output and \$0.20 for cache reads, unchanged from Sonnet 5.
- Anthropic reports 70.6% on Terminal-Bench 4.0 versus 10.3% for Sonnet 5, a test of agentic coding.
- Anthropic reports up to 30% lower cost per task because it uses fewer tokens. That is a workload result, not a token-price cut.

Faster everyday agents make permission boundaries more useful: define what they may prepare, send and change.

[Anthropic, Claude Sonnet 5.5, 28 Sep 2026: vendor pricing and benchmark results](#)

Fable 5.1 separates capability from access

Claude Fable 5.1 and Claude Mythos 5.1, released September 1, 2026

- Fable 5.1 is generally available. Mythos 5.1 uses the same model with different safeguards and restricted trusted access.
- Fable API pricing per million tokens: \$10 input, \$50 output and \$0.25 for cache reads.
- Anthropic estimates 25% lower typical cost than Fable 5, or up to 45% for highly agentic work, driven by cheaper cache reads.

The same capability can carry different permissions. Choose the data, tools and actions you will authorise.

[Anthropic, Fable 5.1 and Mythos 5.1; newsroom dates the release to 1 Sep 2026](#) [Release date](#)

Managed agents are becoming a service

OpenAI Agents API, September 10, public beta

- OpenAI now offers the Codex harness through an API for developers.
- It manages long sessions, tools and subagents. Teams can choose hosted or other supported compute environments.
- The service is in public beta. Token and tool usage still incur charges.

Leadership implication: a provider can run the machinery, while your organisation still defines the job and checks the result.

[OpenAI, Introducing the Agents API, 10 Sep 2026](#)

Who builds and supplies your agents?

How are your agents built and provided?

- A. Built by our own team
- B. Built into software we buy
- C. Built or run for us by an external provider
- D. A mix of internal and external approaches
- E. Not using agents, or unsure

Choose the closest answer on your phone. Include agents still in pilots.

Copilot brings more work into one place

Microsoft announcement, September 25

- Home brings Chat and Cowork together, with Word, Excel and PowerPoint in the experience.
- Code supports building apps and workflows. Autopilot is designed to keep working between prompts.
- Home and Code are entering the Frontier rollout. Autopilot is expanding to private preview at month end.

Leadership question: which work belongs inside your existing Microsoft environment, and which capabilities can your team actually access?

[Microsoft, Introducing the new Copilot, 25 Sep 2026](#)

AI monitoring data can stay under customer control

Anthropic Enterprise Frontier Safeguards, September 1

- Anthropic announced controls that let customers keep monitoring data in their own cloud infrastructure.
- The design includes customer-managed keys and automated monitoring, with flags for the customer's own team to review.
- The announced rollout is phased for this fall. Availability needs confirmation for each organisation.

Leadership question: who holds the activity logs, who can inspect them, and who acts when a problem appears?

[Anthropic, Enterprise Frontier Safeguards, 1 Sep 2026](#)

NVIDIA puts agent controls outside the agent

Open Agent Safety Platform, announced September 28, 2026

- OpenShell software sets limits on files, networks and tools, then checks and enforces those limits while an agent works.
- Sentry adds a separate watchdog on BlueField hardware. NVIDIA says it can monitor activity and stop agents that cross their boundaries.
- OpenShell is available now. Sentry is part of the reference system design; confirm the hardware and deployment requirements.

The green light needs an independent control: who can see what the agent did, limit its authority and stop it?

[NVIDIA announcement, 28 Sep 2026](#) [NVIDIA technical explanation; vendor descriptions](#)

The green light is a leadership decision

One workflow, one owner, clear permission

Model progress tells us what AI can do. Leaders decide what it may do, what evidence counts and who can stop it.

Agent governance is already falling behind

EMA for Cequence Security: 202 IT and security leaders at organisations with 1,000 or more employees.

65%

Outside scope

Reported an agent acting beyond its intended scope, including near misses.

46%

Audit gap

Could not easily produce a complete trail of an agent's previous 30 days.

47%

Inventory gap

Could not reliably inventory every deployed agent, according to EMA.

Self-reported survey findings, not a measured failure rate for all agents. The inventory gap includes manual and partial inventories.

[EMA, Agents Without Guardrails, Aug 2026, pp. 4, 9, 16 and 28](#) [Cequence announcement, 31 Aug 2026](#)

Test agents reached real systems

OpenAI and Hugging Face, July 2026; detailed OpenAI report, August 26

- During internal security evaluations, OpenAI agents bypassed isolation controls and reached systems outside their assigned tasks.
- Hugging Face confirmed access to internal datasets and service credentials in its production infrastructure.
- This began in a test environment with reduced safeguards. The intrusion affected real systems, not a simulated victim.

Leadership lesson: test environments need enforced limits on connections and access, plus a way to stop the agent.

[OpenAI incident report, 26 Aug 2026](#) [Hugging Face disclosure, 16 Jul 2026](#)

A routine fix became a production deletion

PocketOS on Railway, April 2026; provider account published April 29

- Railway confirmed an agent used a locally stored access token to delete a customer's live database while trying to fix something else.
- The token had broad account access. The API deleted immediately, bypassing the delay available in the human dashboard.
- Railway recovered all the data and added a 48-hour deletion delay to the API. This was not permanent loss of every backup.

Leadership lesson: a safety check must cover the tool the agent actually uses. A warning in the human interface is not enough.

[Railway, provider account and recovery, 29 Apr 2026](#) [PocketOS case and recovery, Tom's Hardware, 30 Apr 2026](#)

The green light needs four checks

These are business control failures, whatever model sits underneath.

Beyond scope

A request to investigate becomes permission to change or delete.

Unverifiable done

The agent claims success, but nobody checks the actual result.

Invisible actions

Work happens without a record that the owner can inspect.

Unknown agents

No reliable inventory or audit trail shows who acted, with what access.

Before granting authority: know the agent, limit its actions, see what happened and check the result.

What has gone wrong with your agents?

What is the most significant agent issue you have experienced?

- A. Incorrect output or a false claim of completion
- B. Actions beyond agreed permissions
- C. Data exposure or a security incident
- D. Unexpected costs or operational disruption
- E. Missing logs or unclear actions
- F. No issues that we know of
- G. Not using agents, or unsure

Choose one answer on your phone. Share other examples in the live questions.

Hearth makes the work visible

Home is Rob's AI working environment. Hearth coordinates its specialist workspaces, called rooms.

Rooms

Clear responsibility

A request has an assigned workspace and a defined job.

Board

Shared task record

Each ask belongs on a visible dashboard, with an owner, state and evidence.

Follow-up

Work stays visible

Unfinished tasks are followed up. Blockers and decisions can be seen.

Operating rule: every work request belongs on the board. A completion needs an artifact, and someone must still verify it.

Source: Home architecture and Hearth tool reference, checked 29 Sep 2026. Illustrative architecture, no live data.

Latch separates a rule from an enforced gate

The irreversible set covers consequential acts such as sending, paying, deleting and publishing.

- Define the boundary: record which exact action needs human authority and show whether its gate is actually wired.
- On the enforced email path, a send needs recorded approval. Without it, the exact message is held for a visible decision.
- Current limit: payment, bulk deletion and production deployment are defined in policy, but their Latch gates are not wired.

Nothing reads as enforced unless it is. Written instructions and dashboard labels do not create a technical barrier.

Source: Home Latch README and current policy, checked 29 Sep 2026. Enforcement lives in the tool that performs the act.

Alignment must show up in the actions

Alignment means pursuing the intended goal within the agreed boundaries. Governance makes that testable.

- Agree the goal and limits. Finishing a task does not grant permission to invent new actions.
- Make the chain visible: request, owner, decision, action and result. Keep the evidence with the work.
- Verify every done against the real outcome. A completed card or a second model's agreement is not proof.

Home's Latch principle: "An act you cannot see, a promise you cannot list, and a rule you cannot verify are the same failure."

Source: Home Latch principle and architecture. Practical operating guidance, not a claim of complete enforcement.

How confident are you in the controls?

How confident are you in the security and governance of your agents?

- A. High: we have tested controls and evidence
- B. Some: controls exist, but we have gaps
- C. Low: we rely mainly on instructions and trust
- D. Unsure: we cannot assess the controls
- E. Not using agents yet

Choose one answer on your phone. Think about permissions, data access, logs and a way to stop an agent.

Can you account for your agents?

Could you list every agent acting for your company today?

- A. Yes, with an owner and permissions for each
- B. Some, but I could not prove the list is complete
- C. No, or I do not know

Include pilots and agents built by individual teams. What would make your answer verifiable?

Give the green light only where you can prove control

Choose one agent or workflow. Name the person who will check these four things.

Inventory your agents

List the owner, purpose, connected systems and permissions.

Make every act visible

Keep an inspectable record of the request, approval and action.

Gate the irreversible

Require explicit authority before sending, paying, deleting or publishing.

Verify every done

Check the actual result and attach the evidence before closing the task.

Discussion: which of these four checks is missing in your organisation, and who will own it?