aicadium

CHECKLIST FOR LEADERSHIP TEAMS

The agent control checklist

Ten questions your leadership team should be able to answer before AI agents touch production

Why this checklist

Most organisations govern artificial intelligence (AI) with a written policy. An AI agent does not read a policy. It works within whatever limits the systems around it enforce. These ten questions test whether those limits exist in your organisation today. Each one comes from a problem we met while putting agents to work on our own engineering and operations. You do not need a technical background to use it. Put each question to the person who runs the relevant system, and accept only an answer you could verify.

This checklist accompanies our article, **Your AI agents do not read your AI policy**. Read it at aicadium.ai/your-agents-do-not-read-your-ai-policy

Key terms

Plain-language definitions for the terms the questions use

AI agent

Software that carries out a task on its own, using the same systems an employee would. It might write code, draft a document or open a change request.

Route to production

The steps a piece of work passes through before it takes effect for customers or employees.

Agent identity

A separate, verifiable identity for each agent, so that its actions can be told apart from an employee's.

Sandbox

A sealed, temporary workspace created for one task and removed when the task ends.

Human approval checkpoint

A defined point where a person must approve before the work continues, however confident the agent is.

Audit trail

A lasting record of who started an agent, what it was asked to do and what it did.

Guardrail

A control that limits what an agent can do, enforced by a system and not by a document.

Credential

A key, token or password that grants access to a system.

Least privilege

Giving an agent access only to the systems and data its task needs.

Review gate

A checkpoint where a proposed change must pass review before it can take effect.

Behaviour test

A realistic scenario run against an agent to check how it acts. This is sometimes called an evaluation.

The ten questions

Answer each question yes or no. Score one point for each yes. A yes counts only when a system enforces the control. A written rule that people are expected to follow counts as a no.

Visibility and ownership

See it in practice

1 **Can you list every AI agent running in your organisation today, and name the person answerable for each?** ☐ Yes ☐ No

In an Okta survey of 306 security executives in 2026, fewer than half were confident they could identify all the AI agents in their environment.

2 **Can a person stop any agent within minutes, and does it stop without help when its time limit passes?** ☐ Yes ☐ No

The Kiteworks 2026 Forecast reports that 60% of organisations cannot quickly terminate a misbehaving agent.

Access and credentials

See it in practice

3 **Does each agent work under its own identity, separate from any employee's login?** ☐ Yes ☐ No

4 **Are credentials kept out of the agent's reach, so that it is unable to read, copy or pass them on?** ☐ Yes ☐ No

In a Gravitee survey of 900 technology leaders in October 2026, 45% relied mainly on shared accounts, borrowed human credentials or fixed keys.

5 **Is each agent limited to the systems and data its task names, and nothing more?** ☐ Yes ☐ No

Review and approval

See it in practice

6 **Is every change an agent proposes reviewed before it takes effect?** ☐ Yes ☐ No

7 **Have you defined the changes that always need a person's approval, and does a system enforce that rule?** ☐ Yes ☐ No

In an EY survey of 202 senior AI executives at large United States companies, 47% admitted they had bypassed their AI governance policy for urgent deployments.

8 **Is it impossible for any agent to publish to all employees, to customers or to production without a person approving?** ☐ Yes ☐ No

Testing

See it in practice

9 **Do you test how your agents behave in realistic scenarios, and not only whether their files are well formed?** ☐ Yes ☐ No

VentureBeat surveyed 157 enterprise respondents in June 2026. Half had shipped an agent or AI feature that passed internal evaluations and then failed a customer, in a small self-selected sample.

Evidence

10 **After an agent has finished, can you reconstruct who started it, what it was asked to do and what it did?** ☐ Yes ☐ No

Your score Count one point for each yes _____ out of 10

What your score tells you

9 to 10

Your controls are enforced by systems.
Review them whenever you add a new kind of agent.

6 to 8

Most controls are in place.
Take each question you answered no in the access and approval groups first.

0 to 5

Your controls rely mainly on written policy.
Start with questions 1, 4 and 7, which give the largest reduction in risk for the least change.

How these questions map to Singapore's framework

Singapore's Infocomm Media Development Authority published a voluntary governance framework for agentic AI in January 2026. Its recommendations line up with the question groups.

QUESTIONS 1 AND 2	QUESTIONS 3 TO 5	QUESTIONS 6 TO 8	QUESTION 9
Make humans meaningfully accountable	Limit an agent's tools and permissions	Require human approval at defined checkpoints	Test before deployment

About this checklist

Aicadium runs these controls on its own work. The checklist reflects what we built and what we are still deciding. Read how each control works in practice in the companion article.

Your AI agents do not read your AI policy
aicadium.ai/your-agents-do-not-read-your-ai-policy

Sources

Okta, Global CISO Insights 2026, okta.com

Kiteworks, 2026 Forecast, kiteworks.com

Gravitee, State of AI agent security, October 2026, gravitee.io

EY survey of senior AI executives, as reported by CDO Magazine, September 2026

VentureBeat, VB Pulse survey, June 2026, venturebeat.com

Infocomm Media Development Authority, Model AI Governance Framework for Agentic AI, January 2026