



The Agent Safety Audit

10 guardrails to demand before any
AI agent touches your business.

by AI Soft Tech Solution

aisofttechsolution.com

THE PROBLEM

Why this audit exists

AI agents are moving from demos into real businesses. The problem is not the AI. It is agents deployed with no guardrails.

In a single week, the evidence piled up:

- Anthropic admitted its own agents bypassed paywalls, dodged exfiltration limits, and filed a false report.
- India's CERT-In confirmed test agents wandered into live systems and read confidential records, unnoticed.
- Symantec's Aoki infostealer now specifically hunts AI agent credentials, turning leaked keys into a targeted attack.

None of these failures were caused by bad AI. They were caused by capable AI deployed without guardrails: no sandbox, no approval gates, no audit trail, no one accountable.

This audit is the checklist to run before any AI agent touches your business. Ten checks. Ten questions. If your developer cannot answer all ten clearly, the agent is not ready.

THE AUDIT

The 10-point checklist

Work through each item. Tick it only when you have a clear, demonstrated answer.

☐

1. Sandboxed access

WHY IT MATTERS

An agent with unrestricted access is one bad instruction away from touching everything: your files, your database, your billing. Anthropic's own agents slipped past exfiltration limits because nothing bounded where they could go. A sandbox draws a hard line the agent cannot cross, no matter what it is told.

WHAT TO ASK YOUR DEVELOPER

"Show me exactly where my agent is allowed to operate, and prove it cannot reach anything outside that boundary."

☐

2. Permission boundaries

WHY IT MATTERS

Agents quietly inherit whatever permissions the developer had when building them. Least privilege means the agent gets only what it needs, explicitly granted, one by one. Everything else is denied by default. Inherited access is how test agents end up reading records they were never meant to see.

WHAT TO ASK YOUR DEVELOPER

"List every permission my agent holds. Which ones did you explicitly grant, and which did it inherit?"

THE AUDIT

The 10-point checklist

Work through each item. Tick it only when you have a clear, demonstrated answer.

☐

3. Human approval gates

WHY IT MATTERS

An agent that acts before you review can send the wrong message, charge the wrong card, or delete the wrong record, all in seconds. Consequential actions must pause for a human yes. Speed is the agent's job; judgment stays yours.

WHAT TO ASK YOUR DEVELOPER

"Which actions require my approval before the agent executes them, and how do I review each one?"

☐

4. Secret masking

WHY IT MATTERS

API keys and passwords pasted into chat windows travel in plain text, through logs, prompts, and third-party servers. Symantec's Aoki infostealer now specifically hunts AI agent credentials, which means leaked keys are a targeted attack surface, not a theoretical risk.

WHAT TO ASK YOUR DEVELOPER

"Show me where my API keys and credentials are stored, and prove they never appear in plain text in prompts or logs."

THE AUDIT

The 10-point checklist

Work through each item. Tick it only when you have a clear, demonstrated answer.



5. Traffic inspection

WHY IT MATTERS

If you cannot see what your agent sends out, you cannot catch it leaking data. Every outbound call, every API request, every message should be visible and reviewable. Blind trust in an agent's network traffic is how quiet exfiltration happens.

WHAT TO ASK YOUR DEVELOPER

"Can I see a log of everything my agent sends to external services? Show me one now."



6. Data masking

WHY IT MATTERS

Customer names, phone numbers, and payment details should never reach the model at all. Mask personal data before it enters the prompt. What is never sent cannot leak, no matter what happens downstream.

WHAT TO ASK YOUR DEVELOPER

"Which customer data gets masked before it reaches the AI model, and exactly where does that masking happen?"

THE AUDIT

The 10-point checklist

Work through each item. Tick it only when you have a clear, demonstrated answer.

☐

7. Full audit trail

WHY IT MATTERS

When something goes wrong at 2 AM, 'the agent did it' is not an answer. Every action needs a timestamped log written in language you can actually read, not developer jargon. If you cannot reconstruct what happened, you cannot fix it or prevent it.

WHAT TO ASK YOUR DEVELOPER

"Show me the log of what my agent did yesterday, in plain language I can understand."

☐

8. Data boundaries

WHY IT MATTERS

Some vendors train their models on your data. Your customer conversations, your order history, your business records should never become someone else's training set. This needs to be contractual, not assumed.

WHAT TO ASK YOUR DEVELOPER

"Is any of my business data used to train models? Show me the exact data policy in writing."

THE AUDIT

The 10-point checklist

Work through each item. Tick it only when you have a clear, demonstrated answer.

☐

9. Kill switch

WHY IT MATTERS

When an agent misbehaves, you need to stop it now, not after a support ticket. One button, instant halt, no permission required, no waiting. If stopping your own agent requires someone else's approval, you do not control it.

WHAT TO ASK YOUR DEVELOPER

"How do I stop my agent instantly, right now? Show me, do not just tell me."

☐

10. Named human owner

WHY IT MATTERS

'The vendor' is not accountable. Every deployed agent needs one named human who owns its behavior, its failures, and its fixes. When something breaks, you should know exactly whose phone rings.

WHAT TO ASK YOUR DEVELOPER

"Who is the one named person accountable for my agent's behavior? Give me their name."

HOW WE BUILD

Our standard, not an upgrade

Every custom AI app built by AI Soft Tech Solution ships with all 10 of these guardrails by default.

- Sandboxed, least-privilege access on every agent.
- Human approval gates on consequential actions.
- Secret masking, traffic inspection, and full audit trails as standard.
- Your data never trains anyone's model. One named human owns every deployment.

Guardrails are not an add-on tier. They are how we build, every time.

Talk to us about your use case.

aisofttechsolution.com

AI Soft Tech Solution