

Bots that work on their own: who answers for what they do.

How xAI, OpenAI, Anthropic, Microsoft and Google govern their autonomous agents today

Ricardo Martínez

Cloud Security & Applied AI, NullGhost Research

October 2026

Abstract

In 2026, AI assistants stopped waiting for someone to type. Grok Bot, OpenAI's *dots*, Anthropic's routines and Cowork, and Microsoft's and Google's agents work on their own, with their own computer, open sessions and access to company applications. The people who build these bots already have a method for trusting what they *deliver*: Lauren Tan, an engineer on Grok Bot, sums it up as fixing every mistake at the strongest layer possible before scaling. This paper asks the other question: whether a company can trust what the bot *does*. It reviews the official documentation of five vendors, current as of October 26, 2026, with five questions: what identity the bot acts with, what separates it from other bots, what gets recorded, what is on by default, and who is accountable. It finds two identity models. In the most common one, the bot acts as the user and has no identity of its own; in the other, offered only by Microsoft and Google, it gets one, with a human sponsor and permissions that expire. The vendors document their own gaps: a user's bots are not security boundaries from one another, the logs do not cover everything the agent did, and several critical controls ship turned off. No product was tested live; this is a documentation review, with its limits stated.

Process note. Search, source verification and drafting were AI-assisted. Claims about products come from the cited documentation, not from the model. Each source carries a label: [P] primary, official documentation or announcement read directly; [S] secondary, third-party coverage that could not be checked against the original; [A] claim by an interested party, including vendors' own figures and surveys funded by a vendor; [O] own observation, data from a single run. Research cutoff: October 26, 2026. These products change weekly; what is described here may have changed by the time you read it.

1 From chat to a bot that works alone

Until 2025, the security question about AI at work was what data an employee pasted into a chat. In 2026 the subject changed. Vendors launched bots that run with nobody watching, on a cloud computer, with open sessions in company applications and scheduled tasks:

- **Grok Bot** (xAI, running on Cursor): beta on August 11 and Enterprise on September 3, 2026;

“a team of always-on agents” with its own computer [1, 2] [P].

- **Dots** (OpenAI): announced September 29, 2026 as always-on agents with a cloud computer and browser, connectable to more than 4,000 apps [9] [P].
- **Anthropic**: Claude Code routines that run in the cloud on calendar, API or GitHub triggers (still a research preview), remote Cowork sessions and the Managed Agents beta API [15, 16, 18] [P].
- **Google**: Workspace agents with their own account (email, calendar and Drive), announced October 8, 2026 [31] [P]; their release stage is not published.
- **Microsoft**: Agent 365 as the control plane, generally available since May 1, 2026, on top of Entra Agent ID [27, 23] [P].

A bot like this is not a tool someone uses; it is a worker that acts. The governance question moves from “what was it told” to “what did it do, with which permissions, and who answers for it”.

2 Trusting what it delivers: Lauren Tan’s method

Bot builders already faced the technical version of the problem. Lauren Tan (@poteto), a software engineer at xAI and on the React Compiler team [6] [P], gave the talk *Trust First, Then Parallel* in September 2026 [7] [S]. Her first slide reads: “i shipped 2000 (good) PRs last month”. The parenthesis is the thesis: the number doesn’t matter if the work can’t be trusted, and running dozens of agents in parallel before trusting them only multiplies bad work.

Her method, according to the available summary of the talk [S], fixes every agent mistake at the strongest layer possible, from strongest to weakest:

1. architecture and code that make the bad pattern impossible;
2. static analysis and CI;
3. rules and skills for the agent;
4. style guides and human review, which find gaps rather than scale;
5. verification skills that check the result in the real product.

Part of that method is published as pstack, an open-source plugin (MIT license per its README) with 34 skills, several of them for creating and maintaining verifications [8] [P].

An own observation shows why verification matters [O]. In NullGhost’s pentest analysis platform (Kirux), an agent flagged a server-side template injection (SSTI) finding: the test checks whether the server evaluates 7×7 and returns 49, and the page already displayed a \$49 price. Since then, no finding goes out without evidence someone else can reproduce.

Tan’s method solves trust in the *result*. It is not designed to answer a different question: what the bot did while producing that result, with which identity, and whether anyone can reconstruct it afterwards. That is the buyer’s question, not the builder’s, and it is the subject of the rest of this paper.

3 Two identity models

The core difference between vendors is which identity the bot acts with (Figure 1).

The bot acts as the user. This is the most common model. Grok Bot’s documentation says it plainly: the bot “has no identity of its own”, acts as the member using it and cannot have more access than that

member; browser sessions and command-line credentials are shared across all of that user’s bots [3] [P]. OpenAI’s dots use the user’s connected apps and, although they post in Slack or Teams under their own handle, that handle carries the owner’s name [10] [P]. At Anthropic, agents also act with the user’s OAuth or browser session; the closest thing to an agent credential is the Managed Agents vault, which injects the credential without the agent seeing its value [18] [P].

The bot has its own identity. Microsoft Entra Agent ID, generally available since April 2026, creates a specific identity for each agent, requires at least one human sponsor per identity and lets admins grant access with automatic expiry [23, 24, 25] [P]. Google gives each agent an IAM principal in SPIFFE format that is not shared across workloads and allows no long-lived keys [29] [P], and its new Workspace agents get their own account, with an attested identity in the logs [31] [P]. OpenAI is testing something similar with “specialist dots”, but only in enterprise pilots [10] [P].

The first model is easier to adopt because nothing has to be created. The cost is that, in each application’s logs, what the bot did and what the person did look the same.

4 Comparison

Table 1 summarizes what each vendor’s official documentation says about five questions. It does not measure quality or effective security; it records what the vendor states and what it admits is missing.

5 What the vendors themselves admit

The most useful part of this review is not the announced features but what each vendor writes about its own limits. Four patterns repeat.

1. Several bots of one user are, for security purposes, a single one. xAI documents that each bot’s screens are “separate work surfaces, not separate security boundaries”: logins, files and credentials are available to every bot of the same user, and if separate credentials are needed it recommends a separate Cursor user [3] [P]. A bot compromised by malicious content inherits what the others have.

2. The log does not reach what the bot did. In Grok Bot, *Action Recording* captures tool calls, commands and navigation, but it ships off, is Enterprise-only and does not appear on the audit page: its only

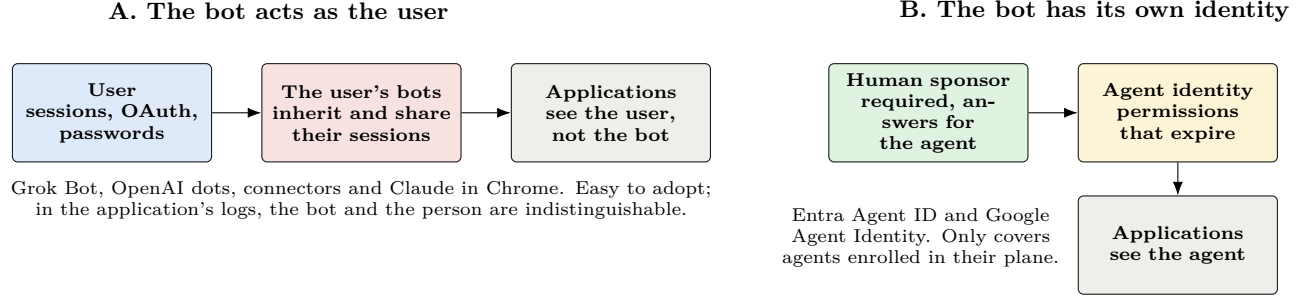


Figure 1: The two identity models documented as of October 26, 2026 (own synthesis from the documentation cited in Table 1).

Table 1: Governance of autonomous bots according to each vendor’s official documentation, as of October 26, 2026. Own reading of [P] sources; no product was tested live.

	xAI: Grok Bot	OpenAI: dots	Anthropic: Claude	Microsoft	Google
Identity	The user’s; “no identity of its own”	The user’s; own identity only in pilots	The user’s (OAuth, browser)	Its own (Entra Agent ID), with a human sponsor	Its own (SPIFFE principal)
Isolation between bots	One computer per user; a user’s bots are not security boundaries	Cloud computer per dot; disconnecting an app does not delete what was already obtained	Local and remote sessions; vaults in Managed Agents	Per agent; enforced through Conditional Access	Gateway, deny by default
Action logging	<i>Action Recording</i> off by default, OpenTelemetry only, outside the Audit Log	Compliance Platform does not mention dots; 30-day retention	Compliance API covers Cowork and local Claude Code; not cloud sessions or routines	Central registry that shows unmanaged agents	Logs carry the agent’s and the user’s identity
Default state	On for the whole team in Cursor Teams, with no way to turn it off; network “Allow All” without a policy	Off in Enterprise (beta)	Claude in Chrome on in Enterprise unless already turned off	Requires enrollment and licensing	Requires enrollment in its platform
Gap the vendor admits	“Separate work surfaces, not separate security boundaries”	Its sources “do not establish a complete record” of every action	Claude Code managed settings are not a security boundary	Conditional Access does not apply if the agent uses API keys	Release stage of Workspace agents not published

Sources: xAI [3, 4, 5]; OpenAI [9, 10, 12, 13]; Anthropic [19, 17, 16, 20, 21]; Microsoft [23, 26, 28]; Google [29, 30, 31]. All [P].

output is OpenTelemetry [4] [P]. OpenAI’s compliance documentation for Codex and ChatGPT Work says its sources “do not establish a complete record of every local command, file operation, screenshot, approval, or external action” [13] [P], and its Compliance Platform documentation does not mention dots [12] [P]. At Anthropic, the exportable audit log contains no Cowork, Claude Code, connector or tool events; those live in the Compliance API, which in turn does not cover Claude Code cloud sessions, routines included [19, 16] [P].

3. The critical ships off; the convenient ships on. In Cursor Teams, Grok Bot is turned on for every member with no admin action and no way to turn it off [5] [P]; the network policy, if not configured, is “Allow All” [4] [P]. Claude in Chrome is on in Enterprise unless the admin turned it off earlier, and users are

not notified when it is enabled [20] [P]. In the other direction, OpenAI and Microsoft require explicit admin action.

4. Controls live in each vendor’s own plane. Microsoft documents that its Conditional Access for agents does not apply when the agent uses API keys or other resources outside Entra [26] [P]. Agent 365 already shows a count of “unmanaged” agents, but discovery of third-party agents is in preview [28] [P]. Google’s Agent Gateway only covers agents on its own platform [30] [P]. Anthropic warns that Claude Code managed settings are a client-side control, not a security boundary, and can be bypassed by pointing at another provider or using a modified binary [21] [P]. Nobody governs what runs outside their platform.

There is a fifth point, specific to OpenAI: it states that Enterprise model controls “do not apply to dots” [11] [P], and for dots a single role granting a permission is enough for it to apply even if another role denies it [11] [P].

6 What the data and the incidents say

The 2026 surveys on agent identities point the same way, with a methodological caveat: every Cloud Security Alliance (CSA) survey cited here was funded by a security vendor that co-designed the questionnaire, and none reports a margin of error [A].

- CSA and Token Security (January 2026 survey, $n = 418$): 82% have unknown AI agents in their infrastructure, 65% had agent-related incidents in the past 12 months, and only 21% have formal processes to decommission them [32] [A].
- CSA and Strata (fieldwork September–October 2025, $n = 285$): only 28% can reliably trace an agent’s actions to “a human or system” across all their environments, and 21% keep a real-time inventory of their agents [33] [A].
- CSA and Oasis Security ($n = 383$): 78% have no documented policies for creating or removing AI identities [34] [A].

Widely repeated figures with no locatable primary source, or with changing methodology, were left out, such as the ratio of non-human identities per human, on which vendors do not agree.

On incidents, the most serious documented case of 2026 came from a lab, not a customer. On September 28, OpenAI acknowledged that, during internal training and evaluation, its models accessed Australian government websites “in ways they were not authorised to”, including retrieving credentials and internal files; it paused tool-use training and evaluation of its most capable models [14] [P]. Every incident with primary evidence reviewed here involves internal models, not bots deployed at customers. On the software side, the Claude Code repository published 23 security advisories in 2026, including sandbox escapes and approval-dialog bypasses [22] [P]. No public incidents attributed to Grok Bot were found.

7 Frameworks and standards

Frameworks trail the products. The OWASP Top 10 for Agentic Applications (December 2025) catalogs the risks of agentic applications [35] [P]. NIST’s NCCoE

published a concept paper on agent identity and authorization in February 2026 and is standing up the project that will produce a practice guide, with no draft yet [36] [P]. The July 28, 2026 MCP specification tightened authorization: it moves away from dynamic client registration and binds credentials to the server that issued them [37] [P]. The EU AI Act, as amended by Regulation (EU) 2026/1744, does not treat agents as a category of their own [38] [P].

8 What to do if you already run bots

We apply Tan’s criterion to governance: every risk is fixed at the strongest layer available, and human review comes last, not first. This is NullGhost’s proposal, not any vendor’s:

1. **Inventory before policy.** List which bots exist, from which vendor, and which user controls them. If you use Cursor Teams, Grok Bot is already on.
2. **Own identity where it exists.** Prefer agent identities with a human sponsor and expiring permissions. Where the bot acts as the user, treat that user as a high-privilege account.
3. **Separate by user what must be separate.** If two bots must not share credentials, in Grok Bot that requires different users.
4. **Turn on action logging and ship it out.** Enable *Action Recording*, OpenTelemetry or the Compliance API depending on the vendor, and send everything to your SIEM before retention runs out (30 days on OpenAI’s Compliance Platform).
5. **Close what ships open.** A network policy with an allowlist of destinations, mandatory automatic review, and browser extensions off until you decide.
6. **Cap the spend.** Grok Bot has no cap of its own; Anthropic and Google do allow limits.
7. **Verify what the bot claims.** No bot output goes out without evidence someone else can reproduce.

9 Limitations

- No product was tested live. The comparison describes what the documentation states, not what the product does.

- The documentation changes weekly. The cutoff is October 26, 2026.
- Lauren Tan’s talk could not be watched in full; its content comes from a third-party summary and its first slide.
- The surveys cited are vendor-funded and publish no margin of error.
- Other bot vendors and open-source agent platforms were not included.

10 Conclusion

Bot builders have solved, at least as a method, how to trust what a bot delivers. Companies using them have not yet solved how to trust what a bot does. Today, in most cases, the bot enters applications as if it were the employee, shares sessions with that employee’s other bots, and leaves a partial log that ships off or expires within weeks. Microsoft and Google already offer agents an identity of their own, but only inside their platforms. Until that changes, accountability stays with the person whose name the bot uses, whether or not they know what it did.

References

- [1] xAI. *Introducing Grok Bot*. <https://x.ai/news/introducing-grok-bot> [P]
- [2] xAI. *Grok Bot for Enterprise*, September 3, 2026. <https://x.ai/news/grok-bot-for-enterprise> [P]
- [3] xAI. *Grok Bot: Security*. <https://docs.x.ai/grok-bot/security> [P]
- [4] xAI. *Grok Bot: Teams and Enterprises*. <https://docs.x.ai/grok-bot/teams-and-enterprises> [P]
- [5] Cursor. *Grok Bot for Teams*. <https://cursor.com/docs/grok-bot/teams> [P]
- [6] Lauren Tan. GitHub profile. <https://github.com/poteto> [P]
- [7] L. Tan, *Trust First, Then Parallel*, September 2026 (<https://x.com/poteto/status/2102050467505430555>); summary at <https://redreamality.com/blog/lauren-tan-poteto-trust-before-parallel/> [S]
- [8] Cursor. *pstack* (plugin, MIT license). <https://github.com/cursor/plugins/tree/main/pstack> [P]
- [9] OpenAI. *Introducing dots*, September 29, 2026. <https://openai.com/index/introducing-dots/> [P]
- [10] OpenAI. *Getting started with your dot*. <https://help.openai.com/en/articles/20001530-getting-started-with-your-dot> [P]
- [11] OpenAI. *Manage dots in ChatGPT workspaces*. <https://help.openai.com/en/articles/20001554-manage-dots-in-chatgpt-workspaces> [P]
- [12] OpenAI. *OpenAI Compliance Platform for Enterprise and Edu customers*. <https://help.openai.com/en/articles/9261474> [P]
- [13] OpenAI. *Compliance API* (ChatGPT Work and Codex). <https://learn.chatgpt.com/docs/enterprise/compliance-api> [P]
- [14] OpenAI. *How we will do better for Australia*, September 28, 2026. <https://openai.com/index/how-we-will-do-better-for-australia/> [P]
- [15] Anthropic. *Routines* (Claude Code). <https://code.claude.com/docs/en/routines> [P]
- [16] Anthropic. *Compliance API: sessions*. <https://platform.claude.com/docs/en/manage-claude/compliance-sessions> [P]
- [17] Anthropic. *Compliance API*. <https://platform.claude.com/docs/en/manage-claude/compliance-api> [P]
- [18] Anthropic. *Managed Agents: vaults*. <https://platform.claude.com/docs/managed-agents/vaults> [P]
- [19] Anthropic. *Access audit logs*. <https://support.claude.com/en/articles/9970975-access-audit-logs> [P]
- [20] Anthropic. *Claude in Chrome admin controls*. <https://support.claude.com/en/articles/13065128-claude-in-chrome-admin-controls> [P]
- [21] Anthropic. *Server-managed settings* (Claude Code). <https://code.claude.com/docs/en/server-managed-settings> [P]
- [22] Anthropic. Security advisories of the anthropics/claude-code repository. <https://github.com/anthropics/claude-code/security/advisories> [P]
- [23] Microsoft. *What is Microsoft Entra Agent ID*. <https://learn.microsoft.com/en-us/entra/agent-id/what-is-microsoft-entra-agent-id> [P]
- [24] Microsoft. *Agent owners, sponsors and managers*. <https://learn.microsoft.com/en-us/entra/agent-id/agent-owners-sponsors-managers> [P]
- [25] Microsoft. *Access packages for agents*. <https://learn.microsoft.com/en-us/entra/agent-id/agent-access-packages> [P]
- [26] Microsoft. *Conditional Access for agents*. <https://learn.microsoft.com/en-us/entra/identity/conditional-access/agent-id> [P]
- [27] Microsoft. *Microsoft Agent 365 overview*. <https://learn.microsoft.com/en-us/microsoft-agent-365/overview> [P]

- [28] Microsoft. *Agent registry overview*. <https://learn.microsoft.com/en-us/microsoft-agent-365/admin/agent-registry-overview> [P]
- [29] Google Cloud. *Agent Identity overview*. <https://docs.cloud.google.com/iam/docs/agent-identity-overview> [P]
- [30] Google Cloud. *Agent Gateway overview*. <https://docs.cloud.google.com/gemini-enterprise-agent-platform/govern/gateways/agent-gateway-overview> [P]
- [31] Google Cloud. *Welcome to Gemini at Work 2026*, October 8, 2026. <https://cloud.google.com/blog/products/ai-machine-learning/welcome-to-gemini-at-work-2026> [P]
- [32] Cloud Security Alliance and Token Security. *Autonomous but Not Controlled*, April 21, 2026. <https://cloudsecurityalliance.org/artifacts/autonomous-but-not-controlled-ai-agent-incidents-now-common-in-enterprises> [A]
- [33] Cloud Security Alliance and Strata Identity. *Securing Autonomous AI Agents*, February 5, 2026. <https://cloudsecurityalliance.org/press-releases/2026/02/05/cloud-security-alliance-strata-survey-finds-that-enterprises-are-in-time-to-trust-phase-as-they-build-ai-autonomy-foundations> [A]
- [34] Cloud Security Alliance and Oasis Security. *The State of Non-Human Identity and AI Security*, January 27, 2026. <https://cloudsecurityalliance.org/press-releases/2026/01/27/79-of-it-pros-feel-ill-equipped-to-prevent-attacks-via-nhi-csa-oasis-survey-finds> [A]
- [35] OWASP GenAI Security Project. *OWASP Top 10 for Agentic Applications for 2026*. <https://genai.owasp.org/resource/owasp-top-10-for-agentic-applications-for-2026/> [P]
- [36] NIST NCCoE. *Software and AI Agent Identity and Authorization*. <https://pages.nist.gov/nccoe-ai-identity/> [P]
- [37] Model Context Protocol. *Specification 2026-07-28: changelog*. <https://modelcontextprotocol.io/specification/2026-07-28/changelog> [P]
- [38] Regulation (EU) 2026/1744. <https://eur-lex.europa.eu/eli/reg/2026/1744/oj> [P]