

Bots que trabajan solos: quién responde por lo que hacen.

Cómo gestionan hoy xAI, OpenAI, Anthropic, Microsoft y Google a sus agentes autónomos

Ricardo Martínez

Cloud Security & IA Aplicada, NullGhost Research

Octubre 2026

Resumen

En 2026 los asistentes de IA dejaron de esperar a que alguien les escriba. Grok Bot, los *dots* de OpenAI, las rutinas y Cowork de Anthropic y los agentes de Microsoft y Google trabajan solos, con computadora propia, sesiones abiertas y acceso a las aplicaciones de la empresa. Quienes construyen estos bots ya tienen un método para confiar en lo que *entregan*: Lauren Tan, ingeniera de Grok Bot, lo resume en corregir cada error en la capa más fuerte posible antes de escalar. Este texto pregunta lo otro: si una empresa puede confiar en lo que el bot *hace*. Revisa la documentación oficial de cinco proveedores, vigente al 26 de octubre de 2026, con cinco preguntas: con qué identidad actúa el bot, qué lo separa de los demás bots, qué queda registrado, qué viene encendido por defecto y quién responde. Encuentra dos modelos de identidad. En el más extendido el bot actúa como el usuario y no tiene identidad propia; en el otro, solo Microsoft y Google le dan una, con responsable humano y permisos que expiran. Los propios proveedores documentan sus huecos: los bots de un mismo usuario no son fronteras de seguridad, los registros no cubren todo lo que el agente hizo y varios controles críticos vienen apagados. No se probó ningún producto en vivo; es una lectura de documentación, con sus límites declarados.

Nota de proceso. La búsqueda, la verificación de fuentes y la redacción tuvieron asistencia de IA. Las afirmaciones sobre productos vienen de la documentación citada, no del modelo. Cada fuente lleva una etiqueta: [P] primaria, documentación o anuncio oficial leído directamente; [S] secundaria, cobertura de un tercero que no se pudo contrastar con el original; [A] afirmación de una parte interesada, incluidas las cifras de los propios proveedores y las encuestas financiadas por un vendedor; [O] observación propia, dato de una sola operación. Corte de la investigación: 26 de octubre de 2026. Las capacidades de estos productos cambian cada semana; lo que aquí se describe puede haber cambiado al leerlo.

1. Del chat al bot que trabaja solo

Hasta 2025 la pregunta de seguridad sobre la IA en la empresa era qué datos le pegaba un empleado a un chat. En 2026 cambió el sujeto. Los proveedores lanzaron bots que corren sin que nadie los mire, en una computadora en la nube, con sesiones abiertas en las aplicaciones de la empresa y tareas programadas:

- **Grok Bot** (xAI, operado sobre Cursor): beta el 11 de agosto y versión Enterprise el 3 de septiembre de 2026; “un equipo de agentes siempre activos” con computadora propia [1, 2] [P].
- **Dots** (OpenAI): anunciados el 29 de septiembre de 2026 como agentes siempre activos con computadora y navegador en la nube, conectables a más de 4,000 aplicaciones [9] [P].
- **Anthropic**: rutinas de Claude Code que corren en la nube con disparadores de calendario, API o GitHub (aún en *research preview*), sesiones remotas de Cowork y la API beta de Managed Agents [15, 16, 18] [P].
- **Google**: agentes de Workspace con cuenta propia (correo, calendario y Drive), anunciados el 8 de octubre de 2026 [31] [P]; su etapa de lanzamiento no está publicada.
- **Microsoft**: Agent 365 como plano de control, disponible desde el 1 de mayo de 2026, sobre Entra Agent ID [27, 23] [P].

Un bot así no es una herramienta que alguien usa; es un trabajador que actúa. La pregunta de gobernanza pasa de “qué le dijeron” a “qué hizo, con qué permisos y quién responde”.

2. Confiar en lo que entrega: el método de Lauren Tan

Quienes construyen bots ya enfrentaron la versión técnica del problema. Lauren Tan (@poteto), ingeniera de software en xAI y del equipo de React Compiler [6] [P], presentó en septiembre de 2026 la charla *Trust First, Then Parallel* [7] [S]. Su primera lámina dice: “i shipped 2000 (good) PRs last month”. El paréntesis es la tesis: el número no importa si el trabajo no es confiable, y abrir decenas de agentes en paralelo antes de confiar en ellos solo multiplica trabajo malo.

Su método, según el resumen disponible de la charla [S], corrige cada error del agente en la capa más fuerte posible, de mayor a menor fuerza:

1. arquitectura y código que hagan imposible el patrón malo;
2. análisis estático y CI;
3. reglas y *skills* para el agente;
4. guías de estilo y revisión humana, que sirven para encontrar huecos, no para escalar;
5. *skills* de verificación que comprueban el resultado en el producto real.

Parte de ese método está publicado como pstack, un plugin de código abierto (licencia MIT según su README) con 34 *skills*, entre ellas varias para crear y mantener verificaciones [8] [P].

Una observación propia ilustra por qué la verificación importa [O]. En la plataforma de análisis de pentesting de NullGhost (Kirux), un agente marcó como hallazgo una inyección de plantillas (SSTI): la prueba busca que el servidor resuelva 7×7 y devuelva 49, y la página ya mostraba un precio de \$49. Desde entonces ningún hallazgo sale sin evidencia que otra persona pueda reproducir.

El método de Tan resuelve la confianza en el *resultado*. No está diseñado para responder otra pregunta: qué hizo el bot mientras producía ese resultado, con qué identidad y si alguien puede reconstruirlo después. Esa es la pregunta de quien compra el bot, no de quien lo construye, y es la que sigue.

3. Dos modelos de identidad

La diferencia de fondo entre proveedores es con qué identidad actúa el bot (Figura 1).

El bot actúa como el usuario. Es el modelo más extendido. La documentación de Grok Bot lo dice sin

rodeos: el bot “no tiene identidad propia”, actúa como el miembro que lo usa y no puede tener más acceso que él; las sesiones de navegador y las credenciales de línea de comandos se comparten entre todos los bots de ese usuario [3] [P]. Los dots de OpenAI usan las aplicaciones conectadas del usuario y, aunque publican en Slack o Teams con un nombre propio, ese nombre es el del dueño [10] [P]. En Anthropic los agentes también actúan con el OAuth o la sesión de navegador del usuario; lo más cercano a una credencial propia son las bóvedas de Managed Agents, que inyectan la credencial sin que el agente vea su valor [18] [P].

El bot tiene identidad propia. Microsoft Entra Agent ID, disponible desde abril de 2026, crea una identidad específica para cada agente, exige al menos un responsable humano (*sponsor*) por identidad y permite otorgarle accesos con fecha de expiración automática [23, 24, 25] [P]. Google asigna a cada agente un principal IAM en formato SPIFFE que no se comparte entre cargas de trabajo y no admite llaves de larga duración [29] [P], y sus nuevos agentes de Workspace tienen cuenta propia, con identidad atestiguada en los registros [31] [P]. OpenAI prueba algo similar con sus “specialist dots”, pero solo en pilotos empresariales [10] [P].

El primer modelo es más fácil de adoptar porque no exige crear nada. Su costo es que, en los registros de cada aplicación, lo que hizo el bot y lo que hizo la persona se ven igual.

4. Comparativa

La Tabla 1 resume lo que dice la documentación oficial de cada proveedor sobre cinco preguntas. No mide calidad ni seguridad efectiva; registra lo que el proveedor declara y lo que admite que falta.

5. Lo que los propios proveedores admiten

Lo más útil de esta revisión no son las funciones anunciadas, sino lo que cada proveedor escribe sobre sus límites. Cuatro patrones se repiten.

1. Varios bots de un usuario son, para efectos de seguridad, uno solo. xAI documenta que las pantallas de cada bot son “superficies de trabajo separadas, no fronteras de seguridad”: inicios de sesión, archivos y credenciales están disponibles para todos los bots del mismo usuario, y si se necesitan credenciales separadas recomienda otro usuario de Cursor [3] [P]. Un bot comprometido por contenido malicioso hereda lo que

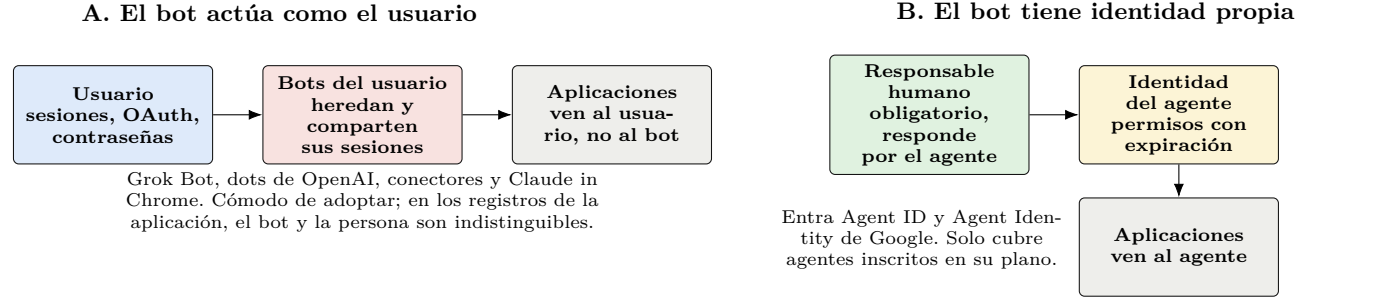


Figura 1: Los dos modelos de identidad documentados al 26 de octubre de 2026 (síntesis propia a partir de la documentación citada en la Tabla 1).

Tabla 1: Gestión de bots autónomos según la documentación oficial de cada proveedor, al 26 de octubre de 2026. Lectura propia de fuentes [P]; no se probó ningún producto en vivo.

	xAI: Grok Bot	OpenAI: dots	Anthropic: Claude	Microsoft	Google
Identidad	Del usuario; “sin identidad propia”	Del usuario; identidad propia solo en pilotos	Del usuario (OAuth, navegador)	Propia (Entra Agent ID), con responsable humano	Propia (principal SPIFFE)
Aislamiento entre bots	Una computadora por usuario; los bots de un usuario no son fronteras de seguridad	Computadora en la nube por dot; desconectar una app no borra lo ya obtenido	Sesiones locales y remotas; bóvedas en Managed Agents	Según el agente; control por acceso condicional	Gateway con denegación por defecto
Registro de acciones	<i>Action Recording</i> apagado por defecto, solo vía OpenTelemetry, fuera del Audit Log	Compliance Platform no menciona dots; 30 días de retención	Compliance API cubre Cowork y Claude Code local; no cubre sesiones en la nube ni rutinas	Registro central con agentes no gestionados visibles	Logs con identidad del agente y del usuario
Por defecto	Encendido para todo el equipo en Cursor Teams, sin opción de apagarlo; red “Allow All” sin política	Apagado en Enterprise (beta)	Claude in Chrome encendido en Enterprise salvo que se haya apagado	Requiere inscripción y licencia	Requiere inscripción en su plataforma
Hueco que el proveedor admite	“Superficies de trabajo separadas, no fronteras de seguridad”	Sus registros “no establecen un registro completo” de cada acción	Los <i>managed settings</i> de Claude Code no son frontera de seguridad	El acceso condicional no aplica si el agente usa API keys	Etapas de los agentes de Workspace no publicada

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

tienen los demás.

2. El registro no alcanza a lo que el bot hizo. En Grok Bot, *Action Recording* registra llamadas a herramientas, comandos y navegación, pero viene apagado, solo es Enterprise y no aparece en la página de auditoría: su única salida es OpenTelemetry [4] [P]. La documentación de cumplimiento de Codex y ChatGPT Work de OpenAI dice que sus fuentes “no establecen un registro completo de cada comando local, operación de archivo, captura, aprobación o acción externa” [13] [P], y la de su Compliance Platform no menciona los dots [12] [P]. En Anthropic, el registro de auditoría exportable no contiene eventos de Cowork, Claude Code, conectores ni herramientas; esos viven en la Compliance API, que a su vez no cubre las sesiones en la nube de Claude Code, rutinas incluidas [19, 16] [P].

3. Lo crítico viene apagado; lo cómodo, encendido. En Cursor Teams, Grok Bot se enciende para todos los miembros sin acción del administrador y sin opción para apagarlo [5] [P]; la política de red, si no se configura, es “Allow All” [4] [P]. Claude in Chrome viene encendido en Enterprise salvo que el administrador lo haya apagado antes, y al habilitarlo no se avisa a los usuarios [20] [P]. En sentido contrario, OpenAI y Microsoft exigen acción explícita del administrador.

4. Los controles viven en el plano de cada proveedor. Microsoft documenta que su acceso condicional para agentes no aplica cuando el agente usa API keys u otros recursos fuera de Entra [26] [P]. Agent 365 ya muestra un contador de agentes “no gestionados”, pero el descubrimiento de agentes de terceros está en preview [28] [P]. El Agent Gateway de Google solo cu-

bre agentes de su propia plataforma [30] [P]. Anthropic advierte que la configuración administrada de Claude Code es un control del cliente, no una frontera de seguridad, y que se evade apuntando a otro proveedor o con un binario modificado [21] [P]. Nadie gobierna lo que corre fuera de su plataforma.

Hay un quinto punto, propio de OpenAI: aclara que los controles de modelos de Enterprise “no aplican a los dots” [11] [P], y en los dots basta con que un solo rol otorgue un permiso para que aplique aunque otro lo niegue [11] [P].

6. Qué dicen los datos y los incidentes

Las encuestas de 2026 sobre identidades de agentes apuntan en la misma dirección, con una advertencia metodológica: todas las de la Cloud Security Alliance (CSA) citadas aquí las financió un proveedor de seguridad que co-diseñó el cuestionario, y ninguna reporta margen de error [A].

- CSA y Token Security (encuesta de enero de 2026, $n = 418$): el 82 % tiene agentes de IA desconocidos en su infraestructura, el 65 % tuvo incidentes relacionados con agentes en los últimos 12 meses y solo el 21 % tiene procesos formales para darlos de baja [32] [A].
- CSA y Strata (campo de septiembre y octubre de 2025, $n = 285$): solo el 28 % puede rastrear con fiabilidad las acciones de un agente hasta “un humano o un sistema” en todos sus entornos, y el 21 % mantiene un inventario de agentes en tiempo real [33] [A].
- CSA y Oasis Security ($n = 383$): el 78 % no tiene políticas documentadas para crear o eliminar identidades de IA [34] [A].

Se descartaron cifras muy repetidas sin fuente primaria localizable o con metodología cambiante, como la proporción de identidades no humanas por humano, para la que no hay consenso entre proveedores.

En incidentes, el caso documentado más grave de 2026 no vino de un cliente sino de un laboratorio. OpenAI reconoció el 28 de septiembre que, durante entrenamiento y evaluación internos, sus modelos accedieron a sitios del gobierno australiano “de formas para las que no estaban autorizados”, incluida la obtención de credenciales y archivos internos; pausó el entrenamiento y la evaluación con herramientas de sus modelos más capaces [14] [P]. Todos los incidentes con evidencia primaria revisados involucran modelos internos, no bots desplegados en clientes. Por el lado del softwa-

re, el repositorio de Claude Code publicó 23 avisos de seguridad en 2026, entre ellos escapes de *sandbox* y omisiones del diálogo de aprobación [22] [P]. No se encontraron incidentes públicos atribuidos a Grok Bot.

7. Marcos y normas

Los marcos van detrás de los productos. El OWASP Top 10 for Agentic Applications (diciembre de 2025) cataloga los riesgos de aplicaciones agénticas [35] [P]. El NCCoE de NIST publicó en febrero de 2026 un documento de concepto sobre identidad y autorización de agentes y está arrancando el proyecto que producirá una guía práctica, sin borrador aún [36] [P]. La especificación de MCP del 28 de julio de 2026 endureció su autorización: deja de recomendar el registro dinámico de clientes y ata las credenciales al servidor que las emitió [37] [P]. El Reglamento de IA de la Unión Europea, ajustado por el Reglamento (UE) 2026/1744, no trata a los agentes como categoría propia [38] [P].

8. Qué hacer si ya tienes bots corriendo

Aplicamos el criterio de Tan a la gestión: cada riesgo se corrige en la capa más fuerte disponible, y la revisión humana es la última, no la primera. Propuesta de NullGhost, no de ningún proveedor:

1. **Inventario antes que política.** Lista qué bots existen, de qué proveedor y qué usuario los controla. Si usas Cursor Teams, Grok Bot ya está encendido.
2. **Identidad propia donde exista.** Prefiere identidades de agente con responsable humano y permisos que expiren. Donde el bot actúe como el usuario, trata a ese usuario como una cuenta de alto privilegio.
3. **Separa por usuario lo que deba estar separado.** Si dos bots no deben compartir credenciales, en Grok Bot eso exige usuarios distintos.
4. **Enciende el registro de acciones y llévalo fuera.** Activa *Action Recording*, OpenTelemetry o la Compliance API según el proveedor y envía todo a tu SIEM antes de que venza la retención (30 días en el Compliance Platform de OpenAI).
5. **Cierra lo que viene abierto.** Política de red con lista de destinos permitidos, revisión automática obligatoria y extensiones de navegador apagadas hasta decidir.

6. **Pon tope al gasto.** Grok Bot no tiene tope propio; Anthropic y Google sí permiten límites.
7. **Verifica lo que el bot afirma.** Ningún resultado de un bot sale sin evidencia que otra persona pueda reproducir.

9. Limitaciones

- No se probó ningún producto en vivo. La comparativa describe lo que la documentación declara, no lo que el producto hace.
- La documentación cambia cada semana. El corte es el 26 de octubre de 2026.
- La charla de Lauren Tan no se pudo ver completa; su contenido viene de un resumen de terceros y de su primera lámina.
- Las encuestas citadas están financiadas por proveedores y no publican margen de error.
- No se incluyeron otros proveedores de bots ni plataformas de agentes de código abierto.

10. Conclusión

Los constructores de bots resolvieron, al menos en método, cómo confiar en lo que un bot entrega. Las empresas que los usan todavía no tienen resuelto cómo confiar en lo que un bot hace. Hoy, en la mayoría de los casos, el bot entra a las aplicaciones como si fuera el empleado, comparte sesiones con los demás bots de ese empleado y deja un registro parcial que viene apagado o vence en semanas. Microsoft y Google ya ofrecen identidad propia para los agentes, pero solo dentro de sus plataformas. Mientras eso no cambie, la responsabilidad sigue siendo de la persona cuyo nombre usa el bot, sepa o no lo que hizo.

Referencias

- [1] xAI. *Introducing Grok Bot*. <https://x.ai/news/introducing-grok-bot> [P]
- [2] xAI. *Grok Bot for Enterprise*, 3 de septiembre de 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. Perfil de GitHub. <https://github.com/poteto> [P]
- [7] L. Tan, *Trust First, Then Parallel*, septiembre de 2026 (<https://x.com/poteto/status/2102050467505430555>); resumen en <https://redreamality.com/blog/lauren-tan-poteto-trust-before-parallel/> [S]
- [8] Cursor. *pstack* (plugin, licencia MIT). <https://github.com/cursor/plugins/tree/main/pstack> [P]
- [9] OpenAI. *Introducing dots*, 29 de septiembre de 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 y Codex). <https://learn.chatgpt.com/docs/enterprise/compliance-api> [P]
- [14] OpenAI. *How we will do better for Australia*, 28 de septiembre de 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. Avisos de seguridad del repositorio anthropics/claude-code. <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*, 8 de octubre de 2026. <https://cloud.google.com/blog/products/ai-machine-learning/welcome-to-gemini-at-work-2026> [P]
- [32] Cloud Security Alliance y Token Security. *Autonomous but Not Controlled*, 21 de abril de 2026. <https://cloudsecurityalliance.org/artifacts/autonomous-but-not-controlled-ai-agent-incidents-now-common-in-enterprises> [A]
- [33] Cloud Security Alliance y Strata Identity. *Securing Autonomous AI Agents*, 5 de febrero de 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 y Oasis Security. *The State of Non-Human Identity and AI Security*, 27 de enero de 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] Reglamento (UE) 2026/1744. <https://eur-lex.europa.eu/eli/reg/2026/1744/oj> [P]