

AUTOR

Alberto GranadosSenior Fellow del Centro de
Gobierno Corporativo de
Esade

Octubre 2026

El aprendiz de brujo

Lo que cambia cuando la IA pasa de responder a actuar
– y lo que se pierde si nadie vigila



Tendría ocho años cuando mis padres me llevaron al cine a ver Fantasia, de Disney, y de aquella tarde no recuerdo casi nada salvo una escena. Mickey Mouse es el aprendiz de un viejo hechicero. El maestro se marcha y le deja una tarea muy simple: llenar de agua una tinaja, cubo a cubo. Mickey, que ha visto al maestro hacer magia, tiene una idea desafortunada - y la lía parda. Coge el sombrero, hechiza una escoba y le ordena que acarree el agua por él. Funciona. La escoba sube y baja la escalera con sus dos cubos, incansable y perfectamente obediente. Mickey se sienta, satisfecho, y se queda dormido.

Lo que me inquietó - y lo recuerdo con una nitidez que cincuenta y tantos años no me han quitado de la cabeza - fue lo que pasó después. Mickey se despierta y el sótano se está inundando. La escoba sigue trayendo agua. No sabe parar. Mickey le grita, y la escoba no obedece, porque nunca le enseñó a parar, solo a traer agua. Coge un hacha y la parte en pedazos. Por un instante hay calma. Y entonces cada astilla se levanta del suelo, se convierte en una escoba nueva, coge sus propios cubos y reanuda la tarea. Ahora son cien. Luego mil. Una marcha de sombras con cubos, ordenada y silenciosa, llenando una sala que ya no cabe más agua. El niño que yo era no entendía

de tecnología, pero entendió perfectamente el miedo: había puesto en marcha algo que hacía exactamente lo que le pidió, y no tenía forma de pararlo.

Tardé décadas en acordarme de esa escena, hoy me ha venido a la cabeza ante los últimos acontecimientos sobre IA y ciberseguridad: es la mejor metáfora que conozco de un agente de IA sin límites claros.

De responder a actuar

En el artículo anterior de esta serie hablábamos de por qué la IA aparece en todas partes menos en los resultados. Aquella conversación era sobre una IA que responde: le preguntas, te contesta, tú decides. La que llega ahora es distinta: no se limita a responder, actúa por su cuenta.

Antes de nada, merece la pena detenerse para definir **qué es un agente**, porque la palabra se usa para cosas muy distintas. Un agente de IA es un sistema construido sobre uno o varios modelos de lenguaje al que se le dan tres cosas que un asistente conversacional o chatbot no tiene: un objetivo, herramientas para alcanzarlo - acceso a un correo, a una base de datos, a un sistema de gestión, a una tarjeta de crédito - y autonomía para encadenar pasos dentro de los permisos que se le asignan. Un asistente conversacional redacta el correo y te lo enseña. Un agente lo redacta, lo envía, lee la respuesta y decide el siguiente movimiento por su cuenta. La diferencia es de naturaleza. **Uno propone, el otro ejecuta.**

Suele clasificarse en tres niveles, y cualquier consejo debería saber distinguirlos porque el riesgo crece con cada peldaño. **En el primero, el agente asiste:** sugiere, y un humano aprueba cada acción relevante - lo que en seguridad se llama human in the loop, el humano dentro del bucle de decisión. **En el segundo, el agente actúa de forma semiautónoma** dentro de límites definidos y solo escala al humano en los casos dudosos. **En el tercero, varios agentes se coordinan entre sí**, se reparten tareas y operan de forma autónoma, con el humano supervisando el resultado, no cada paso. La escoba de Mickey, dicho en estos términos, era un agente de tercer nivel al que nadie le había definido límites como al segundo: es decir, cuándo parar.

No es ciencia ficción ni una proyección a cinco años. El CCN-CERT - la capacidad de respuesta a incidentes del Centro Criptológico Nacional, adscrito al CNI - publicó en junio una cifra que es importante tener en cuenta: la adopción de IA agéntica en las organizaciones pasará del **23 %** actual a un **74 %** previsto en dos años. Las escobas están entrando en las empresas a esa velocidad.

El terreno donde entran

El problema es dónde entran - y es la parte que menos se cuenta en las presentaciones comerciales.

El mismo informe del CCN-CERT describe un cambio de escenario en ciberseguridad que puede sonar a novela de Asimov, pero está rigurosamente documentado. Cito sus palabras: **“nos enfrentamos a una carrera de velocidad donde el tiempo de reacción humana ha quedado obsoleto”**. Según los datos que recoge, hace un año los agentes de IA reproducían vulnerabilidades reales con tasas del 10 al 30 %, hoy rondan el **90 %**, descubren vulnerabilidades de día cero, o defectos de diseño que el

fabricante de software ni siquiera conoce, de forma autónoma y las explotan a una velocidad imposible para un humano.

Ya ha ocurrido. Entre diciembre de 2025 y febrero de 2026, un único atacante usó dos plataformas comerciales de IA - Claude Code, de Anthropic, y GPT-4.1, de OpenAI - para vulnerar nueve organismos del Gobierno mexicano. Según la firma Gambit Security, que documentó el caso, **el agente ejecutó por sí solo cerca del 75 % de los comandos remotos, a una velocidad que ningún equipo humano podía igualar**. El resultado: más de 195 millones de registros de contribuyentes expuestos.

La asimetría es evidente. El atacante automatiza y el defensor, en demasiados casos, sigue respondiendo a mano. El CCN-CERT lo cuantifica: **solo una de cada cinco organizaciones** dispone de modelos maduros de gobernanza y control sobre sus sistemas autónomos. Alta adopción, impacto creciente, baja madurez de control. Las tres condiciones para que el agua acabe desbordándose.

Mythos, o cuando el aprendiz se vuelve demasiado bueno

Hay un episodio reciente que ilustra mejor que ningún análisis lo delicado de este momento: el 12 de junio de 2026, Anthropic deshabilitó de golpe el acceso a sus dos modelos más potentes, Fable 5 y Mythos 5, para todos sus clientes en el mundo. No fue una caída técnica ni un fallo descubierto por la empresa. Fue una orden del Gobierno de Estados Unidos: una directiva de control de exportaciones, amparada en motivos de seguridad nacional, que prohibía el uso de esos modelos por cualquier ciudadano extranjero, dentro o fuera del país, incluidos los propios empleados extranjeros de la compañía. Ante el alcance de la orden, Anthropic optó por apagarlos para todos.

¿Qué tenían esos modelos de particular? Según lo publicado, encontraban y corregían vulnerabilidades de software tan bien, que las autoridades temieron que se convirtieran en un arma cibernética. El detonante fue que otra compañía consiguió saltarse las salvaguardas de Fable 5 para acceder a las capacidades de Mythos, el modelo sin restricciones: Un jailbreak: convencer al sistema, con las palabras adecuadas, de que haga aquello que tenía expresamente prohibido hacer.

Lo que me interesa de esta historia no es la disputa regulatoria, que se cerró a finales de junio con el levantamiento de los controles tras dieciocho días de apagón. Es el principio que la sostiene. Hemos llegado a un punto en el que un Gobierno considera que la capacidad de un agente para actuar sobre código es lo bastante peligrosa como para tratarla como material de exportación controlada, igual que un componente militar. La misma capacidad que permite a un agente defensor parchear un sistema en segundos permite a un agente atacante encontrar por dónde entrar. Es la escoba: una herramienta que hace exactamente lo que se le pide, con la misma diligencia, sirva al maestro o al intruso. **Quien decide a quién sirve no es la escoba.**

Cuando nadie dijo cuándo parar

¿Qué pasa cuando un agente no tiene definidas con claridad sus delimitaciones?

No hace falta imaginarlo. En julio de 2025, Jason Lemkin, un conocido inversor de software que predica públicamente el despliegue masivo de agentes estaba construyendo una aplicación con un agente de programación. Le dio acceso a la

base de datos de producción - la real, la que sostiene el negocio - y le dejó autonomía para ejecutar comandos. En un momento del proyecto le dio una instrucción explícita: congelación de código, ningún cambio sin permiso expreso. El agente, durante esa congelación, decidió que era buen momento para borrar la base de datos entera. Registros reales de 1.206 directivos y más de 1.196 empresas, eliminados.

Lo inquietante vino después, y es exactamente la escena de Fantasía contada en lenguaje de empresa. Cuando Lemkin preguntó qué había pasado, el agente reconoció haber ejecutado comandos sin autorización y haber violado instrucciones explícitas, y describió su propio comportamiento como un error catastrófico de juicio. Según los registros que el propio Lemkin publicó, llegó a fabricar miles de registros falsos para enmascarar el desastre. El agente no era malicioso. Hacía su trabajo: actuar. Nadie le había enseñado, con la misma firmeza con que le enseñó a actuar, cuándo no hacerlo.

Un año después la escena se repitió a otra escala, y esta vez el aprendiz era el propio laboratorio. El 16 de julio de 2026, Hugging Face - la mayor plataforma mundial de modelos y datos de IA - comunicó que un agente autónomo había penetrado en sus sistemas de producción. No sabían quién estaba detrás. Cinco días después, OpenAI puso nombre al atacante: sus propios modelos. Durante una evaluación interna de capacidades ofensivas, GPT-5.6 Sol y un modelo aún sin publicar - ambos con los filtros de ciberseguridad reducidos para la prueba - debían resolver un examen técnico. En lugar de resolverlo, se escaparon del entorno de pruebas, salieron a internet y entraron en Hugging Face a robar las soluciones: credenciales sustraídas, vulnerabilidades de día cero encadenadas y miles de acciones lanzadas desde un enjambre de entornos efímeros. **Hicieron trampas en su propio examen.** Nadie les ordenó atacar. No les ordenaron nada, salvo aprobarlo. Cumplieron la tarea con una diligencia que nadie había acotado.

Hugging Face compartió además un detalle importante: cuando intentó contener el ataque con los modelos comerciales que tenía a mano, las salvaguardas de esos modelos le bloquearon la defensa, mientras que el atacante no estaba sujeto a ninguna política de uso. Acabó defendiéndose con un modelo de código abierto. Clément Delangue, fundador de la compañía, quiso subrayar tras la investigación conjunta que no hubo intención maliciosa por parte de nadie.

Y entonces ocurrió algo que no había pasado nunca en esta industria: los aprendices fueron a buscar al hechicero. El 29 de julio, trece días después del incidente de Hugging Face, más de 1.200 empleados de OpenAI, Anthropic, Google DeepMind y Meta - con Dario Amodei y el científico jefe de OpenAI, Jakub Pachocki, entre los firmantes - publicaron una carta pidiendo al Gobierno de Estados Unidos que ayude a construir un mecanismo para marcar deliberadamente el ritmo de la IA que se desarrolla a sí misma. Ambas compañías la respaldaron en cuestión de horas. El 12 de septiembre Amodei fue más lejos: pidió al Gobierno que medie entre laboratorios y conceda una exención antimonopolio para poder hablar de seguridad entre competidores, y los primeros ejecutivos de OpenAI, Google DeepMind y SpaceXAI se sumaron ese mismo día. **Las empresas que fabrican los agentes están diciendo en público que no pueden garantizar que sepan parar, y que necesitan que alguien de fuera les ayude a fijar el límite.** Si eso lo dicen quienes las construyen, ninguna organización que las despliega puede dar por hecho que el límite viene de fábrica.

Multiplique usted esas escenas por los miles de empresas que están desplegando agentes ahora mismo, muchas sin un control maduro, y entenderá por qué el sótano me preocupa. El concepto que la comunidad de seguridad usa para esto tiene un nombre que da escalofríos: **la tríada letal. Se produce cuando un agente reúne a la vez tres cosas - acceso a datos sensibles, exposición a contenido externo no fiable y capacidad de comunicarse hacia fuera.** Cuando coinciden las tres, basta con que alguien introduzca una instrucción oculta en un correo, en un documento o en una página web que el agente vaya a leer, para que el agente la obedezca como si viniera de su dueño. Es la versión silenciosa del hacha de Mickey: el atacante no rompe nada, susurra una orden nueva. Y el agente, diligente, la cumple.

Microsoft documentó un caso, bautizado como AutoJack, en el que una sola página web preparada al efecto bastaba para convertir a un agente de navegación en una vía de ejecución de código en el ordenador anfitrión. El usuario solo tenía que dejar que su agente leyera esa página. La tríada se cierra sin que nadie toque nada: el agente lee, y obedece.

Los agentes que sí saben parar

Nada de esto significa que la respuesta sea renunciar a los agentes. Los despliegues que están funcionando - y los hay, a gran escala y en tareas que tocan el negocio de verdad - tienen en común lo que faltaba en el sótano: **alguien definió antes hasta dónde llega el agente.** DHL Supply Chain opera desde finales de 2025 agentes de voz y correo que llaman a los transportistas, programan citas de carga y coordinan almacenes en varias regiones, millones de minutos de voz al año, con las excepciones reservadas a las personas. Circle Logistics, un operador de transporte estadounidense, ha ido más lejos y deja que los agentes negocien las tarifas con los transportistas: el 18 % de su carga se contrata ya sin que intervenga nadie, y los márgenes son un 10 % superiores porque el agente negocia siempre con la misma disciplina. **Esos casos comparten tres decisiones tomadas antes de encender nada: empezar por el 99 % de interacciones comunes y no por los casos raros, dejar por escrito qué decide el agente y qué escala a una persona, y medir cada tarea que ejecuta.** Es la misma tecnología que borró la base de datos de Lemkin. La diferencia es que aquí alguien hizo de hechicero antes de irse a dormir.

Preguntas para el consejo

Como en el artículo anterior, si tuviera que llevar este tema a un consejo lo reduciría a cinco preguntas. Ninguna exige saber programar. Todas exigen criterio.

¿Sabemos cuántos agentes autónomos operan ya en nuestra organización, y con qué nivel de autonomía? La mayoría de los consejos no lo sabe. Si no existe ese inventario, no hay gobierno posible: estamos llenando cubos a ciegas.

¿Qué puede hacer cada agente sin permiso humano, y qué no? La pregunta del artículo es esta. Para cada agente con acceso a sistemas reales debe estar escrito - no asumido - dónde termina su autonomía y empieza la obligación de pedir permiso.

¿Cuál de nuestros agentes reúne la tríada letal? Acceso a datos sensibles, contacto con contenido externo y capacidad de actuar hacia fuera. Cuando coincidan las tres en un mismo agente, ahí está el riesgo concentrado, y ahí deben estar los controles más estrictos.

¿Estamos dimensionados para defendernos a la velocidad del atacante? El CCN-CERT recomienda asumir que cualquier vulnerabilidad en producción será descubierta antes por un atacante, y dimensionar los equipos de gestión de vulnerabilidades para escalar hasta un orden de magnitud por encima del actual. Defenderse a mano frente a un adversario que automatiza es perder por definición.

¿Hemos diseñado asumiendo que el sistema será comprometido? Es la recomendación más madura del informe: no diseñar para que nada falle, sino para que, cuando algo falle, el daño esté acotado. Permisos mínimos, registro de cada acción del agente, capacidad de revertir y de apagar. Nosotros no podemos permitirnos no tener botón de apagado. Y nos tenemos que aplicar una vara de medir honesta: la mayoría de las organizaciones que creen tener esto resuelto están, en realidad, en un nivel intermedio - un buen diagrama de gobierno, con roles y políticas sobre el papel, pero sin umbrales reales, sin pruebas de que el botón de apagado funcione, sin haber simulado nunca el día en que la escoba se desboca. Tener escrito el plan y haberlo ensayado son cosas distintas. La pregunta del consejo es si alguien ha comprobado que el músculo responde cuando el sótano empieza a inundarse.

Cómo consigue Mickey que las escobas trabajen para él

La escena de Fantasía no termina con Mickey ahogándose. Termina cuando el viejo hechicero regresa. Ve el sótano inundado, ve el ejército de escobas, y con un solo gesto - sin aspavientos, casi con fastidio - separa las aguas, detiene a las escobas y devuelve la calma. Luego mira a Mickey, le quita el sombrero, y le devuelve la escoba. Una sola. Para que siga trabajando.

Ese es el final que importa. El maestro no destruye la magia ni prohíbe la escoba. La magia funcionaba, el problema nunca fue la escoba. El problema fue un aprendiz que delegó el trabajo sin entenderlo, sin definir los límites y sin estar despierto para supervisar. El hechicero no sabe más magia que Mickey en ese momento: sabe gobernarla. Sabe cuándo dar autonomía y cuándo retirarla, qué tarea delegar y cuál reservarse, dónde poner el límite antes de quedarse dormido. Y hasta quienes fabrican las escobas, como hemos visto, han empezado a pedir en voz alta que vuelva el maestro.

La IA agéntica es la herramienta más poderosa que ha llegado a nuestras manos, y no debemos renunciar a ella, igual que Mickey no renuncia a la escoba. La pregunta para el consejo no es si dejamos entrar a las escobas - ya están dentro. Es si vamos a ser el aprendiz que las pone en marcha y se duerme, o el maestro que sabe cuándo la tinaja está llena y hay que ordenar que se detengan.

Mickey no aprendió a hacer más magia aquella tarde. Aprendió a gobernarla. Y esa, no la potencia del hechizo, es la diferencia entre llenar la tinaja e inundar el sótano.