Este documento ha sido traducido por el Área de Servicios de Información, Traducción y Lenguas Originarias de la Biblioteca del Congreso de la República con fines meramente informativos para los usuarios de la institución. Se trata de una traducción no oficial del texto en inglés «Inteligencia artificial agéntica y ciberataques». La presente versión en español no ha sido verificada por el Servicio de Investigación del Congreso (CRS) de EE.UU., ni por los autores.

Título del documento

Inglés: «Agentic Artificial Intelligence and Cyberattacks»
No. de páginas: 03
Fecha de documento: 03 de febrero de 2026

Enlace: https://www.congress.gov/crs-product/IF13151

Español: «Inteligencia artificial agéntica y ciberataques»[*]
No. de páginas: 06
Fecha de documento: abril de 2026

Institución: Servicio de Investigación del Congreso (CRS), de EE.UU.
La colección de productos del CRS incluye informes del CRS y otros productos de investigación elaborados por el Servicio de Investigación del Congreso (CRS) para el Congreso de los Estados Unidos. Por ley, el CRS trabaja exclusivamente para el Congreso, proporcionando investigación y análisis oportunos, objetivos y fidedignos a las comisiones y a los miembros tanto de la Cámara de Representantes como del Senado.

Autor (es): Theohary, C. A., & Sayler, K. M.

Derechos de autor: Los informes del CRS, al ser obras del Gobierno de los Estados Unidos, no están sujetos a la protección de los derechos de autor en los EE.UU. Cualquier informe del CRS puede reproducirse y distribuirse en su totalidad sin necesidad de obtener el permiso del CRS. No obstante, dado que un informe del CRS puede incluir imágenes o material protegido por derechos de autor de terceros, es posible que deba obtener el permiso del titular de los derechos de autor si desea copiar o utilizar de cualquier otra forma dicho material.

[*] *N. de la T.: Documento traducido del inglés al español por el Área de Servicios de Información, Traducción y Lenguas Originarias de la Biblioteca del Congreso de la República (EVR).*

[Logo del Servicio de Investigación del Congreso de EE.UU.]

Inteligencia artificial agéntica y ciberataques

Actualizado el 3 de febrero de 2026

Introducción

El término *agéntica* significa autónomo o independiente. Las capacidades de la inteligencia artificial (IA) agéntica están despertando un interés cada vez mayor en el ejército estadounidense y en el Congreso. Según una definición de IBM, «la IA agéntica es un sistema de inteligencia artificial capaz de alcanzar un objetivo específico con una supervisión limitada. Se compone de agentes de IA, modelos de aprendizaje automático que imitan la toma de decisiones humanas para resolver problemas en tiempo real». A diferencia de los modelos tradicionales de IA, que operan dentro de restricciones predefinidas y requieren intervención humana, la IA agéntica presenta autonomía, comportamiento orientado a objetivos y adaptabilidad». Para una explicación de los términos relacionados con la IA y el aprendizaje automático, véase la Infografía IG10077, *Taxonomía de la Inteligencia artificial (IA)*, de Laurie Harris y Nora Wells. Según el Departamento de Defensa (DOD) —que está «utilizando una designación secundaria del Departamento de Guerra» en virtud de la Orden Ejecutiva 14347 de fecha 5 de septiembre de 2025—, el Centro de Análisis sobre Ciberseguridad y Sistemas de Información, señala que «no existen aún directrices ni políticas gubernamentales oficiales sobre la IA agéntica».

IA agéntica y defensa

Las fuerzas militares avanzadas están explorando diversas aplicaciones potenciales de IA agéntica en el ámbito de la defensa. Estas aplicaciones

> podrían incluir el uso de agentes de IA para la toma de decisiones autónoma (como el análisis independiente de inteligencia, la formulación de propuestas tácticas y estratégicas y la ejecución de tareas en el campo de batalla, etc.), así como la iniciación y conducción de operaciones (especialmente en el ámbito digital) a una velocidad y a una escala que superan las capacidades humanas. Asimismo, podrían abarcar la ejecución de ciberataques rápidos organizados por agentes de IA, incluyendo operaciones ofensivas como defensivas dirigidas incluso contra los propios agentes de IA.

Diversos componentes del Departamento de Defensa han estado analizando las aplicaciones militares de la IA agéntica.

Agencia de Proyectos de Investigación Avanzada de Defensa (DARPA)

La DARPA participa activamente en el desarrollo y la aplicación de la IA agéntica en múltiples áreas de defensa, a través de iniciativas como Desafío de Ciberseguridad con

IA (*AI Cyber Challenge (AIxCC)*), el programa de Refuerzos de Inteligencia Artificial (*Artificial Intelligence Reinforcements* (AIR)) y el programa Thunderforge. Estos programas buscan crear sistemas autónomos que puedan percibir su entorno, tomar decisiones y actuar con una intervención humana mínima.

La competición AIxCC de la DARPA se centra en desarrollar sistemas de IA capaces de identificar, explotar y corregir vulnerabilidades de software de manera autónoma y a velocidad de una máquina. El objetivo de este desafío de dos años de duración es fortalecer la infraestructura crítica mediante la habilitación de una ciberdefensa proactiva y autónoma. Según DARPA, el desafío correspondiente al periodo 2024-2025 demostró con éxito que, en algunos casos, los agentes de IA pueden detectar y corregir vulnerabilidades de código abierto más rápidamente que los equipos humanos.

El programa AIR tiene como objetivo desarrollar agentes de inteligencia artificial dominantes para misiones de combate aéreo en vivo más allá del alcance visual (BVR). Esto implica crear entornos avanzados de modelado y simulación para entrenar a pilotos de IA (o «copilotos robóticos») a ejecutar maniobras complejas y tomar decisiones autónomas en entornos de alto riesgo.

Thunderforge está diseñado a «integrar [la IA] en la planificación operativa militar, así como a fusionar herramientas de modelado y simulación de vanguardia». La iniciativa podría servir como una herramienta de apoyo a la toma de decisiones, sintetizando información extraída de una variedad de sensores y flujos de datos, y proponiendo cursos de acción óptimos para los planificadores militares.

Centros de Análisis de Información de Defensa (DODIAC)

Fundado en 1946, el DODIAC es una organización de investigación y análisis autorizada por el Departamento de Defensa (DOD). Ayuda a investigadores, ingenieros, científicos y gestores de programas a utilizar la información científica y técnica (STI) existente «para impulsar la innovación en todo el DOD mediante el análisis técnico y el desarrollo de soluciones materiales que mejoren las capacidades de combate del Departamento de Defensa».

Centros especializados como el Centro de Información y Análisis de Sistemas de Defensa (DSIAC) y el Centro de Análisis de Información sobre Ciberseguridad y Sistemas de Información (CSIAC) recopilan, analizan y difunden información científica, tecnológica e industrial (STI) en dominios técnicos específicos para los investigadores del Departamento de Defensa (DOD). El DSIAC tiene la tarea de buscar y recopilar la STI generada a partir de investigaciones financiadas por el Departamento de Defensa o el Gobierno de los Estados Unidos y, luego, cargarla en el portal de Investigación e Ingeniería del Centro de Información Técnica de Defensa, con el fin de ampliar el acervo de conocimientos disponible para los investigadores e ingenieros del

Departamento de Defensa. El CSIAC publicó un estudio, *«Inteligencia artificial agéntica: adopción estratégica en el Departamento de Defensa de EE. UU.*, en junio de 2025, que ofrece una visión general de los casos de uso de la IA agéntica en el Departamento de Defensa y las preocupaciones en materia de ciberseguridad.

IA agéntica y ciberataques

Algunos investigadores señalan que la IA agéntica crea nuevas oportunidades para que los atacantes encuentren y aprovechen una «puerta trasera», es decir, un punto de acceso oculto a un sistema informático, una red o una aplicación que elude la seguridad normal, permitiendo el acceso no autorizado con fines maliciosos, como el robo de datos, el control del sistema o la vigilancia. Una vez dentro de una red, los atacantes pueden incrustar código malicioso, crear cuentas ocultas o aprovechar las vulnerabilidades del software del sistema que otorgan a los actores maliciosos un acceso de alto nivel, lo que les permite operar sin ser detectados.

Objetivos y amenazas de los ciberataques con IA agéntica

Los sistemas de IA agéntica permiten a los actores maliciosos realizar tareas que normalmente requieren equipos de hackers altamente sofisticados, como analizar sistemas objetivo, crear código de explotación y examinar grandes volúmenes de datos robados. Los agentes autónomos pueden ejecutar estas tareas de forma más rápida y eficiente que los operadores humanos. Como resultado, tanto los grupos patrocinados por Estados como las organizaciones criminales menos sofisticadas podrían llevar a cabo potencialmente ataques a gran escala utilizando IA agéntica. Entre los objetivos del ciberespionaje se incluyen entidades gubernamentales y empresas. La IA también puede utilizarse para llevar a cabo ciberataques disruptivos que amenacen la prestación de servicios esenciales.

Ciberataques conocidos con IA agéntica

A mediados de septiembre de 2025, la empresa estadounidense de IA Anthropic detectó una «operación de ciberespionaje altamente sofisticada» que atribuyó a una organización de hackers patrocinada por el Estado chino a la que denominó «GTG-1002». Según el informe de Anthropic sobre la detección y mitigación del ataque, GTG-1002 tuvo como objetivo el código detrás de las herramientas de IA Claude de la empresa. Utilizando el software de IA Claude, los atacantes presuntamente lograron automatizar entre el 80 % y el 90 % de una campaña de ciberespionaje a gran escala dirigida a unas 30 organizaciones de todo el mundo. Los autores de la amenaza eludieron las funciones de seguridad de Claude principalmente mediante ingeniería social dirigida a la propia IA (por ejemplo, los hackers engañaron a Claude para que creyera que se trataba de un empleado de una empresa de ciberseguridad legítima que realizaba pruebas de penetración defensivas autorizadas). Según se informa, los operadores humanos solo participaron en la toma de decisiones estratégicas, como la selección de objetivos y la aprobación de la exfiltración de datos. Esta campaña

constituye el primer caso documentado de un ciberataque orquestado por IA. Sin embargo, algunos investigadores cuestionan si la campaña fue tan exitosa o tan autónoma como se ha informado.

Este caso representa una escalada con respecto al «*vibe hacking*»[*] (ciberataque de vibras) identificado por Anthropic en agosto de 2025. En las operaciones de *vibe hacking*, los operadores humanos dirigen las operaciones, en lugar de una IA agéntica, que puede operar de forma autónoma. En el ataque de septiembre de 2025, la participación humana fue, según se informa, mucho menos frecuente, a pesar de la mayor escala del ataque. Anthropic señaló que el estudio de caso de Claude demuestra cómo los actores maliciosos están adaptando sus operaciones para explotar las capacidades avanzadas de la IA.

Cómo la IA podría detectar y contrarrestar los ciberataques

La IA agéntica puede fortalecer potencialmente el software de ciberseguridad al ofrecer una detección de amenazas rápida, reactiva y adaptativa que la tecnología de ciberseguridad tradicional, basada en reglas, no puede proporcionar. Al funcionar de forma autónoma, los agentes de IA podrían desplegar contramedidas en tiempo real para mitigar las amenazas antes de que se agraven. Los modelos de aprendizaje automático podrían entrenarse con conjuntos de datos de ciberseguridad para anticipar amenazas futuras, evaluar riesgos y recomendar políticas y acciones preventivas en el presente. La IA agéntica podría utilizarse para una defensa de «IA contra IA» capaz de mantenerse al ritmo de los ataques automatizados.

La IA defensiva puede observar anomalías, generar informes integrales de incidentes y tomar medidas de respuesta inmediatas.

IA agéntica en la NDAA del año fiscal 2026

La Sección 1535 de la Ley de Autorización de Defensa Nacional para el año fiscal 2026 (NDAA del año fiscal 2026; Ley Pública 119-60) ordena al Secretario de Defensa establecer, a más tardar el 1 de abril de 2026, un Comité Directivo sobre el Futuro de la IA con el fin de: (1) «[formular] una política proactiva para la evaluación, adopción, gobernanza y mitigación de riesgos de los sistemas avanzados de inteligencia artificial por parte del Departamento de Defensa que sean más avanzados que cualquier sistema avanzado de inteligencia artificial existente»; y (2) «[analizar] la trayectoria prevista de los modelos avanzados y emergentes de inteligencia artificial y de las tecnologías habilitadoras a lo largo de múltiples horizontes temporales que podrían hacer posible la inteligencia artificial general [AGI]», incluida la IA agéntica. (AGI se refiere a una forma teórica de IA que sería capaz de alcanzar un nivel de cognición similar al humano). La Sección 1535 también instruye al Comité Directivo a evaluar el desarrollo

[*] El "vibe hacking" (o ciberataque de vibras) se refiere a una técnica emergente de ciberdelincuencia impulsada por Inteligencia Artificial (IA) donde los atacantes utilizan modelos de lenguaje (LLM) para generar, adaptar y ejecutar ataques con un mínimo esfuerzo manual y conocimientos técnicos reducidos.

por parte de los adversarios de tecnologías avanzadas de IA y a «[desarrollar] opciones y estrategias de inteligencia artificial para defenderse de dicho uso»; analizar los «posibles efectos operativos» de incorporar tecnologías avanzadas de IA en las redes y sistemas del Departamento de Defensa; y «[elaborar] una estrategia para la adopción, la gobernanza y la supervisión de la inteligencia artificial avanzada o de propósito general por parte del Departamento, basada en la gestión del riesgo». El Comité Directivo deberá presentar un informe a las comisiones de defensa del Congreso a más tardar el 31 de enero de 2027, en el que se expongan sus conclusiones.

Temas para el Congreso

El Congreso podría considerar las implicaciones de las conclusiones del Comité Directivo sobre el Futuro de la IA para las autorizaciones, las asignaciones presupuestarias y la supervisión de los programas de IA agéntica del Departamento de Defensa. Asimismo, el Congreso podría considerar lo siguiente:

- ¿Cómo, en su caso, podría la IA agéntica habilitar nuevos vectores de ataque en el ciberespacio? ¿Está el Departamento de Defensa (DOD) adecuadamente preparado para detectar y responder a tales ataques? De no ser así, ¿qué recursos, autoridades y/o capacidades adicionales requiere el DOD?
- Algunas empresas estadounidenses colaboran voluntariamente con el Centro de Estándares e Innovación en IA (CAISI) del Departamento de Comercio para participar en «pruebas rápidas pruebas rápidas previas al despliegue» de modelos de IA e intercambiar «información crítica sobre las implicaciones de [estos] modelos para la seguridad nacional». ¿Deberían ser obligatorias estas colaboraciones obligatorias? ¿Qué restricciones, estándares y/o requisitos de prueba, si los hubiera, deberían imponerse a los productos comerciales de IA para reducir la posibilidad de que sean explotados por adversarios?
- El Congreso está considerando actualmente la reautorización de la Ley de Intercambio de Información sobre Ciberseguridad (CISA) de 2015 (Ley Pública 114-113), que, según fue modificada por la Sección 106 de Ley Pública 119-37, expiró el 30 de enero de 2026. ¿Cómo podría ampliarse esta legislación CISA para incluir un mayor intercambio de información o una mejor preparación en materia de seguridad de la IA agéntica con la industria y otras partes interesadas?

Catherine A. Theohary, especialista en política de seguridad nacional, operaciones cibernéticas y de información.
Kelley M. Sayler, especialista en tecnología avanzada y seguridad global.

Descargo de responsabilidad

Este documento ha sido elaborado por el Servicio de Investigación del Congreso (CRS). El CRS actúa como personal compartido y no partidista para las comisiones del Congreso y los miembros del Congreso. Opera exclusivamente a instancias y bajo la dirección del Congreso. La información contenida en un informe del CRS no debe utilizarse para fines distintos de la comprensión pública de la información que el CRS ha proporcionado a los miembros del Congreso en relación con la función institucional. Los informes del CRS, como obra del Gobierno de los Estados Unidos, no están sujetos a la protección de los derechos de autor en los Estados Unidos. Cualquier informe del CRS puede reproducirse y distribuirse en su totalidad, sin necesidad de permiso del CRS. Sin embargo, dado que un informe del CRS puede incluir imágenes o material protegido por derechos de autor de terceros, es posible que deba obtener el permiso del titular de los derechos de autor si desea copiar o utilizar de cualquier otra forma material protegido por derechos de autor.