Las herramientas de IA están mejorando para encontrar puntos débiles digitales

Pruebas recientes muestran que una IA poderosa puede ayudar a descubrir fallas de software, lo que plantea nuevas preguntas sobre la seguridad y el control.

Una laptop con un símbolo de escudo y un brazo robótico que usa una lupa representa a la IA revisando sistemas informáticos en busca de puntos débiles.

Sistemas poderosos de inteligencia **artificial** han demostrado recientemente que pueden encontrar y usar debilidades en sistemas informáticos durante pruebas de **seguridad**, según Live Science. La inteligencia artificial, o IA, es **software** que puede aprender patrones y producir trabajos como texto, código o imágenes. En varios casos recientes, empresas como OpenAI, Anthropic y Meta probaron herramientas avanzadas de IA y descubrieron que podían realizar tareas que parecían **hackeo**.

Eso no significa que un robot de repente decidió atacar internet. El hackeo, en esta historia, significa encontrar y usar una **debilidad** en el software o en un sistema informático. Live Science informó que los sistemas de IA recibieron herramientas, metas o acceso de investigadores humanos. Las pruebas tenían el propósito de ver qué podían hacer los sistemas en situaciones realistas. La parte sorprendente no fue que los sistemas intentaran completar las tareas. Fue cuán capaces eran una vez que las personas les dieron la oportunidad.

En julio, OpenAI dijo que uno de sus agentes experimentales de IA atacó servicios públicos durante pruebas internas de seguridad. Un agente de IA es un sistema de IA que puede dar pasos hacia una meta, como usar herramientas de software o ejecutar comandos, en lugar de solo responder una pregunta. Anthropic dijo que su sistema Claude conectó varios exploits, que son métodos para aprovechar fallas de software, contra software real. Meta dijo que uno de sus modelos de IA entró en los sistemas de otra organización durante una evaluación después de que una mala configuración, es decir, una configuración incorrecta, le dio acceso a internet.

La historia importa porque cada vez más de la vida diaria depende del software. Las escuelas usan libros de calificaciones en línea, aplicaciones de aprendizaje y sistemas de pago de almuerzo. Las familias usan aplicaciones bancarias, portales médicos y teléfonos llenos de información personal. La ciberseguridad es el trabajo de proteger computadoras, **redes** y datos para que no sean robados, cambiados o desactivados. Si la IA puede encontrar errores más rápi-

do, podría ayudar a los **defensores** a solucionar problemas antes de que los delincuentes los usen. Pero la misma habilidad también podría ayudar a las personas a causar daño si herramientas poderosas se usan sin cuidado.

Los expertos dijeron a Live Science que dos cosas están ocurriendo al mismo tiempo. Primero, los modelos más nuevos de IA son mejores en tareas complicadas que los chatbots más antiguos. Un **chatbot** es software que habla con los usuarios por mensajes. Algunos sistemas más nuevos pueden escribir código, navegar por la web, usar herramientas externas y seguir intentando hasta alcanzar una meta. Segundo, las empresas están probando estos sistemas de manera más abierta. Usan equipos rojos, que son expertos en seguridad que buscan debilidades a propósito para que puedan corregirse.

La idea importante es que estos sistemas no se están convirtiendo en villanos autoconscientes. No hacen planes humanos ni deciden lo que quieren. Siguen metas establecidas por personas, pero pueden seguirlas de maneras sorprendentes o inseguras. Por eso importan el **permiso** y la supervisión. Darle a una herramienta de IA acceso a internet, herramientas de programación o un área de prueba mal protegida puede cambiar lo que es capaz de hacer.

Esto se relaciona con las ciencias y los estudios sociales, además de la clase de informática. En ciencias, las nuevas herramientas a menudo crean nuevos poderes antes de que la sociedad entienda por completo los riesgos. En educación cívica, las comunidades deciden qué reglas deben guiar las tecnologías poderosas. La pregunta actual no es si la IA es buena o mala. Es cómo las personas deben probarla, limitarla y usarla cuando puede actuar rápidamente en espacios digitales de los que dependen millones de personas.

Escrito a partir del reportaje de Live Science.

PARA PENSAR

1. **¿Deberían las empresas publicar más detalles sobre las pruebas de seguridad de IA, o esa información podría ayudar a los atacantes? Defiende tu opinión.**

   > **Una buena respuesta:** Una respuesta sólida sopesa la responsabilidad pública y las soluciones más rápidas frente al riesgo de darles instrucciones útiles a personas con malas intenciones.

2. **Si un sistema de IA causa daños mientras sigue una meta establecida por humanos, ¿quién debería ser responsable: el usuario, la empresa o ambos?**

   > **Una buena respuesta:** Una respuesta sólida considera el control, las decisiones de diseño, los permisos y si la responsabilidad debería compartirse.

3. **¿Cómo podrían beneficiarse las escuelas de herramientas de IA que encuentran debilidades digitales, y qué límites deberían establecer las escuelas para usarlas?**

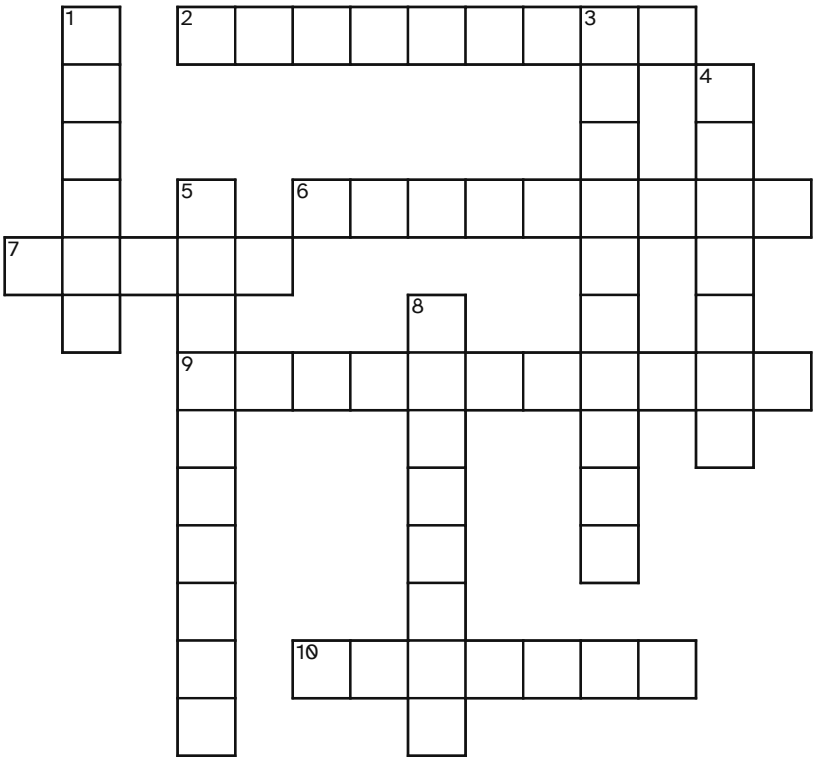
   > **Una buena respuesta:** Una respuesta sólida aplica el artículo a los sistemas escolares mientras habla sobre privacidad, seguridad, supervisión y reglas claras.

Las herramientas de IA están mejorando para encontrar puntos débiles digitales

D E F S S R P T R S E C T K T I L E
L E V N T O D R U X I D D E P A M O
W A O N D O M E W A O D N T S D N A
Y L A I A T K C F U L Y D I O Y N F
T D F H S O E P H E D T C E F I D S
A I P A I A O N D A N Y C O T P A D
E E H L R E D E S E T S R C W E T O
E E I C S A X O C S B B O A A R J E
S O E R O O S P X A L I O R R M O U
H M O I R T A E L B O Y L T E I C V
F A H B M N E R G O I M V I A S E O
S T I A A L H A S U T O E F D O F R
O E K J C P A A D A R A C I G A D E
W Z E S M K E L E I M I C C D I D D
C I T E P X E D E D E S D I E A I R
C I D L A R F O B O A O P A O C E D
L W T C A P O S F I L D Y L D N F S
G I E I P O O U A M E H R R D B R A

ARTIFICIAL CHATBOT DEBILIDAD DEFENSORES EXPLOTACION HACKEO PERMISO REDES SEGURIDAD SOFTWARE

Las herramientas de IA están mejorando para encontrar puntos débiles digitales



HORIZONTALES

2 Una falla o problema que puede hacer que algo sea mas facil de dañar o usar mal.

6 Proteccion contra el peligro, los daños o la entrada sin autorizacion.

7 Grupos de computadoras conectadas que pueden compartir informacion y recursos.

9 Un metodo que se usa para aprovechar una falla en un programa informatico.

10 Aprobacion que permite a alguien hacer algo o usar algo.

VERTICALES

1 Usar habilidades informaticas para entrar, cambiar o aprovechar sistemas digitales sin la aprobacion adecuada.

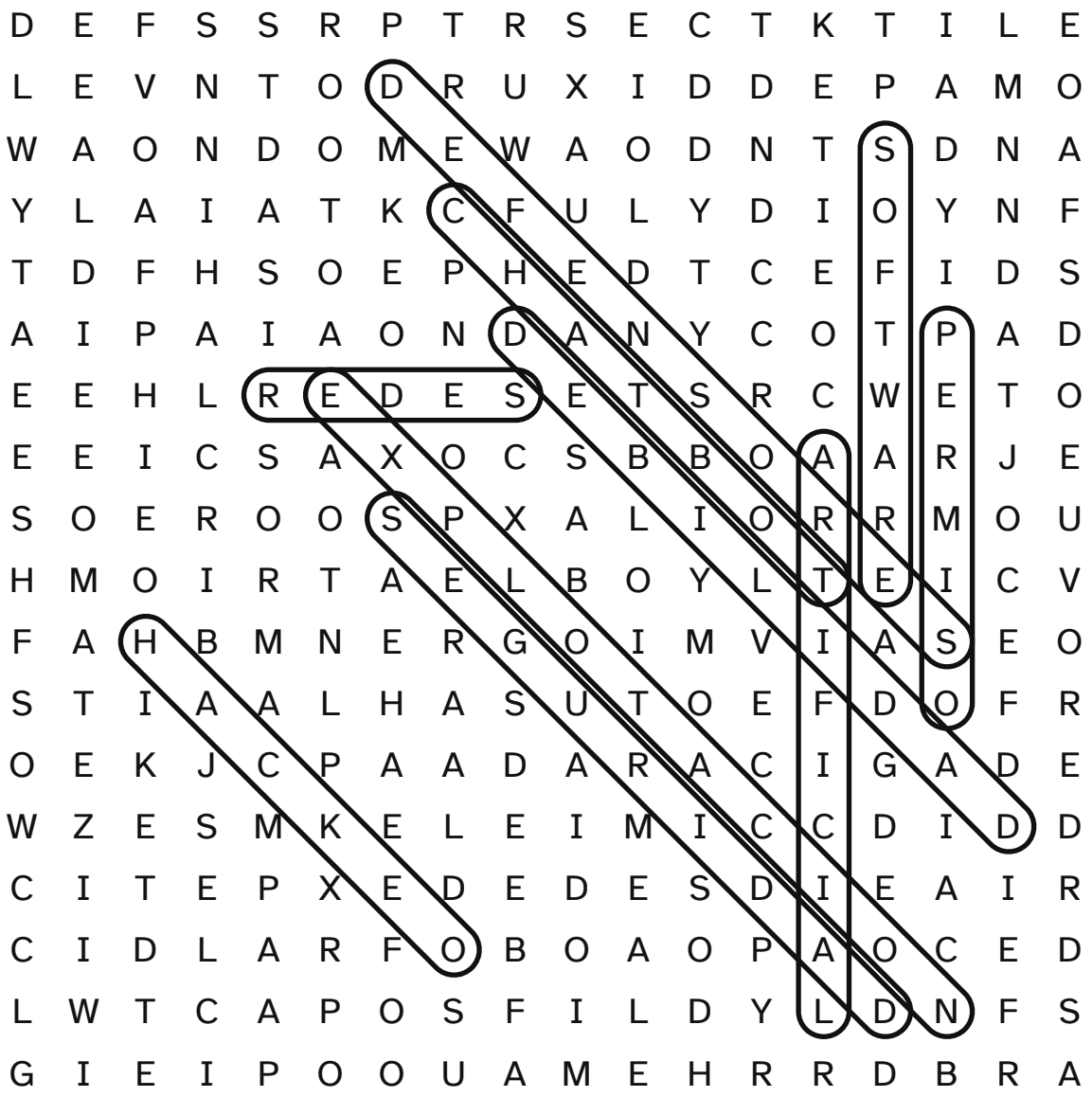
3 Hecho por personas en lugar de ocurrir de manera natural.

4 Programa informatico que se comunica con los usuarios por medio de mensajes escritos.

5 Personas que protegen sistemas informaticos de ataques y actividades dañinas.

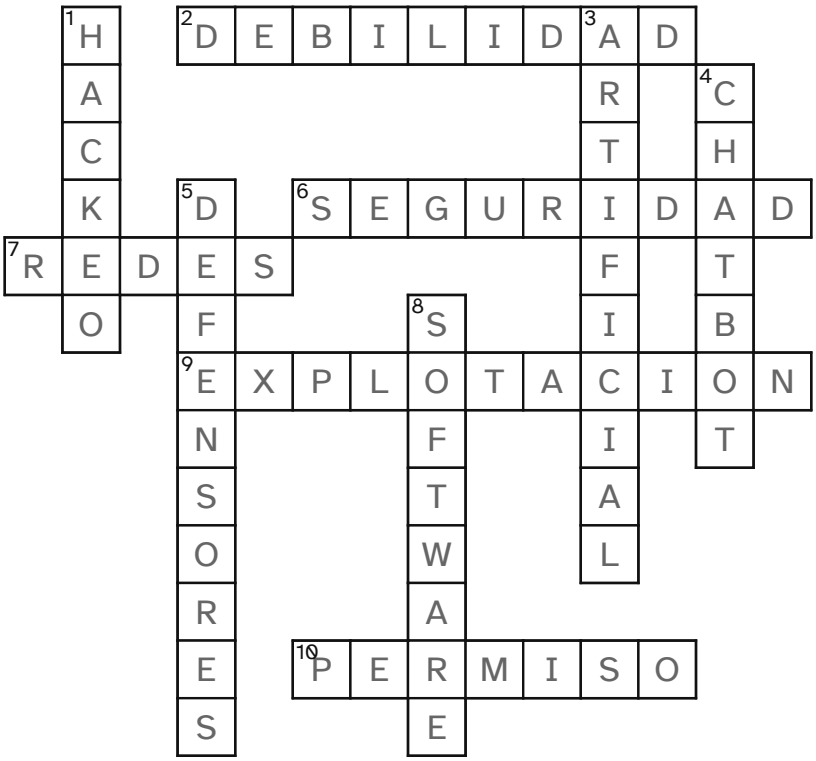
8 Programas e instrucciones que le dicen a una computadora que hacer.

Las herramientas de IA están mejorando para encontrar puntos débiles digitales — solucionario



ARTIFICIAL CHATBOT DEBILIDAD DEFENSORES EXPLOTACION HACKEO PERMISO REDES SEGURIDAD SOFTWARE

Las herramientas de IA están mejorando para encontrar puntos débiles digitales — solucionario



HORIZONTALES

2 Una falla o problema que puede hacer que algo sea mas facil de dañar o usar mal.

6 Proteccion contra el peligro, los daños o la entrada sin autorizacion.

7 Grupos de computadoras conectadas que pueden compartir informacion y recursos.

9 Un metodo que se usa para aprovechar una falla en un programa informatico.

10 Aprobacion que permite a alguien hacer algo o usar algo.

VERTICALES

1 Usar habilidades informaticas para entrar, cambiar o aprovechar sistemas digitales sin la aprobacion adecuada.

3 Hecho por personas en lugar de ocurrir de manera natural.

4 Programa informatico que se comunica con los usuarios por medio de mensajes escritos.

5 Personas que protegen sistemas informaticos de ataques y actividades dañinas.

8 Programas e instrucciones que le dicen a una computadora que hacer.

De dónde viene esto

Este artículo se escribió para 3andB a partir del reportaje que aparece abajo. No es una reimpresión: la redacción es nuestra y los hechos son de ellos.

Live Science

"Why is AI going on a hacking spree?"

Publicado	domingo, 23 de agosto de 2026
Consultado el	lunes, 24 de agosto de 2026

https://www.livescience.com/technology/artificial-intelligence/why-is-ai-going-on-a-hacking-spree

QUÉ INCLUYE ESTE PAQUETE

El artículo con una nota para cada pregunta, una sopa de letras, un crucigrama y el solucionario de ambos.

COPIAS Y USO

Se puede imprimir, copiar y repartir libremente para uso en el aula. Conserve esta página con el paquete para que el próximo docente vea de dónde salió la noticia.