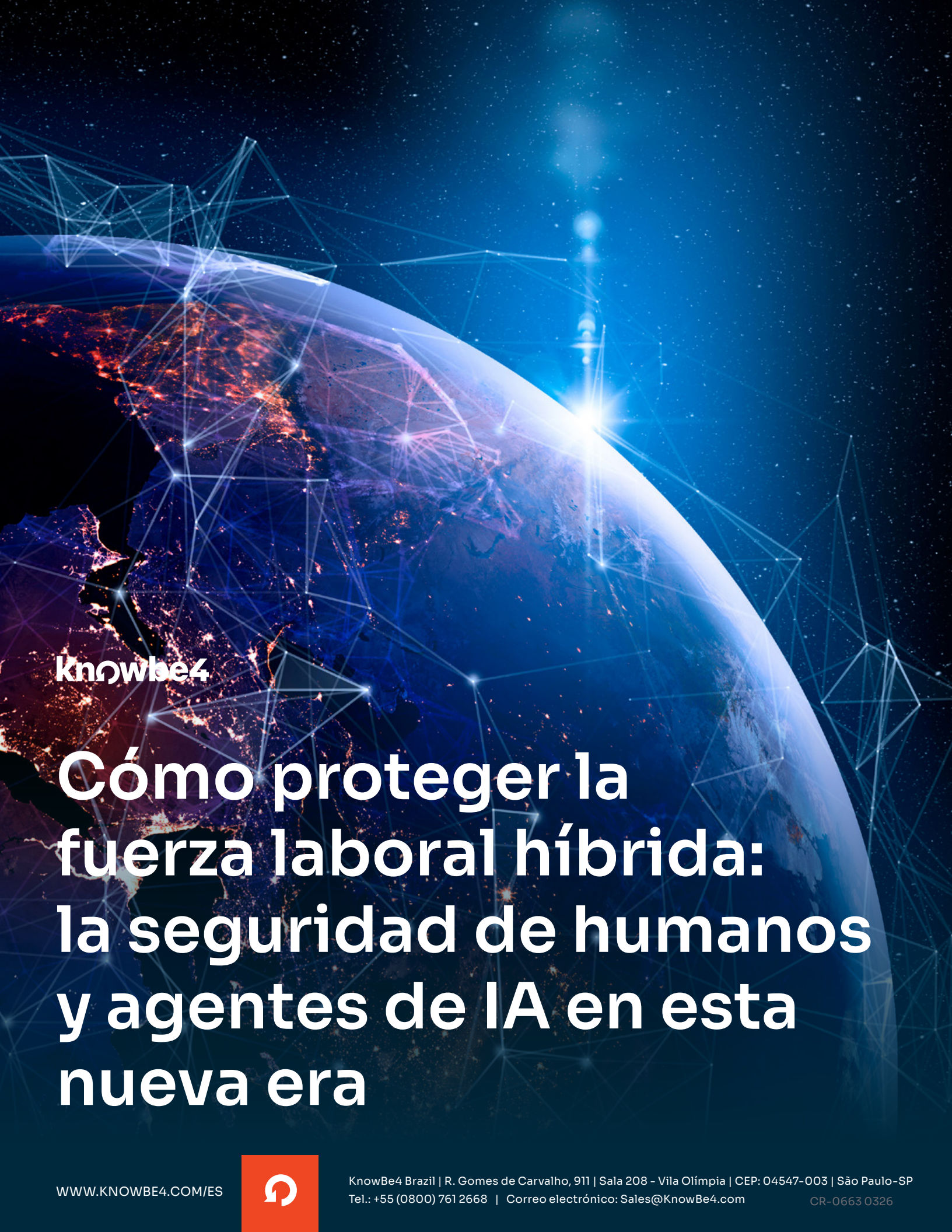


knowbe4

Cómo proteger la fuerza laboral híbrida: la seguridad de humanos y agentes de IA en esta nueva era

WWW.KNOWBE4.COM/ES

KnowBe4 Brazil | R. Gomes de Carvalho, 911 | Sala 208 - Vila Olímpia | CEP: 04547-003 | São Paulo-SP
Tel.: +55 (0800) 761 2668 | Correo electrónico: Sales@KnowBe4.com

CR-0663 0326

Introducción

La fuerza laboral cambió.

El personal ya no trabaja en soledad. Colabora cada vez más con asistentes, agentes autónomos y copilotos de IA que ayudan a escribir código, resumir incidentes, redactar comunicaciones, analizar datos y responder las preguntas de la clientela. Aquello que comenzó como una simple herramienta de productividad se convirtió rápidamente en una nueva clase de colegas digitales que se integran en todos los flujos de trabajo de las empresas.

Este cambio avanza a gran velocidad. Goldman Sachs estima que la IA agéntica podría representar alrededor del 60 % del valor de mercado del software para 2030, lo cual indica un cambio fundamental en cómo se realiza el trabajo. La fuerza laboral moderna ya no es puramente humana, sino que es un equipo híbrido de personas y agentes de IA que trabajan en conjunto para impulsar la productividad y la toma de decisiones.

Pero esta transformación introduce un nuevo reto de seguridad que muchas organizaciones no están preparadas para afrontar.

Durante décadas, los equipos de ciberseguridad se concentraron en proteger los sistemas, las redes y los terminales, mientras reducían también el riesgo humano mediante la concientización sobre seguridad, el desarrollo de la cultura y las capacitaciones. Estas labores se diseñaron para la defensa contra el phishing, la ingeniería social y otras formas de manipulación cognitiva.

Hoy en día, esa misma superficie de riesgos se ha expandido. Las personas ya no son las únicas que interactúan con los datos delicados, generan contenido o toman decisiones dentro de una empresa. Ahora, los agentes de IA realizan muchas de estas mismas tareas, a menudo con rapidez y confianza, pero sin comprender por completo las políticas, el contexto o la tolerancia al riesgo de la organización.

A diferencia del personal, los agentes de IA no se pueden mejorar a través de las capacitaciones o los programas de concientización tradicionales. Las organizaciones no pueden inspeccionar ni controlar por completo los *prompts* que les dan forma a las respuestas, y a pesar de eso, estos agentes ya están influyendo en los resultados comerciales.

Esto plantea un desafío conocido, pero con un nuevo enfoque. Los equipos de seguridad deben gestionar de qué manera interactúan las personas con la tecnología que utilizan. En la era de la IA, el riesgo radica cada vez más en la capa de interacción entre personas y sistemas inteligentes.

Resulta útil pensar en un agente de IA como un pasante muy competente pero influenciable. Procesa cantidades masivas de información y sigue las instrucciones con rapidez y seguridad. Demuestra interés por ayudar, pero le falta criterio, contexto y una comprensión inherente del riesgo de la organización.

Sin las protecciones adecuadas, hasta el asistente más competente puede hacer algo inesperado.

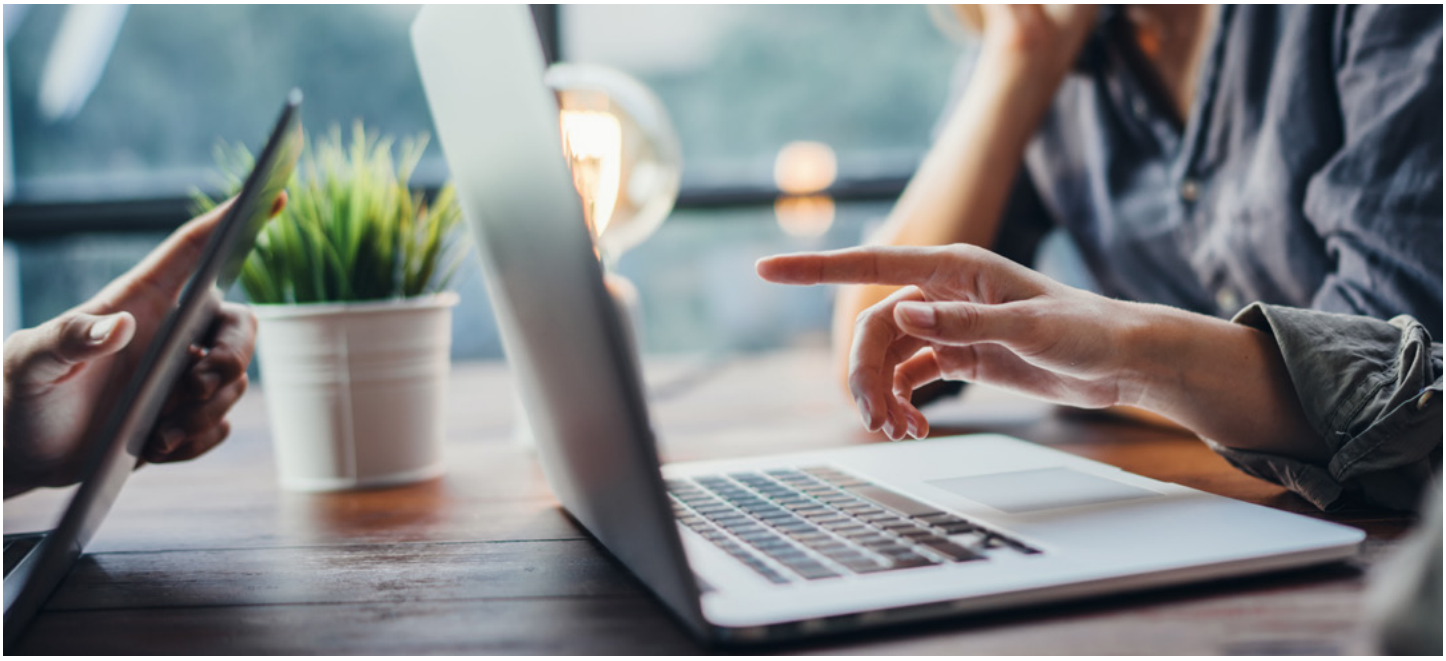
La superficie de ataque emergente entre personas e IA

Los agentes de IA son increíblemente poderosos, pero tienen una vulnerabilidad inherente. Las personas atacantes ya están aprendiendo a aprovecharse de la dinámica entre las personas y la IA. En muchos casos, estos ataques no apuntan solo a la persona o a la IA, sino que se aprovechan del vínculo de confianza que existe entre ellos.

La integración de la IA en los flujos de trabajo diarios ha generado una nueva categoría de riesgo para la seguridad. Las personas responsables de las amenazas utilizan la IA para ampliar los ataques conocidos al generar correos electrónicos de phishing más convincentes, crear contenido deepfake para ingeniería social y automatizar el reconocimiento. Al mismo tiempo, están desarrollando técnicas diseñadas específicamente para manipular los sistemas de IA:

- **Ataques de inyección de *prompts***
  Son entradas maliciosas, diseñadas para influir en el comportamiento de los agentes de IA, que tienen el potencial de lograr que evadan controles, divulguen datos delicados o realicen acciones no deseadas.
- **Suplantación de identidad del agente de IA**
  Hay agentes maliciosos que imitan herramientas legítimas de empresas para recopilar credenciales, información delicada o datos sobre flujos de trabajo de los miembros desprevenidos del personal.
- **Ingeniería social entre personas e IA**
  Se trata de ataques que se aprovechan de la confianza en las respuestas generadas por la IA, con el potencial de convertir a los agentes comprometidos en amenazas internas.

En muchos casos, no se utiliza malware ni se roban credenciales. El miembro del personal simplemente introduce una instrucción, y el agente de IA la ejecuta según su diseño. Tanto la persona como la IA se comportan de la manera esperada, lo cual es precisamente lo que da lugar a la vulnerabilidad.



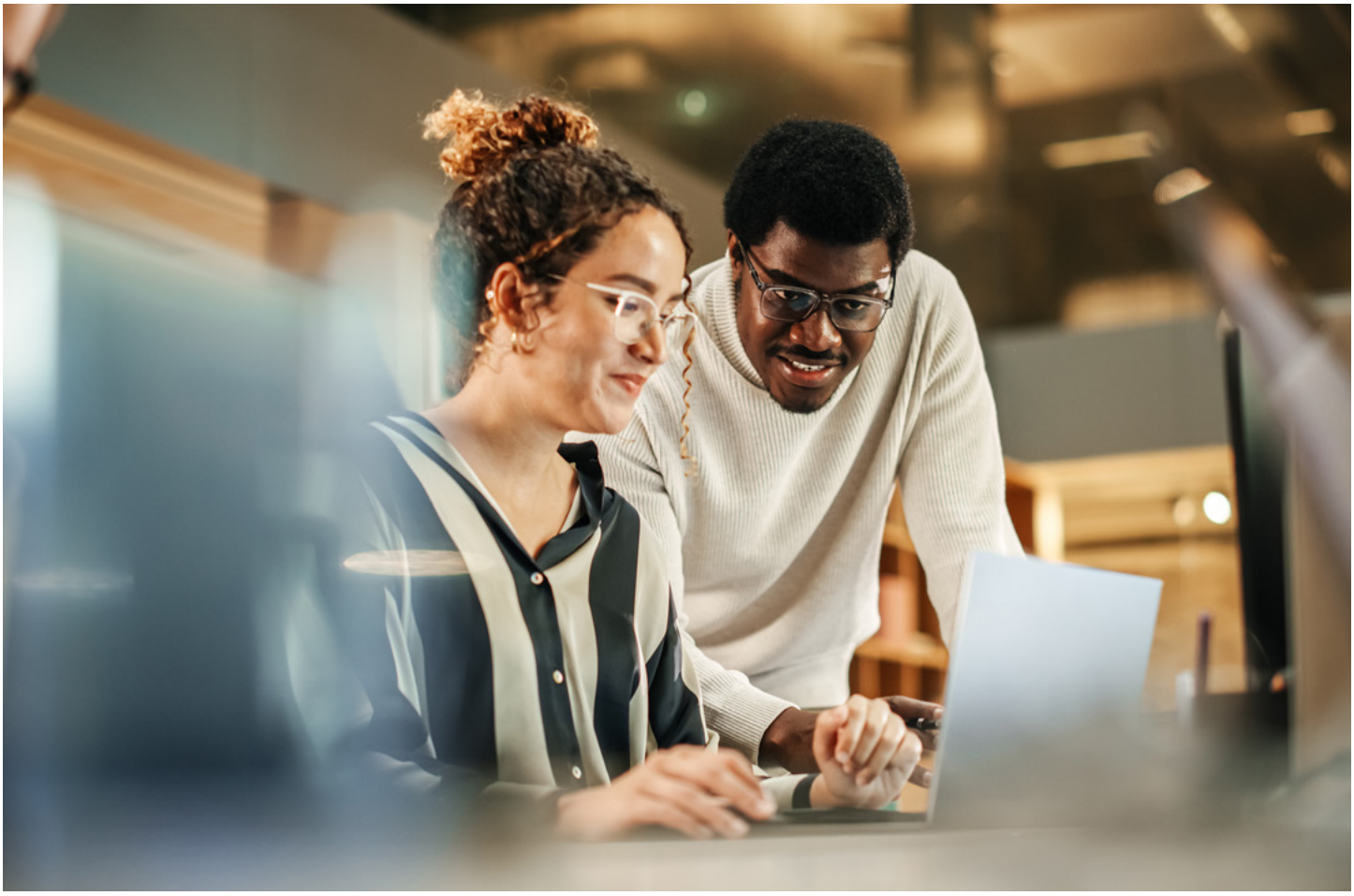
El problema de la confianza

Las personas confían por naturaleza en las herramientas que las ayudan de manera consistente a triunfar. Cuando los sistemas de IA generan resúmenes claros, recomendaciones útiles o contenido prolijo, los usuarios rápidamente comienzan a confiar en sus respuestas. Con el transcurso del tiempo, esto puede provocar un sesgo de automatización, es decir, la tendencia a confiar en las decisiones automatizadas incluso cuando estas podrían ser incorrectas o estar manipuladas.

Las personas atacantes pueden aprovecharse de esta confianza. Un *prompt* malicioso oculto dentro de un documento, conjunto de datos o sitio web puede influir en una respuesta generada con IA, que presentará la información manipulada en un tono confiado y contundente. Para el miembro del personal, la respuesta se ve creíble, aunque haya sido manipulada por alguien con intención de atacar.

Las arquitecturas tradicionales de seguridad se diseñaron para proteger a los sistemas, las redes y los terminales. Si bien estos controles siguen siendo esenciales, no abarcan por completo los riesgos que se generan en la colaboración entre las personas y la IA. La brecha más importante ahora está en la capa de la interacción, que es donde el personal depende de la IA para interpretar información, generar contenido y respaldar la toma de decisiones.

La protección de este entorno implica proteger a ambas partes de la fuerza laboral moderna. Las organizaciones deben capacitar al personal para que pueda reconocer los intentos de manipulación y usar las herramientas de IA de manera segura. Y al mismo tiempo deben garantizar que los agentes de IA sean resistentes ante la inyección de *prompts*, la exposición de datos y las acciones no autorizadas.



El futuro de la seguridad es la defensa doble

Los límites entre las personas y la IA dentro de la ciberseguridad están desapareciendo. A medida que los agentes de IA se incorporan en los flujos de trabajo diarios, las organizaciones que adapten sus estrategias de seguridad ante esta realidad serán las que logren permanecer firmes. El mensaje es claro: la ciberseguridad ya no se trata solo de proteger a los sistemas de las personas o a las personas de los sistemas. Se trata de proteger una interacción entre ambas partes, porque dentro de esa interacción se encuentra nuestra mayor vulnerabilidad, pero también nuestra mejor defensa.

Para enfrentar este desafío, las organizaciones deben adoptar una estrategia de defensa doble que refuerce tanto al elemento humano como a los mismos sistemas de IA.

1. Fortalecer la resiliencia humana

En un lugar de trabajo donde cada vez se usa más la IA, la función del personal está pasando de la ejecución de tareas a la supervisión y el buen criterio. Las capacitaciones en concientización sobre seguridad deben evolucionar de manera acorde. Ya no alcanza con enseñarles a los usuarios cómo detectar un correo electrónico de phishing. El personal debe desarrollar la consciencia digital: una combinación entre el escepticismo saludable, la sensibilidad contextual y la capacidad de evaluar de manera crítica las respuestas generadas con IA.

Esto comienza con la educación. El personal necesita comprender tanto las capacidades como las limitaciones de los agentes de IA, lo cual incluye de qué manera se pueden manipular. Debe aprender a reconocer el comportamiento anómalo o inesperado de la IA, como las respuestas que entran en conflicto con las políticas o el contexto. Es igual de importante que las organizaciones establezcan protocolos de verificación simples pero eficaces para las acciones de riesgo alto, en particular aquellas que inicie la IA o se realicen con su asistencia. Si algo parece ser urgente, inusual o fuera de alcance, el personal debe saber cómo tomarse un momento para validar la información y derivar el hecho.

2. Reforzar al agente de IA

Al mismo tiempo, los sistemas de IA se deben tratar como una nueva clase de activo empresarial que requiere control, gobernanza y supervisión. Esto comienza con la implementación de validaciones sólidas de entradas para impedir que los *prompts* maliciosos o los datos envenenados influyan en el comportamiento del agente. Las respuestas de la IA también se deben analizar continuamente para detectar anomalías, violaciones de las políticas o indicios de manipulación.

Definir límites claros para la función de los agentes de IA es igual de importante. Los sistemas deben estar diseñados para operar dentro de límites estrictos y rechazar aquellas solicitudes que se encuentren por fuera de las tareas autorizadas. Esto se debe reforzar mediante políticas bien definidas sobre el uso de la IA, que rijan de qué manera interactúan los agentes con los datos, sistemas y usuarios.

Capacidades críticas

La gestión del riesgo en este entorno híbrido requiere visibilidad hacia cómo operan los agentes de IA, cómo interactúan con los usuarios y de qué manera acceden a datos y sistemas. Para poder abarcarlo todo, desde el descubrimiento hasta la defensa, los productos eficaces de seguridad para agentes de IA se deben desarrollar a partir de cuatro pilares fundamentales:

Descubrimiento
Identificar y catalogar de manera automática a todos los agentes de IA que operan dentro de su entorno, lo cual incluye a los agentes introducidos mediante plataformas autorizadas y aquellos implementados sin una aprobación formal. El descubrimiento constante ayuda a eliminar los puntos ciegos y asegura que los equipos de seguridad comprendan el alcance completo del uso de la IA.

Monitoreo
Registrar todas las invocaciones de agentes, interacciones de *prompts* y llamadas a herramientas en una pista de auditoría completa con capacidad de búsqueda. La telemetría enriquecida, incluidos los metadatos de asistentes digitales de IA como Microsoft Copilot, les permite a los equipos comprender de qué manera se utilizan los agentes y dónde pueden emerger los riesgos potenciales.

Detección
Aplicar capacidades avanzadas de detección para identificar el comportamiento riesgoso en tiempo real. Esto incluye la identificación de los intentos de inyección de *prompts*, la escalada de privilegios, la exposición de datos delicados y la actividad anómala de los agentes en diferentes idiomas y casos de uso. El análisis del comportamiento ayuda a destacar las amenazas conocidas y emergentes que los controles tradicionales podrían no detectar.

Protección
Permitir a las organizaciones definir y aplicar su postura de seguridad deseada. Los equipos pueden elegir monitorear de manera pasiva las actividades, generar alertas o interceptar e impedir de manera activa las operaciones de riesgo alto en tiempo real. Una aplicación flexible permite que los equipos de seguridad equilibren la productividad con la protección a medida que evoluciona la adopción de la IA.

Por último, para profesionales de la seguridad y de TI, la escalabilidad es esencial. La gestión del riesgo en una creciente población de usuarios humanos y agentes de IA requiere de una visibilidad centralizada y eficacia operativa. Las capacidades clave deben incluir lo siguiente:

Visibilidad y auditoría
En los tableros centralizados se brinda una vista clara de las interacciones entre personas y agentes, y entre agentes y sistemas. Los registros detallados de auditoría permiten que los equipos de seguridad detecten el comportamiento anómalo, investiguen los incidentes e identifiquen las actividades que estén por fuera de las políticas definidas o los casos de uso previstos.

Aplicación dinámica de las políticas
Los equipos de seguridad pueden definir políticas basadas en riesgos mediante controles impulsados por API, lo cual permite un control consistente del uso de la IA en los diferentes entornos. Las políticas se pueden adaptar a medida que surgen nuevos casos de uso, a fin de garantizar que la adopción de la IA siga estando alineada con los requisitos de cumplimiento y seguridad de la organización.

Puntaje de riesgo integrado
Los datos de interacción alimentan el marco general de gestión de riesgos relacionados con el factor humano, y eso permite que las organizaciones evalúen el riesgo en el comportamiento de las personas y la IA. Al incorporar las interacciones entre las personas y los agentes en modelos de puntaje del riesgo, el liderazgo de seguridad comprende de manera más completa la exposición, y puede priorizar las tareas de mitigación con más eficacia.

Conclusión

Las personas y los agentes de IA ahora trabajan a la par. Las organizaciones necesitan un enfoque unificado para gestionar estos riesgos. Si extiende sus principios existentes de gestión de riesgos relacionados con el factor humano para incluir a los agentes de IA, los equipos de seguridad pueden incorporar innovaciones sin dejar de conservar el control, y así reducir la exposición sin impedir el progreso.

Más información

Prueba de seguridad contra el phishing (PST) gratuita

Descubra qué porcentaje de sus empleados tienen predisposición para ser víctimas de phishing (phish-prone) con la prueba de seguridad contra el phishing (PST) gratuita.

Automated Security Awareness Program gratuito

Cree un programa en concientización sobre seguridad personalizado para su organización.

Phish Alert Button gratuito

Ahora sus empleados tienen una forma segura de denunciar los ataques de phishing con un solo clic.

Comprobación de exposición del correo electrónico gratuita

Descubra antes que los malhechores cuál de las direcciones de correo electrónico de sus usuarios está expuesta.

Domain Spoof Test gratuita

Descubra si hackers pueden falsificar una dirección de correo electrónico de su propio dominio.

Acerca de KnowBe4

KnowBe4 capacita a la fuerza laboral actual para que pueda tomar decisiones de seguridad más acertadas todos los días. Con la confianza de más de 70 000 organizaciones en todo el mundo, KnowBe4 es pionero en seguridad digital para la fuerza laboral, protegiendo tanto a agentes de IA como a personas. La Plataforma KnowBe4 ofrece simulación de ataques y capacitación, seguridad de la colaboración y seguridad de los agentes impulsada por AIDA (Artificial Intelligence Defense Agents), además de un puntaje de riesgo propio. La plataforma aprovecha 15 años de datos de comportamiento para combatir amenazas avanzadas, como ingeniería social, inyección de *prompts* e IA oculta. Al proteger a personas y agentes, KnowBe4 lidera la industria en confianza y defensa de la fuerza laboral.

Obtenga más información knowbe4.com/es. Siga a KnowBe4 en LinkedIn y X.

Follow us on Linkedin

Follow us on X

knowbe4

KnowBe4 Brazil | R. Gomes de Carvalho, 911 | Sala 208 - Vila Olímpia | CEP: 04547-003 | São Paulo-SP
Tel.: +55 (0800) 761 2668 | www.KnowBe4.com/es | Sales@KnowBe4.com