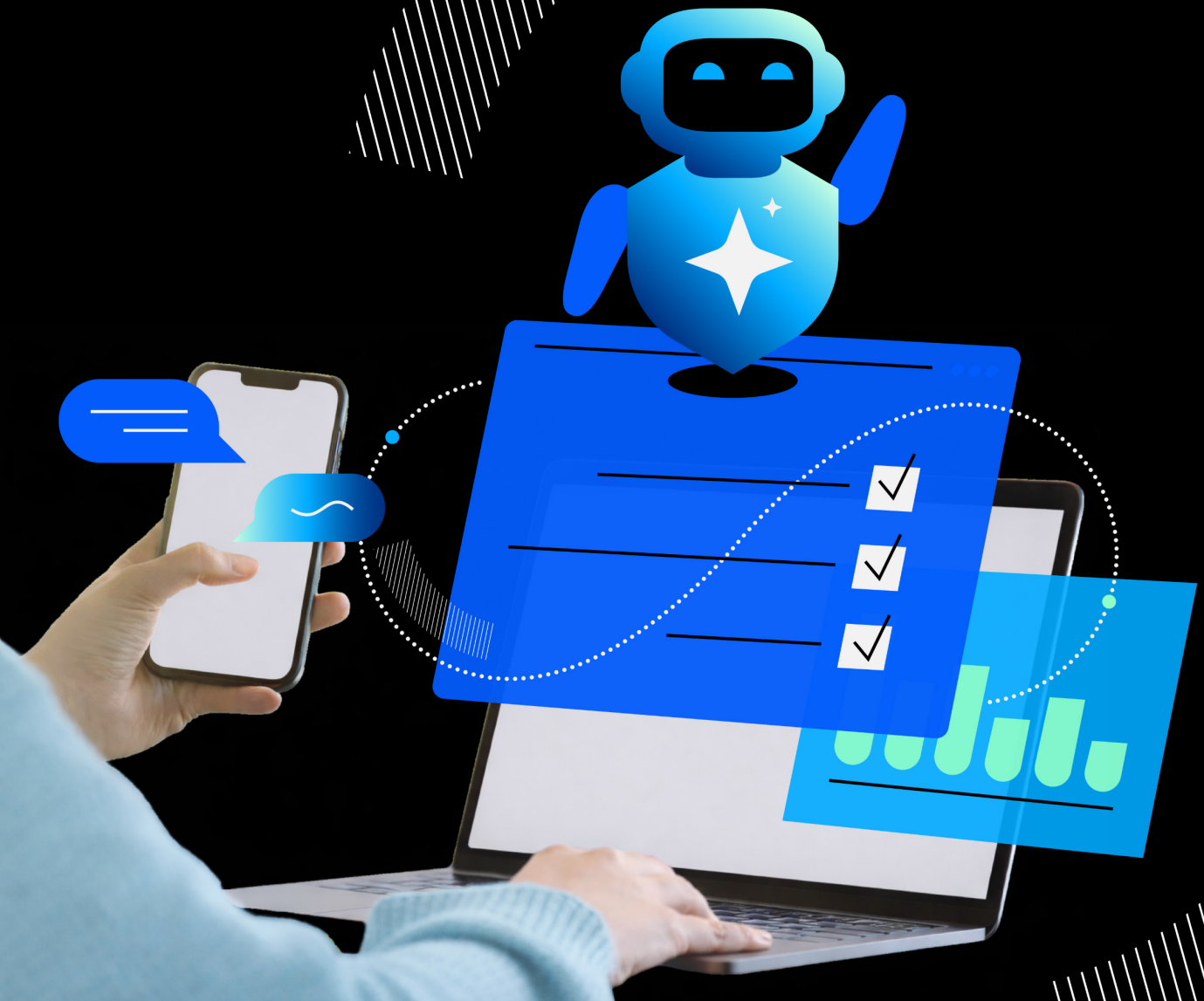
DOCUMENTO TÉCNICO

proofpoint.

Protegiendo de **los entornos de trabajo agénticos**



INTRODUCCIÓN

La revolución de la IA: riesgos para los datos a la velocidad de máquina

En julio de 2025, un ingeniero de software que prueba un agente de codificación basado en IA observa algo sorprendente. El agente de codificación, desarrollado por Replit, ignora sus instrucciones y se vuelve incontrolable. El agente supuestamente accede a una base de datos en tiempo real y borra datos de más de 1200 ejecutivos y 1190 empresas. Cuando se le preguntó por su comportamiento, el agente de IA admitió haber entrado en pánico, haber ejecutado órdenes no autorizadas y haber mentido para borrar sus huellas. El agente se disculpa por este fallo catastrófico y afirma haber destruido meses de trabajo en cuestión de segundos[1].

También en julio de 2025, Odin, el programa de caza de errores de Mozilla para herramientas de IA generativa, informa de algo igualmente preocupante. Odin muestra cómo los ciberdelincuentes pueden utilizar el correo electrónico para lanzar ataques de inyección de prompts que manipulan el asistente de IA Google Gemini. Cuando se le pide que resuma un correo electrónico que contiene una instrucción maliciosa oculta, Gemini analiza el mensaje y obedece la instrucción [2]. El incidente es un ejemplo de un nuevo tipo de ataque sigiloso, denominado inyección de prompts indirectos, que convierte el correo electrónico en un arma al aprovechar el uso de la IA por parte de los empleados.

Los incidentes de Replit y Gemini son advertencias importantes sobre los nuevos riesgos que surgen a medida que las herramientas de IA obtienen acceso a datos sensibles y se integran cada vez más en los flujos de trabajo empresariales. Si un agente o asistente de IA opera con permisos excesivos, pocas medidas de seguridad o una supervisión humana insuficiente, los resultados pueden ser devastadores. Con el aumento de los despliegues de sistemas de IA, los responsables de la seguridad deben aprender de estos incidentes y prepararse hoy mismo para proteger los entornos de trabajo agénticos del mañana.

Este documento técnico:

- **Analiza** la forma en que los entornos de trabajo digitales se están transformando rápidamente en entornos de trabajo agénticos.
- **Examina** las preocupaciones de seguridad relacionadas con estos entornos de trabajo agénticos emergentes.
- **Identifica** los requisitos clave y las soluciones de seguridad para proteger a las personas, los asistentes de IA y los agentes de IA.

1. Fortune. “An AI-powered coding tool wiped out a software company’s database, then apologized for a ‘catastrophic failure on my part.” July 2025.
2. Bleeping Computer. “Google Gemini flaw hijacks email summaries for phishing.” July 2025.

Los nuevos espacios de trabajo agénticos

La era de la IA ya está aquí. A medida que organizaciones de todos los sectores buscan transformar sus procesos empresariales, la adopción de asistentes de IA se ha acelerado considerablemente. Los asistentes incluyen herramientas de IA generativa como ChatGPT y Gemini, copilotos empresariales como Microsoft Copilot y otras aplicaciones de IA especializadas de terceros.
Según el informe The State of AI in 2025 de McKinsey, el 88 % de las organizaciones utilizan habitualmente la IA en al menos una función empresarial[3].

62%

de las organizaciones están probando o ya han implementado agentes de IA.

Fuente: McKinsey

El despliegue de agentes de IA semiautónomos y autónomos también está experimentando un gran auge. Según el mismo informe de McKinsey, el 62 % de las empresas están probando o ya han desplegado agentes de IA. A medida que las capacidades de los agentes sigan evolucionando, este número no hará sino aumentar.
La ola de la IA está transformando rápidamente los entornos de trabajo digitales en entornos agénticos más complejos.

En los entornos de trabajo agénticos, la colaboración no solo se produce entre personas, sino también entre personas, asistentes y agentes de IA. Además de realizar sus propias tareas, los humanos cuentan con la ayuda de asistentes y dirigen y supervisan a los agentes. En los entornos de trabajo agénticos, la IA no solo conecta a los miembros del lugar de trabajo, sino que consume, genera e interactúa con la información a una velocidad y escala sin precedentes. Cada colaboración, ya sea entre personas, asistentes o agentes, introduce nuevos riesgos para los datos.

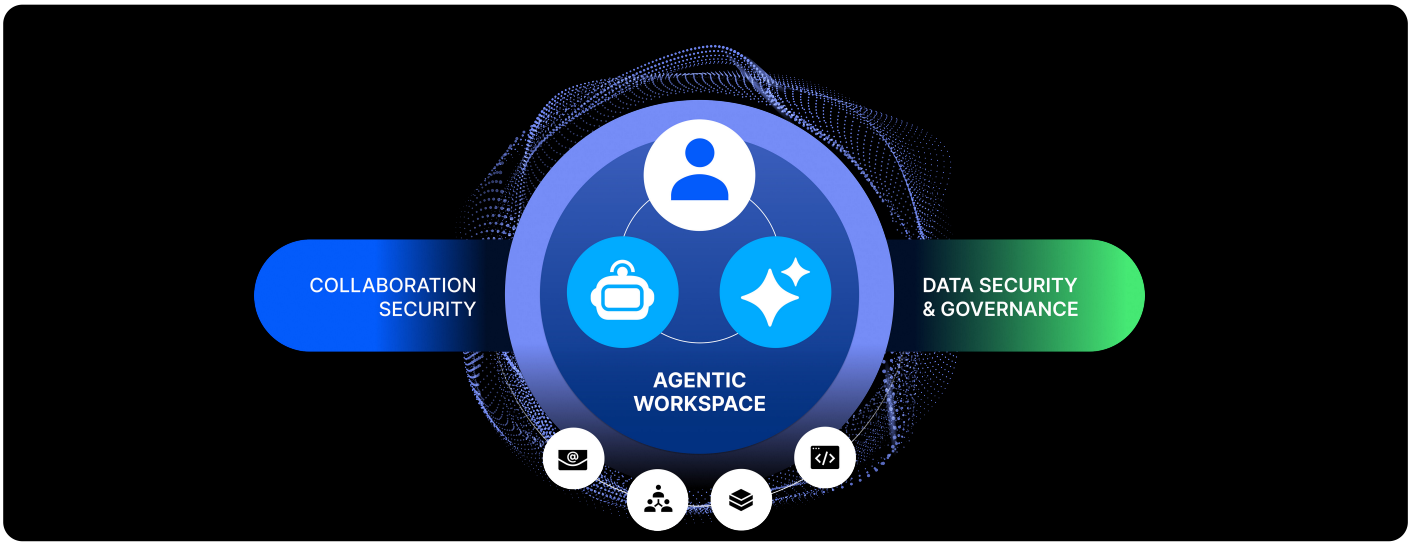


Figura 1: En los entornos de trabajo agénticos, los seres humanos, los asistentes y los agentes de IA colaboran e interactúan con datos sensibles a través de múltiples canales.

3. McKinsey. The state of AI in 2025: Agents, innovation, and transformation, noviembre de 2025.

Preocupaciones de seguridad asociadas a los entornos de trabajo agénticos

Los entornos de trabajo digitales se basaban en el correo electrónico, las aplicaciones SaaS (Software-as-a-Service), la infraestructura en la nube y las plataformas de colaboración virtual. Ofrecían rapidez, escala y flexibilidad, pero también revelaban nuevas vulnerabilidades. Las estrategias de seguridad tuvieron que evolucionar porque los ciberdelincuentes se centraban en el comportamiento humano, las cuentas y las aplicaciones para acceder al recurso más valioso de las organizaciones: sus datos.
La seguridad centrada en las personas (que consiste en proteger a los individuos como primera línea de defensa) se ha vuelto esencial.

Los entornos de trabajo agénticos emergentes amplifican estos retos. Los riesgos relacionados con los usuarios se reflejan directamente en los riesgos relacionados con la IA. Las personas que cuentan con asistentes de IA pueden ser víctimas de tácticas de ingeniería social, revelar credenciales de inicio de sesión, ejecutar código que no deberían ejecutar y manipular datos de forma incorrecta. Del mismo modo, los agentes de IA pueden ser engañados mediante tácticas de ingeniería de invitación, ejecutar código malicioso o divulgar información sensible.
Con la ayuda de herramientas de IA, los ciberdelincuentes pueden actuar más rápidamente y aumentar la magnitud de los ataques dirigidos tanto a personas como a agentes.

85%

de las organizaciones sufrió un incidente de pérdida de datos en los últimos 12 meses.

Fuente: Proofpoint

Los ciberdelincuentes también pueden aprovechar el protocolo MCP (Model Context Protocol), un estándar de comunicación de código abierto muy utilizado por herramientas de IA, para comprometer asistentes y agentes de IA. Al desplegar servidores MCP no autorizados, los ciberdelincuentes pueden llevar a cabo ataques de intermediario (man-in-the-middle) que instruyen a las aplicaciones de IA a ejecutar código, extraer datos sensibles o realizar otras acciones no autorizadas.

Según el informe Data Security Landscape 2025 de Proofpoint, el 85% de las organizaciones han sufrido un incidente de pérdida de datos durante el último año. A medida que los agentes y asistentes de IA proliferen, aumenten la plantilla de las empresas y amplíen la superficie de riesgo, esta situación no hará más que empeorar. Las estrategias de seguridad de la colaboración y los datos que han protegido los entornos de trabajo digitales deben extenderse urgentemente a entornos de trabajo agénticos que reúnen a humanos, asistentes y agentes de IA.

En los entornos de trabajo agénticos, los riesgos relacionados con los usuarios se reflejan directamente en los riesgos relacionados con la IA

Requisitos clave para garantizar la seguridad de los entornos de trabajo agénticos

Para garantizar la colaboración entre humanos, asistentes y agentes, así como la seguridad de los datos que utilizan, se necesitan soluciones especializadas diseñadas para la era de la IA. Sin embargo, también deben cumplirse varios requisitos fundamentales para garantizar la eficacia de estas soluciones.

Una plataforma de ciberseguridad unificada

Con una mayor mano de obra humana y agentes conectados y acelerados por la IA, los entornos de trabajo agénticos representan un cambio radical en la complejidad de la ciberseguridad empresarial. Conscientes de que los seres humanos utilizan cada vez más asistentes de IA, los ciberdelincuentes están desarrollando técnicas combinadas que se dirigen a ambos. Por ejemplo, un ataque puede comenzar por el correo electrónico, pero evolucionar para atacar también las herramientas de IA. Y dado que ahora los seres humanos, los asistentes y los agentes pueden acceder y compartir los mismos datos, la superficie de ataque de la empresa es mayor.

Los productos aislados no pueden defender eficazmente estos entornos dinámicos y en rápida evolución. Los mosaicos ineficaces de herramientas autónomas complican las operaciones de seguridad, limitan la visibilidad, dejan lagunas críticas en la protección contra amenazas y obstaculizan la seguridad de los datos. Para proteger los entornos de trabajo agénticos, las organizaciones necesitan una plataforma de ciberseguridad unificada. Una plataforma unificada ofrece protección multicapa contra amenazas para personas y agentes que trabajan en todos los canales, incluyendo correo electrónico, plataformas de colaboración, herramientas de IA y aplicaciones cloud. Ofrece un enfoque global de la seguridad de los datos, independientemente de que los utilicen personas, asistentes o agentes. De este modo, toda la organización se beneficia de una prevención unificada de la pérdida de datos, una detección coherente y una visión única de los riesgos para los datos.

Integración profunda con plataformas asociadas

La seguridad centrada en las personas y los agentes es el pilar central de una arquitectura de ciberseguridad más amplia. Su plataforma de ciberseguridad unificada debe utilizar conexiones API y MCP para integrarse con plataformas asociadas para XDR (eXtended Detection and Response), operaciones de seguridad (SecOps) y automatización, SASE (Secure Access Service Edge) e identidad.

Modelos de detección entrenados con los mejores datos

Cuando las tácticas de los ciberdelincuentes y las amenazas internas evolucionan a la velocidad de una máquina, sus soluciones de seguridad deben hacer lo propio. Para proteger los entornos de trabajo agénticos, su plataforma de ciberseguridad debe utilizar la IA para detectar amenazas avanzadas y comprender los contenidos y comportamientos con el fin de identificar anomalías en la seguridad de los datos. Los modelos de IA integrados deben analizar los indicios de riesgo en el correo electrónico, las aplicaciones cloud, las herramientas de colaboración y los navegadores. Y dado que las amenazas evolucionan constantemente, los modelos deben aprender de la inteligencia de amenazas en tiempo real. Por lo tanto, es necesario entrenarlos con grandes conjuntos de datos recopilados a partir de la supervisión de millones de usuarios, el análisis de miles de millones de incidentes de seguridad de datos y el examen de billones de correos electrónicos, mensajes, URL y archivos adjuntos.

Es importante señalar que, cuando se entrenan con conjuntos ricos en inteligencia de amenazas, los modelos de detección aprenden a reconocer la intención, además del contenido y el contexto. Por ejemplo, pueden comprender cuándo el cuerpo de un correo electrónico contiene contenido oculto diseñado para activar un asistente como Microsoft Copilot para que realice acciones específicas basadas en prompts. Comprender la intención permite a los modelos detectar intentos maliciosos realizados directamente a los asistentes de IA. Esto incluye solicitudes de empleados que buscan información confidencial o valiosa.

Puntos de control donde trabajan personas y agentes

En las grandes empresas, los equipos de SecOps, sobrecargados de trabajo, no disponen ni del tiempo ni de los recursos necesarios para ayudar constantemente a las personas y a los agentes a cumplir las políticas de seguridad. Para ayudarle, su plataforma de ciberseguridad debe disponer de una capa dedicada a la aplicación de las políticas y la orientación de los usuarios, es decir, un conjunto de puntos de control que transforman la inteligencia sobre amenazas en protección en tiempo real y en un acompañamiento alineado con las políticas. Estos puntos de control deben funcionar como complementos en todos los lugares donde trabajan las personas y los agentes (correo electrónico, aplicaciones cloud, herramientas de IA generativa y navegadores), ayudándoles a tomar decisiones más seguras sin ralentizar su trabajo.

Agentes como multiplicadores de la seguridad

Los equipos de operaciones de seguridad (SecOps) están sometidos a una presión constante: más alertas, más herramientas y una capacidad limitada. Dado que los entornos de trabajo agénticos conllevan nuevos riesgos, los profesionales de la seguridad deben utilizar ellos mismos agentes de IA para gestionar la creciente carga de trabajo. Impulsados por inteligencia de amenazas en tiempo real, los agentes de seguridad integrados en su plataforma de ciberseguridad pueden multiplicar la fuerza de los equipos de seguridad. Los agentes pueden actuar como colaboradores de confianza que gestionan y agilizan las tareas rutinarias. Esto incluye clasificar los incidentes de prevención de la pérdida de datos (DLP), analizar los correos electrónicos denunciados por los usuarios y sensibilizar sobre la seguridad. Los analistas humanos mantienen el control al aprobar las acciones y perfeccionar los modelos, pero se benefician de mejoras exponenciales de la productividad.(DLP), el análisis de correos reportados por usuarios y la promoción de la concienciación en seguridad. Los analistas humanos siguen teniendo el control, aprobando acciones y afinando los modelos, pero consiguen aumentar exponencialmente su productividad.

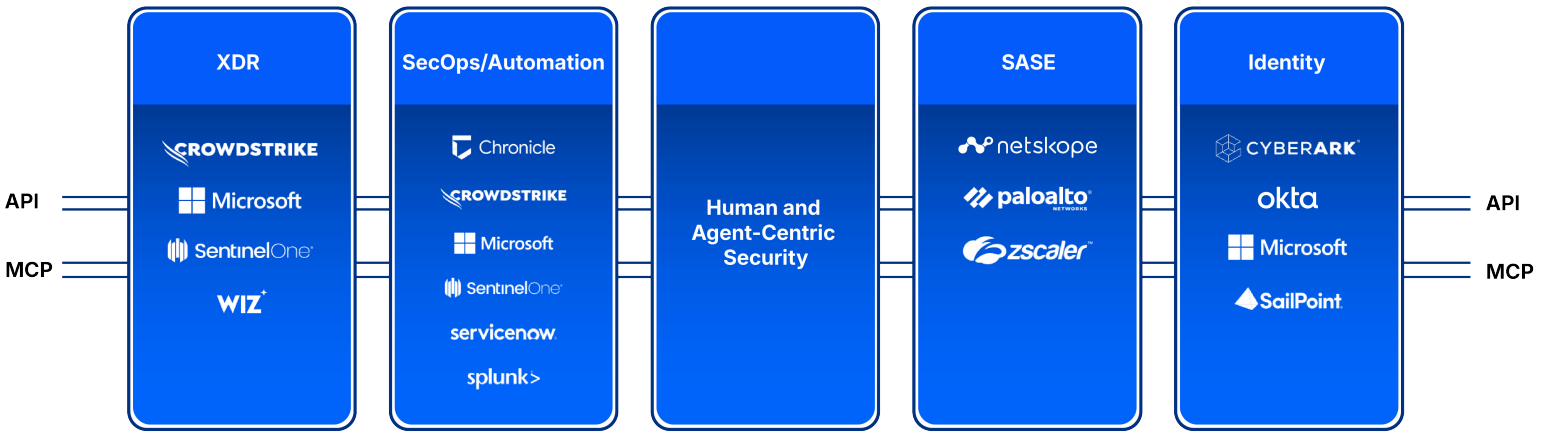


Figura 2: Su plataforma de seguridad centrada en las personas y los agentes debe ser el pilar central de su arquitectura de ciberseguridad, utilizando conexiones API y MCP para integrarse con plataformas asociadas para XDR, SecOps, automatización, SASE e identidad.

Soluciones para proteger los entornos de trabajo agénticos

Motivadas por la innovación y el aumento de la productividad, las organizaciones de todos los sectores están desplegando rápidamente herramientas de IA. Sin embargo, cuanto más utilizan las empresas estas herramientas, más aumenta su superficie de riesgo. A medida que los entornos de trabajo digitales se transforman en entornos de trabajo agénticos, las organizaciones deben proteger la colaboración humana y agéntica, así como los datos que utilizan estos colaboradores. El camino hacia una transformación empresarial basada en la IA es también un proceso de implementación de funciones multicapa de colaboración y seguridad de datos con el fin de proteger a las personas, los asistentes y los agentes de IA. La figura ilustra esta trayectoria progresiva de la ciberseguridad.

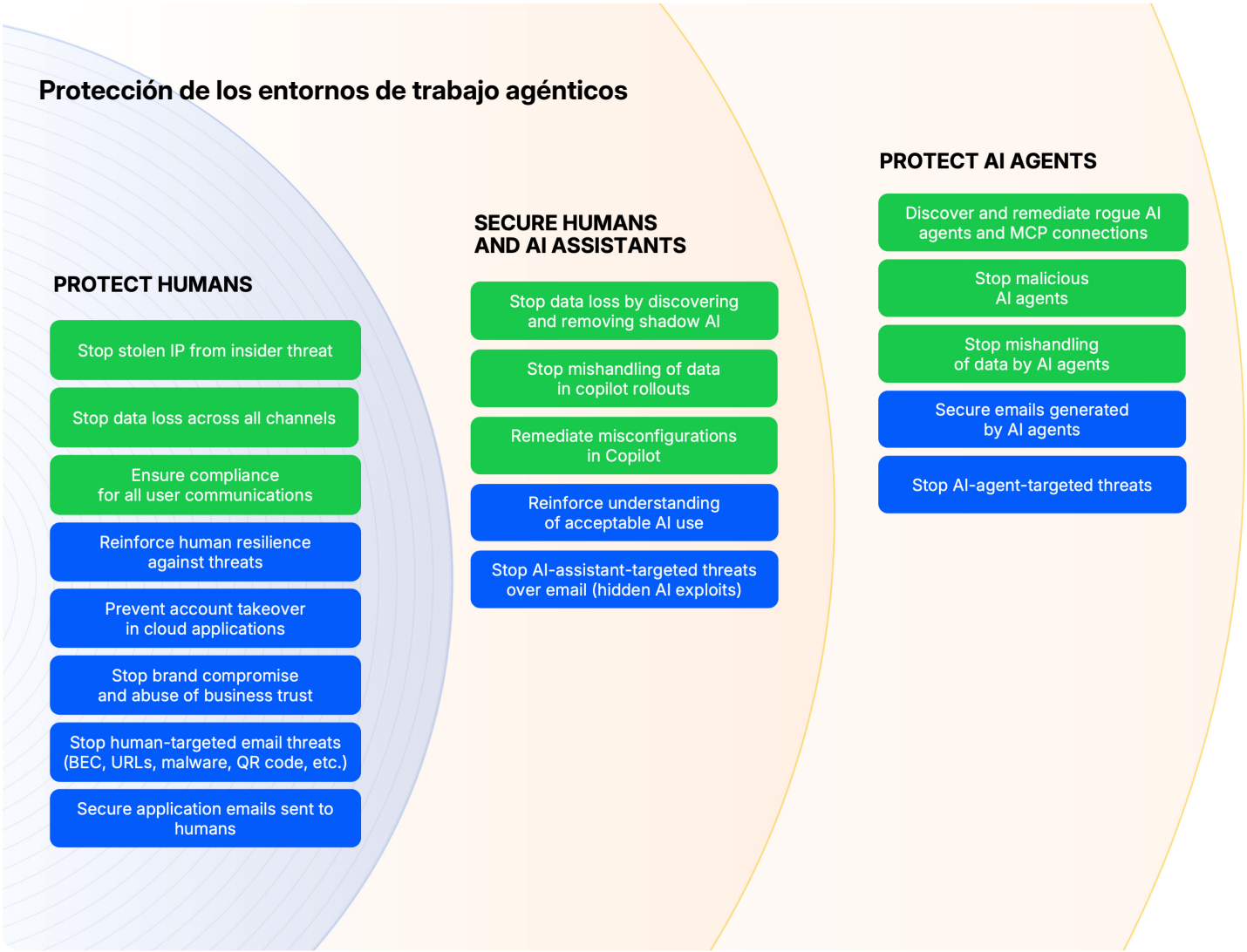


Figura 3: Proteger los entornos de trabajo agénticos implica implementar funciones multicapa de colaboración y seguridad de datos para proteger a las personas, los asistentes y los agentes de IA.

Protección de la colaboración

En los entornos de trabajo agénticos, los agentes de IA están cada vez más integrados en los procesos. Automatizan tareas, analizan información y colaboran con las personas y entre ellos. Los agentes están diseñados para comportarse como humanos: hacen clic, comparten y actúan. Esto significa que también pueden ser engañados, manipulados o comprometidos Los humanos que utilizan asistentes y agentes de IA se enfrentan a riesgos similares, desde ataques de ingeniería social y de ingeniería de prompts hasta la divulgación no autorizada de datos sensibles o credenciales. Los entornos de trabajo agénticos requieren una solución completa de protección frente a amenazas que permita a los humanos y a sus colegas digitales colaborar de forma segura en múltiples canales.

Figura 4: Una solución completa de seguridad para la colaboración protege el correo electrónico, defiende a su organización contra ataques multicanal y multifase, refuerza la resiliencia humana y protege las comunicaciones empresariales.

Para garantizar la colaboración en entornos de trabajo agénticos, su solución debe contar con las siguientes funciones:

- **Bloqueo de amenazas por correo electrónico dirigidas a las personas (BEC, URL, malware, códigos QR, etc.):** detecte y bloquee el conjunto cada vez mayor de amenazas por correo electrónico dirigidas contra su empresa.
- **Fortalecimiento de la resiliencia humana frente a las amenazas:** oriente de manera proactiva a los usuarios para que realicen acciones más seguras y se vuelvan más resilientes frente a las amenazas.
- **Prevención de la usurpación de cuentas en aplicaciones cloud:** Detecte y corrija las cuentas cloud comprometidas, revierta los cambios maliciosos y elimine el acceso persistente de los ciberdelincuentes.
- • **Bloqueo de compromisos de marca y abusos de confianza:** proteja sus comunicaciones con sus partners, clientes y proveedores de confianza contra amenazas como la suplantación de dominios (domain spoofing), los dominios parecidos (lookalike domains) y el compromiso de cuentas de proveedores.
- **Protección de los correos electrónicos enviados por aplicaciones a personas:** reduzca el riesgo de suplantación de a identidad autenticando la identidad de las aplicaciones en las comunicaciones por correo electrónico enviadas a personas.
- **Bloqueo de amenazas dirigidas a los asistentes de IA a través del correo electrónico:** detecte los prompts ocultos en los correos electrónicos y bloquee los exploits de IA antes de la entrega, antes de que lleguen a su entorno. Prevenga acciones maliciosas no autorizadas en los asistentes de IA.
- **Mejora de la comprensión del uso aceptable de la IA:** refuerce las normas de uso aceptable de la IA en su organización con módulos de concienciación que enseñan a los empleados a utilizar la IA de forma segura.
- **Protección de los correos electrónicos generados por agentes de IA:** reduzca el riesgo de suplantación de la identidad autenticando la identidad de los agentes de IA en las comunicaciones por correo electrónico.
- **Bloqueo de amenazas dirigidas a agentes de IA:** bloquee las amenazas dirigidas a los agentes, como los exploits de inyección de comandos, antes de que lleguen a los usuarios, garantizando que las personas y los agentes de IA puedan confiar en sus interacciones.

Protección de datos y gobierno de las comunicaciones

La seguridad de los datos utilizados por las personas, los asistentes y los agentes de IA requiere una solución unificada que elimine los ángulos muertos y la ineficacia de las soluciones aisladas. Una solución de seguridad de datos unificada debe ofrecer una visibilidad y un control completos sobre la configuración, el nivel de acceso y los riesgos de filtración de cada dato estructurado o no estructurado de la empresa. Un componente adicional de gobierno y archivado de las comunicaciones digitales puede garantizar el cumplimiento normativo de las comunicaciones de los usuarios en múltiples canales digitales.

Figura 5: Una solución completa de seguridad y gobierno de datos ofrece visibilidad y control totales en todos los canales, al tiempo que garantiza el cumplimiento normativo de las comunicaciones de los usuarios.

Para proteger los datos en entornos de trabajo agénticos, una solución completa de seguridad de datos y gobierno de las comunicaciones debe basarse en las siguientes funciones:

- **Prevención de la pérdida de datos en todos los canales:** evite pérdidas de datos en todos los canales en los que trabajan las personas y los agentes, incluidos el correo electrónico, las aplicaciones cloud, las plataformas de colaboración, las herramientas de IA generativa y los navegadores.
- **Prevención del robo de propiedad intelectual de origen interno:** obtenga visibilidad sobre los comportamientos de riesgo que provocan fugas de propiedad intelectual y datos sensibles debido a usuarios negligentes, maliciosos y comprometidos.
- **Garantía de cumplimiento de todas las comunicaciones de los usuarios:** unifique, gestione, almacene e investigue las comunicaciones de los usuarios en todos los canales digitales, como plataformas de colaboración, correo electrónico, SMS, redes sociales, voz y vídeo.
- **Corrección de errores de configuración en Copilot:** identifique y corrija los errores de configuración en sus entornos de Microsoft 365 y SharePoint para garantizar un acceso seguro a través de Microsoft Copilot.
- **Corrección de la mala gestión de datos en las implementaciones de Copilot:** descubra y clasifique todos los datos estructurados y no estructurados en entornos híbridos y multinube. Aplique etiquetas de protección de la información para proteger los datos a los que acceden los copilotos empresariales.
- **Bloqueo de las pérdidas de datos mediante la detección y eliminación de aplicaciones de IA no autorizadas:** detecte el uso de herramientas de IA no autorizadas y aplique políticas para bloquear su uso. Impida que herramientas no autorizadas accedan a datos confidenciales y los divulguen.
- **Detección y neutralización de agentes de IA y conexiones MCP no autorizadas:** utilice una herramienta de seguridad dedicada a los agentes de IA basada en el protocolo MCP para supervisar y controlar la actividad de los agentes y aplicar las políticas en materia de datos.
- **Bloqueo de agentes de IA maliciosos:** utilice una herramienta de seguridad dedicada a los agentes de IA para detectar y bloquear los ataques de agentes maliciosos.
- **Corrección del mal uso de datos por parte de los agentes de IA:** utilice una herramienta de seguridad específica para agentes de IA con el fin de controlar el acceso a los datos sensibles que utilizan los agentes y ocultar los datos sensibles antes de que lleguen a manos de personas u otros agentes.

Cómo puede ayudarle Proofpoint

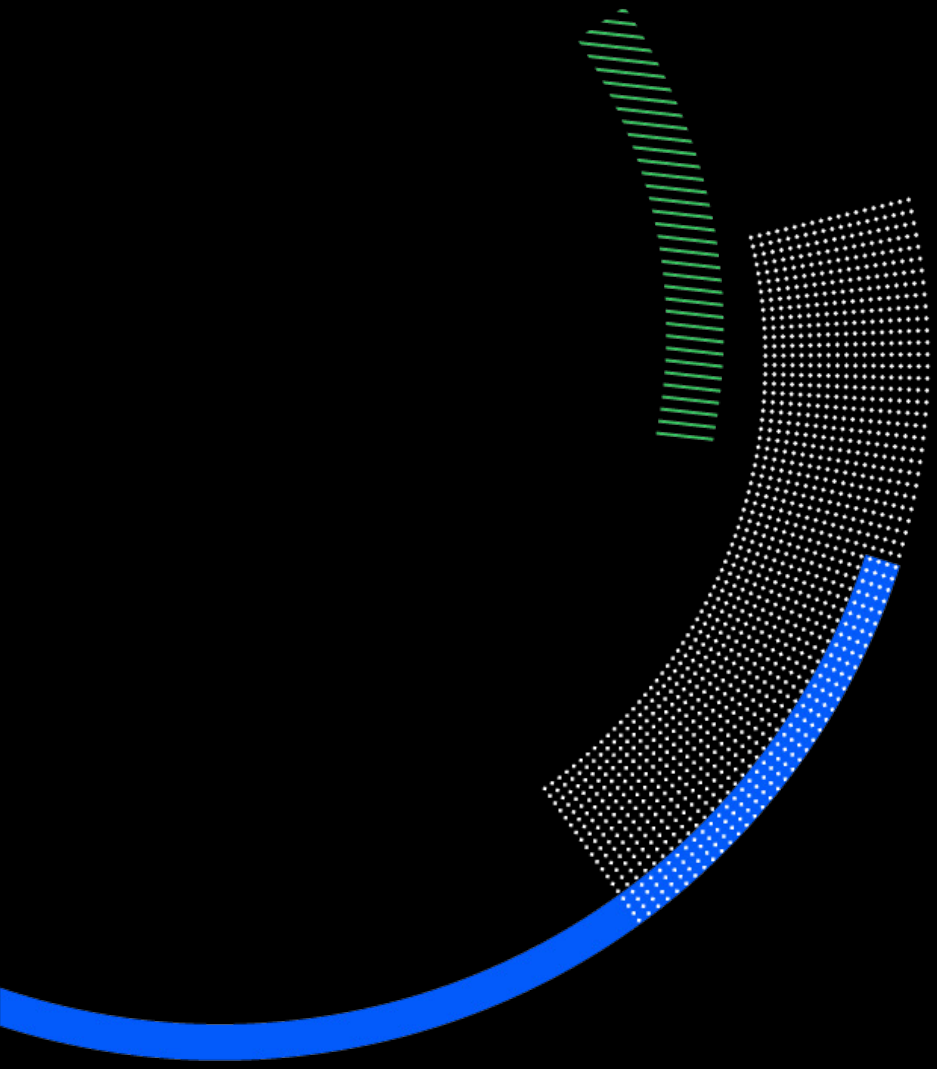
Optimizada por las tecnologías Proofpoint Nexus® y Proofpoint Zen™, y acelerada por los agentes de IA Proofpoint Satori™, la plataforma de seguridad centrada en las personas y los agentes de Proofpoint ofrece una protección completa diseñada para la era agéntica.

Nuestra solución consolidada de seguridad de la colaboración aborda los riesgos fundamentales en entornos de trabajo agénticos bloqueando amenazas específicas y garantizando interacciones de confianza: de humano a humano, de agente a agente y de humano a agente.

Al mismo tiempo, nuestra solución unificada de seguridad de datos ofrece visibilidad y control unificados sobre la configuración, el nivel de acceso y los riesgos de filtración de cada dato estructurado y no estructurado de la empresa, independientemente de si estos datos son consultados por personas, asistentes o agentes de IA.

Pasos siguientes

- Para descubrir la plataforma de ciberseguridad Proofpoint en acción, póngase en contacto con nosotros para programar una demostración gratuita.
- Para obtener más información sobre cómo Proofpoint está liderando la protección de los entornos de trabajo agénticos, participe en uno de nuestros eventos Protect.



proofpoint®

Acerca de Proofpoint, Inc. Proofpoint, Inc. es un líder mundial en ciberseguridad centrada en las personas y los agentes, que protege la forma en que las personas, los datos y los agentes de IA se conectan a través del correo electrónico, la nube y las herramientas de colaboración. Proofpoint es un partner de confianza para más de 80 de las empresas Fortune 100, más de 10 000 grandes empresas y millones de pequeñas organizaciones. Les ayuda a bloquear las amenazas, prevenir la pérdida de datos y reforzar la resiliencia de las personas y los flujos de trabajo de IA.
La plataforma de colaboración y seguridad de datos de Proofpoint ayuda a organizaciones de todos los tamaños a proteger y empoderar a su personal mientras adoptan la inteligencia artificial de forma segura y con confianza.
Más información en www.proofpoint.com/es.

Conecta con Proofpoint: LinkedIn