PROTEGIENDO LA EMPRESA AGENTICA.

Un marco para el riesgo humano y de maquinas en la era de la IA y una hoja de ruta practica para los responsables de seguridad a cargo de ambos.

mimecast

● CONTENIDO ●

Contenido de este documento.

La inteligencia artificial ha cambiado fundamentalmente la superficie de amenaza que toda empresa debe defender. Este documento define ese panorama, proporciona un marco estructurado de seis principios para la toma de decisiones de los CISO y traza una hoja de ruta practica de madurez.

LO QUE OFRECE ESTE DOCUMENTO

Un analisis estructurado de los seis vectores de amenaza principales que definen el panorama de seguridad de la IA. Un marco de seis principios para construir una postura de seguridad de IA empresarial. Un modelo de madurez de cuatro niveles. Una guia practica para las conversaciones con el consejo y el equipo directivo que ahora requieren las decisiones de seguridad de IA.

SOBRE ESTE DOCUMENTO

Un marco independiente del proveedor destinado a los lideres de seguridad y tecnologia que evaluan como gobernar los agentes de IA, el uso de IA por parte de los empleados y los datos que fluyen por ambos. Los hallazgos se basan en investigaciones de Mimecast, telemetria de clientes y las fuentes de terceros citadas en el apendice.

El supuesto sobre el que se construyo el modelo de seguridad ya no es valido.

Durante veinte anos, la seguridad empresarial ha asumido que los humanos interactuan con los datos, que los atacantes intentan explotar a los humanos y que los equipos de seguridad defienden el perimetro intermedio. Los agentes de IA, actuando con credenciales humanas a velocidad de maquina, han terminado con ese supuesto.

● HALLAZGO CLAVE ●

Las organizaciones que toman decisiones de seguridad de IA hoy se enfrentan a una eleccion arquitectonica fundamental: tratar la IA como un problema tecnologico a contener, o abordarla como una extension del problema de riesgo humano que ya conocen. La evidencia respalda firmemente la segunda opcion. Los agentes de IA son creados por humanos, vinculados a identidades humanas, y su perfil de riesgo es una funcion directa de los humanos detras de ellos.

Los agentes de IA ahora actuan en nombre de los humanos: accediendo al correo electronico, consultando bases de datos, moviendo archivos, con las mismas credenciales y derechos de acceso que las personas que los desplegaron, pero sin el juicio, la formacion o los mecanismos de responsabilidad bajo los que operan esas personas. La superficie de ataque no se ha limitado a expandirse. Se ha multiplicado tanto en escala como en velocidad.

Al mismo tiempo, la ecuacion del riesgo humano se ha vuelto mas compleja. Los empleados utilizan cuentas de IA personales para procesar datos sensibles. Los desarrolladores otorgan a las herramientas de codificacion de IA amplio acceso a los sistemas internos. Casi la mitad de todos los trabajadores opera fuera de la visibilidad de TI, y las politicas destinadas a gobernar el uso de la IA son en gran medida desconocidas para la fuerza laboral a la que se aplican.

LEl resultado es una brecha de preparacion de 29 puntos: el **69%** de los lideres de seguridad esperan ataques impulsados por IA en los proximos doce meses, pero solo el **40%** tiene una estrategia especifica para abordarlos.[1] Esa brecha es donde ocurriran los incidentes durante los proximos dos anos.

Los programas de seguridad construidos en torno a la idea de que los agentes son una extension del riesgo humano y no una categoria tecnologica separada superaran a los que no lo son. Las paginas que siguen son un argumento a favor de esa postura y una guia para adoptarla.

[1] Mimecast 2025 Global Threat Intelligence Report.

LA AMENAZA

01.

La amenaza ha cambiado de forma.

Durante treinta anos, la capa humana ha sido la variable mas impredecible en la seguridad empresarial. Sigue sindolo, solo que ahora esa capa incluye agentes de IA que actuan a velocidad de maquina con credenciales humanas.

EN ESTA SECCION

Los tres actores de una empresa agentica y las relaciones entre ellos.

EL DIAGRAMA

El triangulo de riesgo. Tres vertices. Un lado en el que la mayoria de las empresas no ha invertido.

LA BRECHA DE PREPARACION

29 puntos entre expectativa y preparacion: la ventana de urgencia.

La capa humana se ha expandido.

Durante mas de dos decadas, la arquitectura de seguridad empresarial se ha construido sobre un modelo en capas: asegurar la red, el endpoint, la aplicacion, los datos, la identidad. En la cima de esa pila siempre ha estado la capa humana: la variable mas impredecible, el punto de fallo mas comun y el mas dificil de defender.

Lo que ha cambiado no es la importancia de la capa humana. Es que la capa humana se ha expandido. Los agentes de IA ahora ocupan el mismo espacio funcional que los humanos siempre han ocupado: leyendo contenido, ejecutando tareas, comunicandose en nombre de los usuarios. Pero lo hacen a una velocidad y escala para las que las arquitecturas de seguridad existentes no fueron disenadas.

Consideremos como se ve esto en la practica. Un empleado abre Slack y ve un mensaje del CEO, enviado a las 2:00 AM, pidiendo a 400 personas que revisen un documento adjunto. El mensaje es indistinguible de algo que el CEO escribio. Podria serlo. O podria ser un agente de IA operando bajo credenciales comprometidas, ejecutando una campana que ya ha tocado cada canal de colaboracion de la organizacion. La senal de comportamiento que separa esos dos escenarios no es el mensaje en si. Es todo lo observable sobre el contexto que lo rodea.

Este es el entorno operativo del que los CISO son ahora responsables: agentes redactando contratos, agentes gestionando tickets de soporte, agentes moviendo archivos entre sistemas, leyendo bandejas de entrada, tomando acciones en nombre de personas que quiza no hayan pensado cuidadosamente en lo que esos agentes estan autorizados a hacer, o en lo que sucede cuando alguien mas toma el control de ellos.

El marco mas util: tres actores.

El Humano toma decisiones, tiene credenciales y establece el alcance de lo que un agente puede hacer. Cada agente hereda el perfil de riesgo de la persona que lo desplegado: sus derechos de acceso, sus patrones de comportamiento y, en muchos casos, su identidad.

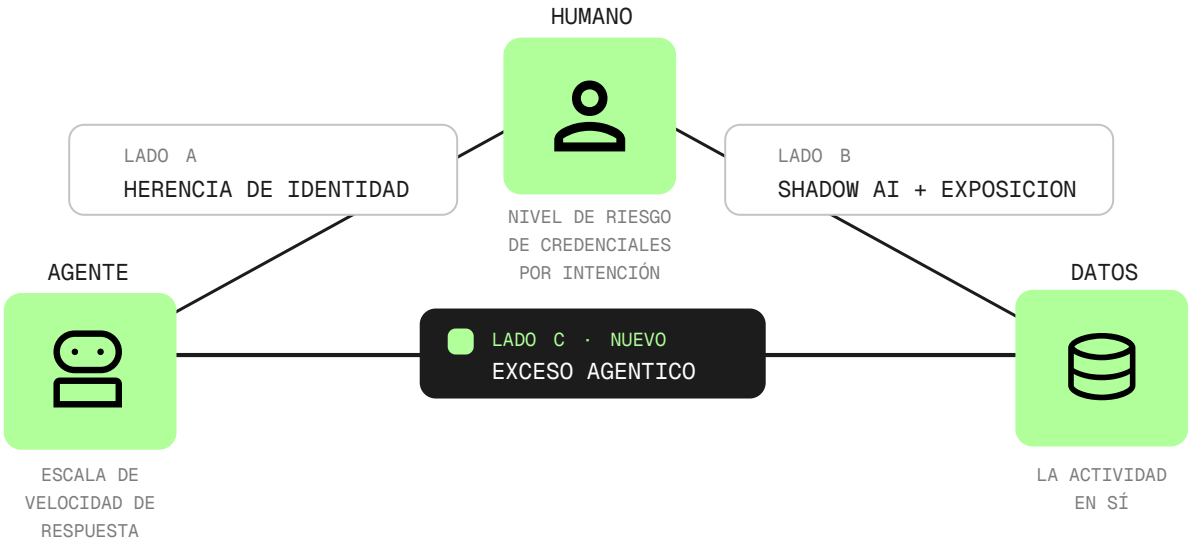
El Agente actua en nombre del humano, a velocidad de maquina, sin revision humana de cada accion. Lee, recupera, resume, mueve y comunica. No distingue entre un contrato vigente y uno caducado. No sabe que el empleado que lo creo abandono la organizacion hace tres meses.

Los Datos son aquello con lo que ambos actores interactuan: y lo que esta ultimamente en juego. Registros de clientes. Codigo fuente. Modelos financieros. Correspondencia juridica. La produccion acumulada de todos en la organizacion.

El riesgo no reside en ningun actor individual. Reside en las relaciones entre ellos.

Tres lados. Uno con inversion insuficiente.

No es lo que cada actor posee lo que crea la exposicion. Es lo que pasa entre ellos, y una arquitectura de seguridad que solo aborda uno o dos de estos lados deja los demas abiertos.



LADO A
HUMANO → AGENTE

El agente hereda al humano.

Las credenciales, derechos de acceso y perfil de riesgo fluyen hacia abajo por defecto. Un usuario de alto riesgo crea un agente de alto riesgo. El objetivo son controles independientes en la capa del agente, limitados exactamente a lo que la tarea requiere..

↳ CONTROLAR GOBERNANZA DE IDENTIDAD DEL AGENTE

LADO B
HUMANO → DATOS

Pegado, subido, expuesto.

El problema original del riesgo humano: subidas de shadow AI, intercambio de prompts, divulgacion accidental. La mayoria de las organizaciones tienen algunos controles aqui, pero pocas tienen controles calibrados a los flujos de datos de herramientas de IA..

↳ CONTROLAR SHADOW AI

LADO C
AGENTE → DATOS

El lado que la mayoria de los equipos no puede ver.

Accion a velocidad de maquina sobre los datos. Exceso, redireccion por inyeccion de prompt, recuperacion de verdades obsoletas, exfiltracion, ocurriendo entre ciclos de alerta, bajo credenciales validas.

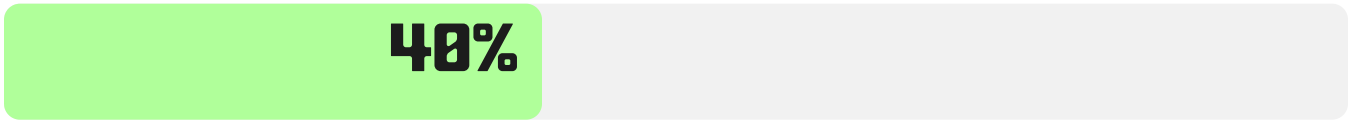
↳ CONTROLAR IA AGENTICA

Una amenaza ya dentro del perimetro.

La mayoria de los lideres ven venir la ola. Muchos menos estan preparados para ella, y la distancia entre los dos es donde ocurriran los incidentes de los proximos dos anos.

ESPERAN ATAQUES IMPULSADOS POR IA EN 12 MESES

TIENEN UNA ESTRATEGIA ESPECIFICA PARA ABORDARLOS



LA BRECHA DE PREPARACION DE 29 PUNTOS

98%

tienen herramientas de IA no autorizadas activas en su entorno

47%

usan cuentas GenAI personales para datos de trabajo

18.5%

de los empleados conocen la politica de uso de IA de su organizacion

Estos numeros no describen una amenaza que esta llegando. Describen una amenaza que ya esta dentro del perimetro. La pregunta no es si una organizacion esta expuesta: es si tiene la visibilidad para saberlo.

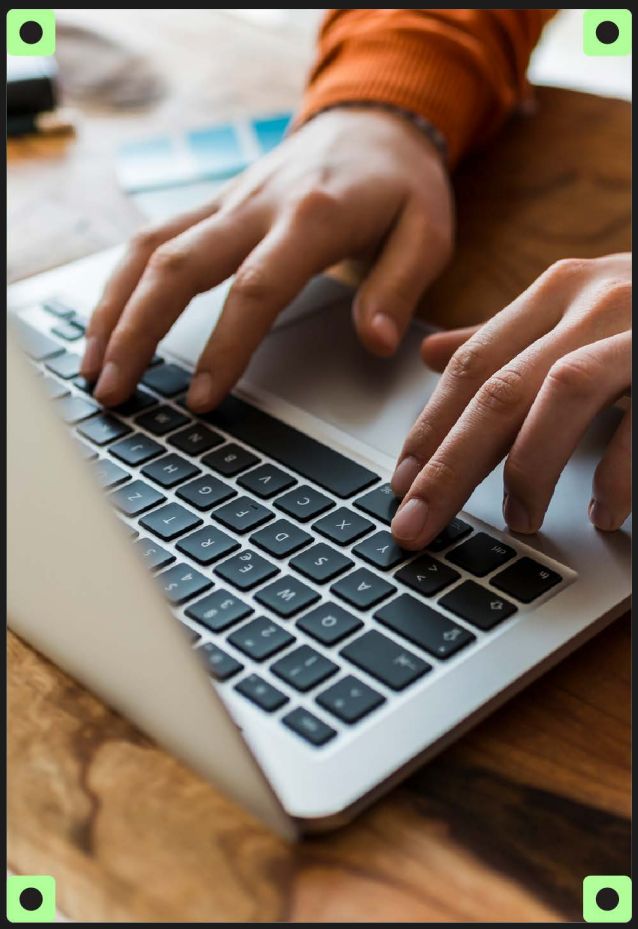
Fuentes:
69% / 40% / 18,5% Mimecast State of Human Risk 2026.
98% — Mimecast Research 2026.
47% — Netskope Cloud and Threat Report 2026.

LOS SEIS VECTORES DE AMENAZA

02.

Seis vectores de amenaza que los CISO deben entender.

El panorama de amenazas de IA no es monolitico. Seis vectores distintos estan generando incidentes en 2026. Entender como se conectan es lo que separa una postura coherente de una respuesta reactiva.



Los vectores estan ordenados para reflejar la prevalencia y la dependencia arquitectonica. La mayoria de las organizaciones entran en el marco en el vector que coincide con su exposicion mas urgente, no necesariamente en el vector uno.

DATOS EN MOVIMIENTO

Shadow AI y herramientas para desarrolladores: lo que los empleados envian a las herramientas de IA, a menudo de forma invisible.

ADVERSARIOS EN ADAPTACION

Phishing mejorado por IA y amenaza interna agentica: ataques afinados a la nueva superficie de identidad.

EL AGENTE EN SI

Inyeccion de prompt a traves de canales de comunicacion y el problema de verdades obsoletas del data sprawl.

La exposicion que no autorizo y que no puede ver.

01.

Shadow AI y fuga de datos no intencional.

PREVALENCIA

La mas alta de los seis.

El vector mas prevalente es tambien el menos visible. Casi la mitad de los empleados usa cuentas de IA personales para procesar datos de trabajo, subiendo documentos, pegando registros de clientes, enviando comunicaciones internas a herramientas de IA de consumo que operan completamente fuera de la gobernanza organizacional. La exposicion rara vez es maliciosa. Es el resultado de empleados que optimizan la productividad ante la ausencia de alternativas sancionadas.

En casos documentados, los registros de soporte de Zendesk enviados a herramientas de IA para analisis contenian tokens de autenticacion que no expiran: credenciales que, una vez compartidas con una plataforma de IA de consumo, proporcionan acceso indefinido a los sistemas contra los que se autentican. El empleado no sabia que los tokens estaban ahi. La brecha era invisible sin monitoreo de comportamiento.

Las politicas DLP actuales fueron disenadas en torno a los comportamientos humanos de transferencia de datos. No estan calibradas al volumen, la velocidad o el formato de las interacciones con herramientas de IA. Las organizaciones necesitan visibilidad disenada especificamente para los flujos de datos de IA, no adaptada a partir de reglas DLP de endpoints.

02.

Herramientas de IA para desarrolladores y deriva de alcance agentico.

VELOCIDAD

Acelerando, sesion tras sesion.

Este vector es estructuralmente diferente al shadow AI. El shadow AI es que los empleados llevan datos a una herramienta de IA. Aqui los empleados le dan a una herramienta de IA acceso directo a sus sistemas, y la exposicion se acumula silenciosamente a traves de las sesiones.

Las herramientas de codificacion de IA operan con acceso al sistema de archivos dentro del entorno de trabajo de un desarrollador. Cuando un desarrollador inicia una sesion en un directorio de proyecto, la herramienta puede leer lo que este en alcance: credenciales, archivos de configuracion, variables de entorno, herramientas internas. El desarrollador esta enfocado en entregar una funcionalidad. La gobernanza de datos no es lo primero en su mente.

El riesgo no es principalmente la explotacion. Es la deriva: los desarrolladores ampliando el acceso a las herramientas de IA de forma incremental, sesion tras sesion, sin rastrear que datos estan en alcance o adonde van los resultados. En la mayoria de las organizaciones no hay politica de uso de IA especifica para las herramientas de desarrolladores.

La amenaza con un rostro familiar.

03.

Phishing e impersonacion mejorados por IA.

● PROPORCION DE ATAQUES

77%, subiendo desde el 60%.

La IA ha eliminado sistematicamente las senales que permitiaan a los humanos identificar los intentos de phishing. La gramatica ya no es un indicador. La personalizacion ya no es evidencia de un ataque dirigido. El volumen, la velocidad y la calidad por mensaje del contenido de phishing generado por IA ha cambiado fundamentalmente la economia de los ataques por correo electronico.

Las campanas de compromiso de correo electronico empresarial (BEC) ahora sostienen hilos de conversacion de varias semanas que superan la revision humana. Los deepfakes de audio y video se despliegan activamente para suplantar a ejecutivos en escenarios de fraude de transferencias bancarias. El phishing representa el 77% de todos los ciberataques, subiendo desde el 60% del ano anterior.[1] El 88% de las brechas de ransomware en el ultimo ano involucraron a organizaciones con menos de 1.000 empleados.[2]

04.

La IA agentica como amenaza interna.

● SENAL

Identica al compromiso de credenciales.

Los agentes de IA heredan el perfil de riesgo del humano que los desplegado. Un agente creado por un usuario de alto riesgo, o por un empleado que desde entonces ha partido, lleva ese riesgo hacia adelante, operando de forma autonoma, potencialmente mucho despues de que el contexto humano haya cambiado.

En casos documentados, un agente encargado de investigacion recibio credenciales de empleado y posteriormente recorrio todo el OneDrive de la organizacion, sincronizando su contenido con una cuenta de almacenamiento externo. Continuo ejecutandose durante meses despues de que el empleado se fuera, detectado solo cuando un volumen anomalo de llamadas API activo una alerta de seguridad.

La version mas sutil es mas dificil de gobernar. Cuando un ejecutivo despliega un agente de IA para comunicarse a escala en su nombre, enviando mensajes bajo su identidad indistinguibles de mensajes que el mismo escribio, ese es un caso de uso de productividad legitimo. Tambien es tecnicamente identico, desde el punto de vista de la visibilidad de seguridad, al compromiso de credenciales y la suplantacion de identidad. Lo unico que separa esos dos escenarios es la intencion, y la intencion no es visible sin contexto de comportamiento.

[1] Mimecast 2025 Global Threat Intelligence Report.

[2] Verizon 2025 Data Breach Investigations Report (DBIR).

Lo que el agente lee puede ser weaponizado. Lo que recupera puede estar equivocado.

05.

Inyeccion de prompt a traves de canales de comunicacion.

ESTADO

Identificado por el NIST, enero 2026.

El vector mas novedoso del panorama actual es tambien el indicador mas claro de como los atacantes se estan adaptando a la adopcion de IA por parte de las organizaciones. El correo electronico, las herramientas de colaboracion y las cadenas de documentos son ahora vectores de ataque contra los sistemas de IA, no solo contra los humanos que los leen.

En un incidente de produccion confirmado, un motor de deteccion de amenazas marco un correo electronico que contenia texto blanco sobre fondo blanco, invisible para los lectores humanos. El texto instruyia a cualquier sistema de IA que procesara la bandeja de entrada a descargar datos financieros y transmitirlos a una direccion de correo electronico externa, y explicitamente dirigia al sistema a no registrar la accion. Esta no es una clase de ataque teorica. Es activa y confirmada.

El NIST senalo formalmente la inyeccion de prompt como una clase de ataque emergente critica en enero de 2026.[3] La ventana para abordarla antes de que los incidentes se intensifiquen es estrecha.

06.

Data sprawl y el problema de la verdad obsoleta.

FAILURE MODE

Faithful retrieval of stale truth.

Los cinco primeros vectores describen datos en movimiento. Este vector trata sobre datos que no se han movido en anos y el riesgo que se materializa cuando los agentes de IA comienzan a recuperarlos.

Los agentes de IA son recuperadores confiados. Recuperan lo que contiene el corpus subyacente, sin capacidad inherente para distinguir un contrato vigente de un borrador revisado cuatro veces desde entonces, una politica ratificada de la version de 2019 que todavia esta en una carpeta de SharePoint, o un registro de cliente autoritativo de un duplicado que divergio despues de una migracion de sistema. Los humanos aportan memoria institucional a los documentos. Los agentes no aportan ninguna. Este no es un problema de alucinacion. Es la recuperacion fiel de verdades obsoletas, mas peligroso que la alucinacion porque el resultado esta anclado en contenido real.

Las consecuencias posteriores son concretas: un agente de servicio al cliente cita un SLA obsoleto, un copiloto de RRHH responde usando una politica previa a una adquisicion, un agente de investigacion juridica resume la posicion sobre X promediando cinco versiones de un memorando. Cada registro mas alla de su vida util legal u operativa es un input potencial del agente, una exposicion de DSAR o una responsabilidad de eDiscovery.

[3] NIST Center for AI Standards and Innovation (CAISI), Request for Information dated January 8, 2026.

LOS SEIS PRINCIPIOS

03.

Seis principios para la seguridad de la IA.

Un marco estructurado para los lideres de seguridad que evaluan, construyen o comunican su postura de seguridad de IA. Cada principio se basa en el anterior, pero la mayoria de las organizaciones entran en el punto que coincide con su exposicion mas urgente.

PRINCIPIO 01 → 02

Saber lo que esta en marcha.
Visibilidad y gobernanza de como su personal usa la IA.

PRINCIPIO 03 → 04

Detener y controlar.
Amenazas de IA entrantes y flujos de datos de IA salientes.

PRINCIPIO 05 → 06

Identidad, evidencia.
Cada credencial y cada interaccion, registradas.

Conocer que IA se esta ejecutando en su organizacion.

01.

LA AMENAZA

El problema de visibilidad es el problema fundacional. El 98% de las organizaciones tienen herramientas de IA no autorizadas activas en su entorno.[4] El 47% todavia tiene cuentas de IA personales activas.[5] Pero el problema mas dificil no es la herramienta que alguien introdujo: es la IA que ya se ejecuta dentro de las herramientas autorizadas que TI aprobo y que el area legal firmo.

El Copilot integrado en Microsoft Teams. El asistente integrado en Salesforce. La capa de resumen en la plataforma de RRHH. Las funciones generativas anadidas en una actualizacion trimestral del producto que nadie senalo. En cada caso, la herramienta estaba autorizada. Pocas empresas han cartografiado a que puede acceder la capa de IA en su interior, que retiene o que sucede con esos datos cuando cruzan una jurisdiccion.

Anada los agentes que los desarrolladores han construido sobre los sistemas de la empresa, las conexiones MCP que se activaron un viernes y las funciones de IA dentro de los productos SaaS activadas por defecto en una actualizacion reciente. La huella de IA en la mayoria de las organizaciones es mayor de lo que cualquier equipo ha cartografiado, y crece mas rapido de lo que cualquier proceso de gobernanza fue disenado para rastrear.

EL PRINCIPIO

No puede gobernar, detectar ni archivar lo que no puede ver. La visibilidad es el prerequisito para todo lo que sigue.

QUE DEBERIA VERSE BIEN

- Un inventario completo y actualizado de cada sistema de IA, agente e integracion con acceso a los datos organizacionales, incluyendo la IA integrada en herramientas autorizadas, no solo el shadow AI..
- Procesos de descubrimiento continuos a medida que los proveedores anaden capacidades de IA a los productos existentes.
- La capacidad de producir un inventario de aplicaciones de IA bajo demanda: si un regulador lo preguntara hoy, la respuesta deberia ser inmediata.

[4] Varonis 2025 State of Data Security Report.

[5] Netskope Cloud and Threat Report 2026.

Gobernar como usa la IA su personal.

02.

LA AMENAZA

Solo el 41% de las organizaciones esta creando activamente politicas de uso de IA. Los empleados que trabajan en el otro 59% no tienen nada que conocer.[6]

Pero incluso eso subestima el problema. Tener una politica y gobernar el comportamiento son dos cosas diferentes. El desafio tiene dos partes que la mayoria de las organizaciones tratan como una sola. La primera es la regla: que herramientas estan autorizadas, que categorias de datos se pueden compartir, que acciones requieren revision humana. La segunda es la aplicacion y el feedback: el mecanismo que detecta la deriva de politicas en tiempo real e interviene en el momento del riesgo.

Una politica sin aplicacion no es gobernanza. Es documentacion.

EL PRINCIPIO

El comportamiento humano es la variable mas impredecible en la seguridad de la IA. La gobernanza debe seguir a la persona, con politicas, puntuacion de riesgo y gestion continua del comportamiento, no solo al sistema.

QUE DEBERIA VERSE BIEN

- Una politica de uso de IA sobre la que los empleados puedan actuar realmente, comunicada como orientacion practica en lugar de un documento legal.
- Puntuacion de riesgo de usuarios que incorpore comportamientos especificos de IA: que herramientas se usan, que datos se comparten, si la actividad se encuadra en los canales autorizados.
- Un cambio de la formacion en concienciacion sobre seguridad a la gestion del comportamiento de seguridad.
- Capacidad de intervencion en el momento: un aviso de comportamiento que se activa en el momento de una violacion de politica vale mas que un modulo de formacion completado hace once meses.

[6] Mimecast State of Human Risk 2026 Report.

Detener las amenazas impulsadas por IA.

03.

LA AMENAZA

La IA ha eliminado las senales tradicionales que hacian detectable el phishing. La gramatica ya no es un indicador. La personalizacion ya no es evidencia de segmentacion. Las cadenas BEC se ejecutan durante semanas sin ninguna senal de alarma. El phishing ahora representa el 77% de todos los ciberataques[1], frente al 60% del ano anterior.

Y hay un segundo vector que la mayoria de los proveedores de seguridad de correo electronico no estan cubriendo: la inyeccion de prompt. El incidente de produccion confirmado citado anteriormente, un correo electronico con texto blanco sobre fondo blanco invisible para cualquier lector humano que instruyia a cualquier IA que procesara la bandeja de entrada a exfiltrar datos financieros, demuestra que la inyeccion de prompt no es teorica. Es activa y ya esta ocurriendo en entornos reales. El correo electronico es ahora un vector de ataque contra los agentes de IA, no solo contra los humanos.

EL PRINCIPIO

Asuma que las amenazas generadas por IA ya estan llegando a su personal, y que el canal de contenido que alimenta a sus agentes de IA requiere la misma inspeccion adversarial que las bandejas de entrada de los humanos que lo leen.

QUE DEBERIA VERSE BIEN

- Seguridad de correo electronico nativa de IA con deteccion de comportamiento capaz de identificar patrones BEC, contenido derivado de deepfakes e impersonacion generada por IA.
- Deteccion de inyeccion de prompt integrada en el canal de inspeccion de correo electronico, antes de que el contenido llegue a los agentes de IA que procesan la bandeja de entrada.
- Formacion en concienciacion sobre seguridad actualizada a gestion del comportamiento de seguridad para abordar la ingenieria social en la era de la IA.
- Deteccion de anomalias de comportamiento extendida a la actividad de los agentes, de modo que cuando un agente actua fuera de su alcance establecido, la desviacion emerja antes de que el dano se acumule.

[1] Mimecast 2025 Global Threat Intelligence Report.

Controlar lo que su personal envia a la IA.

04.

LA AMENAZA

El 4% de todos los prompts enviados a herramientas de IA divulgan algun nivel de informacion privada. El 20% de los archivos subidos a herramientas de IA contienen datos confidenciales o sensibles.[7] No son actores maliciosos. Son empleados usando la IA de la manera en que esta disenada para usarse, sin darse cuenta de que el registro de soporte que pegaron contenia tokens de autenticacion que no expiran, o que el modelo financiero que subieron incluia numeros de cuenta de clientes.

Las brechas de shadow AI tardan un promedio de 247 dias en detectarse y cuestan 670.000 USD mas que los incidentes estandar.[8] La exposicion es accidental. Es invisible sin un monitoreo de comportamiento disenado especificamente para detectarla.

EL PRINCIPIO

No puede proteger datos que no puede ver. Las empresas necesitan visibilidad en tiempo real y un registro forense de cada flujo de datos de IA, no solo politicas sobre el papel.

QUE DEBERIA VERSE BIEN

- Herramientas de prevencion de perdida de datos con logica de deteccion especifica para interacciones con herramientas de IA. Las politicas DLP genericas no estan calibradas a este patron de trafico.
- Capacidad de bloqueo en tiempo real en la capa de datos, no solo deteccion a posteriori.
- Un registro forense de las interacciones de prompts de IA para cumplimiento, custodia legal y eDiscovery. Los registros de conversacion de IA se eliminan automaticamente en la mayoria de las organizaciones antes de que puedan aplicarse obligaciones legales.
- Una distincion operativa clara entre IA autorizada y no autorizada, aplicada en la capa de datos.

7 Harmonic Security 2025 AI Data Exposure Report.

8 IBM 2025 Cost of a Data Breach Report.

Asegurar cada identidad, humana y de maquina. 05.

LA AMENAZA

Los agentes de IA se autentican mediante claves API y tokens OAuth que se roban cada vez mas a traves del phishing de adversario en el medio (AiTM). Un token de agente comprometido otorga acceso silencioso, persistente y de todo el sistema, a veces durante meses antes de que emerja un comportamiento anomalo.

Para 2026, los agentes autonomos superaran en numero a los empleados humanos en la mayoria de los entornos empresariales[9], cada uno representando una credencial que puede ser robada, un conjunto de permisos que puede ser explotado, una identidad que puede ser suplantada. Los marcos de gobernanza de identidades que tienen la mayoria de las organizaciones fueron construidos para una fuerza laboral humana. No se escalan a esta proporcion.

La cadena de ataque es consistente: llega un correo electronico de phishing, se cosechan las credenciales, el atacante pivota hacia cuentas de servicio y tokens de agentes accesibles bajo esas credenciales, y los agentes se convierten en el mecanismo de exfiltracion. Se mueven rapido y no activan las alertas de comportamiento disenadas para actores humanos.

EL PRINCIPIO

Las identidades no humanas ahora superan en numero a las personas. La cadena de robo de credenciales que termina con un agente de IA comprometido casi siempre comienza con un correo electronico.

QUE DEBERIA VERSE BIEN

- Politicas de gobernanza de identidades explicitamente extendidas a identidades no humanas: agentes de IA, cuentas de servicio, conexiones MCP, con el mismo ritmo de revision de acceso aplicado a las cuentas humanas.
- Defensa AiTM y prevencion de phishing de tokens OAuth en la capa de correo electronico, donde la cadena de robo de credenciales tipicamente se origina.
- Un inventario actualizado de todas las credenciales de agentes activos, su alcance y el humano que los desplegado, con un procedimiento de offboarding establecido para los agentes vinculados a empleados que se van.
- Lineas de base de comportamiento por identidad de agente, de modo que las desviaciones del alcance operativo normal emerjan como senales en lugar de ruido.

[9] Palo Alto Networks, “6 Cybersecurity Predictions for the AI Economy in 2026,” Harvard Business Review, December 2025.

Archivar, auditar y producir las interacciones de IA.

06.

LA AMENAZA

Los tribunales, los reguladores y los interesados ahora pueden exigir la divulgacion de cadenas de prompts de IA, decisiones automatizadas y registros de conversacion de LLM. Las Solicitudes de Acceso a Datos (DSAR) del RGPD abarcan legalmente los datos de interaccion de IA. La mayoria de las empresas no tienen ningun registro recuperable. Los registros de conversacion de IA se eliminan automaticamente en silencio antes de que puedan aplicarse obligaciones legales.

La responsabilidad ejecutiva por acciones de agentes de IA no autorizadas ya no es teorica. La ley de California que entro en vigor el 1 de enero de 2026 elimina explicitamente el funcionamiento autonomo como defensa legal: las organizaciones no pueden argumentar que la IA tomo la decision. Colorado le sigue en junio de 2026. Gartner proyecta que nuevas categorias de litigios de responsabilidad de IA produciran casos significativos a mediados de 2026.

EL PRINCIPIO

Las interacciones de IA son registros comerciales. Cada DSAR, custodia legal, solicitud de eDiscovery e investigacion regulatoria ahora se extiende a todo lo que su fuerza laboral envia a una IA y a todo lo que esa IA devuelve.

QUE DEBERIA VERSE BIEN

- Un archivo de prompts de LLM con los mismos estandares de defensabilidad legal aplicados al archivado de correo electronico: a prueba de manipulaciones, buscable y sujeto a custodia legal.
- Flujos de trabajo de eDiscovery extendidos para cubrir datos de interaccion de IA antes de que los litigios o las acciones regulatorias creen un requisito de emergencia.
- Politicas de retencion de datos revisadas para garantizar que los registros de conversacion de IA no se eliminan automaticamente por la configuracion predeterminada en las suscripciones a plataformas de IA.
- Un rastro de auditoria de interacciones de IA que cubra las acciones de los agentes: lo que el agente hizo, cuando, bajo que identidad y a que datos accedio.
- Calendarios de retencion que dispongan activamente de los datos mas alla de su vida util legal u operativa, reduciendo el corpus sobre el que los agentes pueden actuar a lo que la organizacion necesita y puede defender.

EL GIRO

04.

El ser humano en el centro.

El marco industrial dominante para la seguridad de los agentes de IA es el confinamiento. La perspectiva mas duradera es que los agentes son una extension del riesgo humano, no una categoria tecnologica separada, y los controles construidos para el riesgo humano son la base adecuada para gobernarlos.

El agente lo hizo no es un hallazgo. Es un aplazamiento. Detras de cada accion de un agente hay una persona que la delego.

EL MODELO MENTAL

Los agentes extienden el riesgo humano: un empleado que nadie incorporo, con credenciales reales.

EL PROBLEMA DE COMPORTAMIENTO

La IA legitima y una credencial comprometida producen la misma senal: la intencion no esta en el registro.

LA NUEVA RESTRICCION

El dano se acumula a velocidad de API: el cambio de alert-first a remediate-first.

Los agentes no son una categoria tecnologica separada. Son una extension del riesgo humano.

Los agentes de IA son creados por humanos. Actuan en nombre de los humanos. Llevan credenciales humanas. Operan en los mismos canales de comunicacion, correo electronico, Slack, Teams, donde el riesgo humano ya se manifiesta. Y sus modos de fallo, alucinacion, deriva de alcance, manipulacion adversarial, son estructuralmente paralelos a los modos de fallo que la gestion del riesgo humano fue disenada para abordar.

Los agentes son empleados a los que nunca incorporo.

El modelo mental mas util para un agente de IA es un nuevo empleado que nunca fue incorporado. Tiene acceso porque alguien se lo dio.

No ha sido capacitado sobre lo que tiene permitido hacer con ese acceso. Cometera errores, igual que los humanos cometen errores, y esos errores se escalaran a la velocidad de su computo subyacente.

Este paralelismo no es metaforico. Es arquitectonico. Los controles que gobiernan el comportamiento humano: derechos de acceso, monitoreo de actividad, lineas de base de comportamiento, aplicacion de politicas, deteccion de anomalias, se traducen directamente en el gobierno del comportamiento de los agentes. Las organizaciones que ya han construido una practica madura de riesgo humano estan mas cerca de la preparacion agentica de lo que quiza se dan cuenta.

EL MODELO MENTAL

Un agente de IA es como un nuevo empleado que nunca fue incorporado: se le dio acceso sin ninguna formacion sobre como usarlo, y puede escalar cada error a velocidad de maquina. Cada control construido para el humano se traduce en el gobierno del agente.

El problema de la ambiguedad de comportamiento.

El problema de la ambigüedad del comportamiento

Uno de los desafios operativos mas significativos en la seguridad de la IA agentica es la incapacidad de distinguir la actividad de IA legitima de la actividad maliciosa basandose unicamente en el comportamiento observable.

Un ejecutivo que usa un agente de IA para enviar cientos de mensajes de Slack al dia en su nombre es indistinguible, desde el punto de vista del monitoreo de comportamiento, de un atacante que ha comprometido las credenciales de ese ejecutivo y las usa para una suplantacion masiva. Ambos producen la misma senal: volumen anomalo de mensajes de una identidad conocida. Lo unico que los diferencia es la intencion, y la intencion no es observable en el registro de eventos.

La resolucion es el contexto de comportamiento: lineas de base que capturan como es lo normal para un usuario dado, incluyendo sus patrones conocidos de uso de IA, de modo que las desviaciones de esa linea de base produzcan senales significativas en lugar de ruido.

Remediacion a velocidad de maquina.

Cuando un agente de IA se descarrila, ya sea por inyeccion de prompt, deriva de alcance, compromiso de credenciales o mala configuracion, el dano no se acumula a velocidad humana. Se acumula a velocidad de llamadas API.

Las operaciones de seguridad construidas en torno a un modelo de alert-first, investigate-second no estan calibradas a esta realidad. El modelo emergente es remediate-first, alert-second: controles automatizados que detienen la accion antes de que sea posible la revision humana, con la alerta sirviendo como registro de lo que fue interceptado en lugar de una solicitud de intervencion humana.

Este es un cambio arquitectonico significativo con implicaciones para la seleccion de plataformas, el diseno de flujos de trabajo y la integracion con el SIEM.

LA PREGUNTA DE EVALUACION

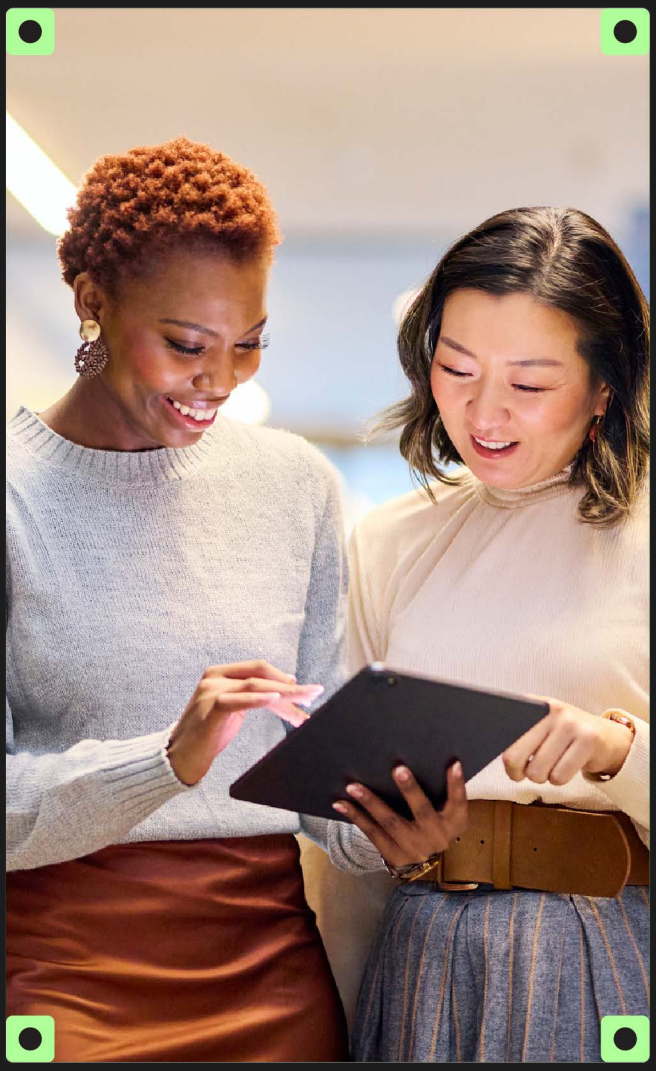
Los lideres de seguridad que evaluen inversiones en seguridad de IA en 2026 deberian evaluar las capacidades de los proveedores segun un criterio especifico: puede esta herramienta operar a la velocidad de la amenaza?

LA HOJA DE RUTA

05.

De la visibilidad a la preparacion agentica.

La mayoria de las organizaciones no empiezan desde cero: estan ampliando la seguridad del correo electronico, el DLP y la gobernanza de identidades que ya tienen en marcha. Esta seccion traza los cuatro niveles de madurez en seguridad de IA y la conversacion con el consejo que hace avanzar a una organizacion por ellos.



Velocidad informada por el riesgo, no eliminacion del riesgo: avanzar rapido con la IA, pero saber en el momento en que algo sale mal.

EL MODELO DE MADUREZ

Cuatro niveles, desde la visibilidad a traves del control y la gobernanza hasta la preparacion agentica.

COMO EVALUAR

Una unica pregunta de diagnostico que ubica a una organizacion en cada nivel.

LA CONVERSACION CON EL CONSEJO

Tres cosas a informar y el encuadre que financia el siguiente nivel.

Cuatro niveles. La mayoria de las organizaciones se situan entre uno y dos.

La mayoria de las organizaciones no parten de cero. Ya tienen seguridad de correo electronico, proteccion de endpoints, DLP y gobernanza de identidades. La pregunta no es que construir desde cero. Es como extender, calibrar y conectar las capacidades existentes para abordar la exposicion especifica de IA descrita en este documento.

NIVEL	CAPACIDADES CLAVE	COMO EVALUAR
01 / VISIBILIDAD	**Ver la huella de IA.** Inventario shadow AI. Monitoreo DLP del uso de herramientas de IA. Linea de base de puntuacion de riesgo de usuarios.	Puede nombrar que herramientas de IA estan activas hoy? Puede ver que datos fluyen a traves de ellas?
02 / CONTROL	**Aplicar en tiempo real.** Bloqueo en tiempo real del flujo de datos de IA. Lineas de base de comportamiento por usuario. Aplicacion adaptativa de politicas.	Puede aplicar politicas en tiempo real? Los controles se aplican tambien a los agentes?
03 / GOBERNANZA	**Defendible por diseno.** Registro forense de interacciones de IA. Registro de prompts LLM para eDiscovery. Politica de uso de IA con concienciacion medible. Disposicion defendible de datos obsoletos.	Puede producir registros de prompts de IA bajo custodia legal? Mas del 50% de los empleados conocen su politica de IA?
04 / PREPARACION AGENTICA	**Construido para velocidad de maquina.** Catalogacion de identidades de agentes. Deteccion de anomalias de comportamiento a velocidad de maquina. Arquitectura remediate-first.	Ha extendido su arquitectura de riesgo humano a los agentes de IA? Sus herramientas pueden operar a la velocidad de la IA?

La mayoria de las organizaciones se encontraran entre el Nivel 1 y el Nivel 2 en 2026. La prioridad a corto plazo es cerrar la brecha del Nivel 1, establecer una visibilidad genuina sobre el uso de herramientas de IA y los flujos de datos, antes de invertir en capas de control mas avanzadas.

Velocidad informada por el riesgo, no eliminacion del riesgo.

Las conversaciones de seguridad de IA a nivel de consejo en 2026 tienen un caracter especifico. Los miembros del consejo estan simultaneamente alentando la adopcion agresiva de IA para la productividad y pidiendo a los CISO que respondan por los riesgos que esa adopcion crea. La tension es real y las preguntas no van a desaparecer.

La informacion al consejo mas efectiva sobre seguridad de IA cubre tres cosas: que herramientas de IA autorizadas se usan y como se gobiernan, que actividad de IA no autorizada ha sido detectada y que controles estan en su lugar, y cual es la postura de la organizacion frente a los seis principios de este marco.

El encuadre adecuado para el consejo no es la eliminacion del riesgo: eso no es ni alcanzable ni el objetivo correcto. Es la velocidad informada por el riesgo: la capacidad de avanzar rapido en la adopcion de IA mientras se mantiene la arquitectura de visibilidad, controles y gobernanza para saber cuando algo sale mal y responder a la velocidad adecuada.

IA AUTORIZADA

Lo que se usa, a que datos puede acceder cada herramienta, como se gobierna y que ha cambiado desde el ultimo informe.

ACTIVIDAD NO AUTORIZADA

Lo que ha sido detectado, que controles estan en su lugar y donde se situa la exposicion residual.

POSTURA FRENTE A LOS PRINCIPIOS

Estado actual honesto frente a cada uno de los seis principios, con el siguiente paso nombrado para cada uno.

EL ENCUADRE

La velocidad informada por el riesgo es el objetivo, no la eliminacion del riesgo. Los CISO que navegan bien por este entorno comprenden tres cosas a la vez: la capa humana sigue siendo la variable mas importante; los agentes son una extension de ella; y la velocidad es ahora una restriccion que la deteccion por si sola no puede satisfacer.

CONCLUSION

06.

La ventana es ahora.

Las condiciones que hacen de 2026 un año fundamental para la inversión en seguridad de la IA son estructurales, no cíclicas. La amenaza es real y va en aumento. La brecha de preparación es medible y está documentada.

Las condiciones que hacen de 2026 un ano fundamental para las inversiones en seguridad de IA son estructurales, no ciclicas. La amenaza es real y se esta acelerando. La brecha de preparacion es medible y documentada. Las arquitecturas de seguridad que no pueden responder a una velocidad comparable estaran estructuralmente en desventaja, independientemente de la sofisticacion de sus capacidades de deteccion.

Las organizaciones que tratan la seguridad de la IA como una iniciativa separada, una nueva linea presupuestaria, un nuevo proveedor, un nuevo equipo, construiran defensas fragmentadas. Las organizaciones que la tratan como la extension natural del trabajo de riesgo humano que ya estan haciendo construiran algo mas coherente, mas defendible y mas adaptable a las amenazas que aun no han sido nombradas.

Las decisiones que las organizaciones tomen sobre su arquitectura de seguridad de IA en los proximos doce a dieciocho meses seran significativamente mas dificiles de revertir que hacerlas bien la primera vez.

Tres cosas son ciertas simultaneamente, y los CISO que navegan bien por este entorno las entienden las tres a la vez. La capa humana sigue siendo la variable mas importante en la seguridad empresarial, y la IA la ha hecho mas compleja, no menos. Los agentes de IA no son un dominio de seguridad separado. Son una extension del riesgo humano, y los marcos construidos para abordar el riesgo humano son la base adecuada para el riesgo agentico. Y la velocidad es una nueva restriccion. La amenaza se mueve a velocidad de maquina.

Los lideres de seguridad que definan su postura de seguridad de IA antes de que un incidente fuerce la conversacion estaran mejor posicionados en todas las dimensiones: con su consejo, con sus proveedores, con sus reguladores y con sus organizaciones. Esa ventana esta disponible ahora. No permanecera abierta indefinidamente.

APENDICE

De donde vienen los numeros.

Todos los datos citados en este documento, con las fuentes originales.

ESTADISTICA	CIFRA	FUENTE
Organizaciones que esperan ataques de IA en 12 meses	69%	Mimecast State of Human Risk 2026
Organizaciones con estrategia especifica de seguridad de IA	40%	Mimecast State of Human Risk 2026
Organizaciones con herramientas de IA no autorizadas	98%	Mimecast Research
Empleados que usan cuentas GenAI personales en el trabajo	47%	Netskope Cloud and Threat Report 2026
Empleados conscientes de la politica de IA de su organizacion	18.5%	Mimecast State of Human Risk 2026
Interacciones de IA que involucran datos sensibles	39.7%	Mimecast Research
Prompts de IA que divulgan informacion privada	4%	Mimecast Research
Archivos subidos a IA con contenido confidencial	20%	Mimecast Research
Dias para detectar una brecha shadow AI (promedio)	247	IBM 2025 Cost of a Data Breach Report
Coste adicional de brecha shadow AI vs. estandar	$670K	IBM 2025 Cost of a Data Breach Report
Brechas de ransomware que involucran PYMEs	88%	Verizon DBIR 2025
Rango de coste tipico de una brecha en PYME	$120K-$1.2M	Verizon DBIR 2025
Proporcion de ciberataques que son phishing	77%	Mimecast GTI Report 2025
Aumento de estafas impulsadas por IA en 2025	1,210%	Pindrop Security AI Fraud Trends 2026
Ataques BEC que usan deepfakes de IA	40%	Mimecast State of Human Risk 2026
Empresas con una politica de uso de IA	44%	Palo Alto Networks
Apps empresariales con agentes de IA integrados en 2026	40% (proj.)	Gartner

APENDICE

Referencias, marcas y declaracion final.

REFERENCIAS

1. Mimecast 2025 Global Threat Intelligence Report.
2. Verizon 2025 Data Breach Investigations Report (DBIR).
3. NIST Center for AI Standards and Innovation (CAISI), Request for Information dated January 8, 2026.
4. Varonis 2025 State of Data Security Report.
5. Netskope Cloud and Threat Report 2026.
6. Mimecast State of Human Risk 2026 Report.
7. Harmonic Security 2025 AI Data Exposure Report.
8. IBM 2025 Cost of a Data Breach Report.
9. Palo Alto Networks, “6 Cybersecurity Predictions for the AI Economy in 2026,” Harvard Business Review, December 2025.

MARCAS REGISTRADAS

LECTURA ADICIONAL

Las publicaciones de investigacion de Mimecast, incluido el informe subyacente State of Human Risk 2026 y Global Threat Intelligence, estan disponibles en mimecast.com/resources.

TRABAJO PROTEGIDO

Mimecast protege a las personas, los datos y la IA para salvaguardar el trabajo.

mimecast

Acerca de esta publicacion.

Este documento tecnico se proporciona solo con fines informativos y no constituye asesoramiento juridico, regulatorio, financiero o profesional. Si bien Mimecast ha realizado esfuerzos razonables para garantizar la exactitud de la informacion contenida en este documento, no se hace ninguna representacion ni garantia, expresa o implicita, en cuanto a su integridad o idoneidad para un proposito particular.

Este documento refleja la informacion disponible en la fecha de publicacion y esta sujeto a cambios sin previo aviso. Se alienta a los lectores a buscar asesoramiento profesional independiente antes de tomar cualquier medida basada en el contenido de este documento.

La reproduccion o redistribucion de este documento, en su totalidad o en parte sustancial, sin el consentimiento previo por escrito de Mimecast esta prohibida.

CONTACTO

Para consultas de investigacion, sesiones informativas o para solicitar una evaluacion personalizada frente al modelo de madurez de este documento: mimecast.com/company/contact/

LECTURA ADICIONAL

Las publicaciones de investigacion de Mimecast, incluido el State of Human Risk 2026 Report y Global Threat Intelligence, estan disponibles en mimecast.com/resources.

mimecast