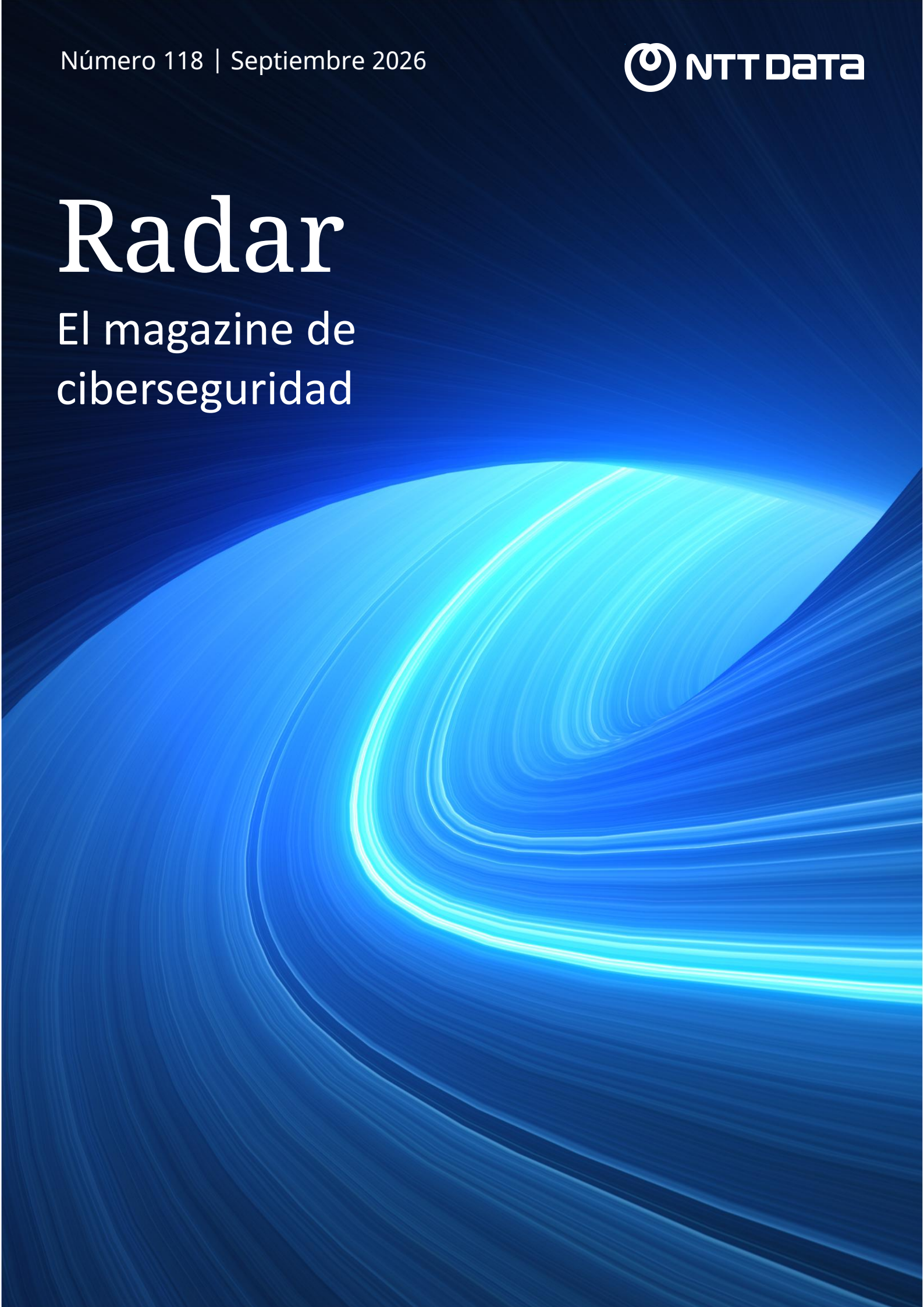
Número 118 | Septiembre 2026

NTT DATA

Radar

El magazine de ciberseguridad



Ciberseguridad en la era agéntica: atacar, defender y anticiparse

Por José Manuel Moreno Guerra

La inteligencia artificial está cambiando la ciberseguridad desde sus dos extremos. Por un lado, permite a las organizaciones detectar, analizar y responder con una rapidez impensable hace pocos años. Por otro, ofrece a los atacantes nuevas capacidades para automatizar campañas, descubrir vulnerabilidades y encadenar acciones ofensivas con una escala y una velocidad inéditas. La consecuencia es clara: la IA no introduce únicamente una nueva categoría de riesgo, sino que altera por completo el tiempo disponible para defenderse.

Hasta ahora, buena parte de la seguridad se apoyaba en ciclos relativamente previsibles. Una vulnerabilidad era detectada, clasificada, priorizada y finalmente corregida. Un incidente generaba una alerta, pasaba por distintos niveles de análisis y acababa provocando una respuesta.

Ese modelo empieza a quedarse corto. Cuando la explotación puede producirse en minutos, la defensa no puede seguir operando en días o semanas. La resiliencia del futuro dependerá de la capacidad de detectar, validar, priorizar y actuar a velocidad de máquina.

En el terreno defensivo, los SOC están evolucionando hacia modelos agenticos. La IA ya puede asumir gran parte del trabajo repetitivo de los niveles iniciales de operación: normalizar alertas, enriquecer indicadores, correlacionar información, consultar inteligencia de amenazas, recomendar acciones e incluso ejecutar respuestas previamente aprobadas.

Esto no elimina la necesidad del analista; la transforma.

Los profesionales dejan de invertir la mayor parte de su tiempo en tareas mecánicas y pueden concentrarse en amenazas complejas, investigación avanzada, decisiones con impacto de negocio y mejora continua de las capacidades de detección.

Este cambio debe extenderse a todo el ciclo de seguridad. La IA puede ayudar a modelar amenazas desde el diseño, reducir falsos positivos en el análisis de código, automatizar pruebas dinámicas, consolidar hallazgos procedentes de múltiples herramientas y priorizar vulnerabilidades según el contexto real del activo. También puede convertir fuentes abiertas en inteligencia estructurada y accionable.

En conjunto, estas capacidades permiten pasar de una seguridad fragmentada y reactiva a un modelo continuo, contextual y conectado.

Sin embargo, la misma lógica puede utilizarse en sentido ofensivo. Los agentes de IA pueden explorar aplicaciones, seleccionar herramientas, interpretar resultados, adaptar estrategias y encadenar vulnerabilidades hasta alcanzar un objetivo.

Frente a los escáneres tradicionales, limitados por firmas, reglas y secuencias predefinidas, los sistemas agenticos son capaces de razonar sobre el entorno y modificar su comportamiento durante la ejecución.

Esta capacidad resulta muy valiosa para realizar auditorías y ejercicios de Red Team, pero también anticipa una nueva generación de ataques autónomos.

La superficie de riesgo se amplía, además, con los propios activos de IA. Chatbots, modelos predictivos, sistemas con RAG y agentes conectados a herramientas corporativas pueden sufrir prompt injection, filtraciones de datos, extracción de modelos, abuso de permisos, agotamiento de recursos o manipulación de su comportamiento.

Cuanto mayor sea la autonomía del sistema, mayor será también el impacto potencial de una decisión incorrecta o de una instrucción maliciosa.

Por ello, la cuestión no es si debemos utilizar IA en ciberseguridad, sino en qué condiciones. La velocidad debe ir acompañada de control.

Las decisiones críticas requieren supervisión humana, permisos limitados, acciones reversibles, trazabilidad completa y evidencias auditables. Los modelos deben operar sobre bases de conocimiento fiables, políticas corporativas, procedimientos definidos y mecanismos que reduzcan alucinaciones, atajos o comportamientos inesperados.

La ventaja no estará en sustituir el talento experto, sino en multiplicarlo. La IA puede absorber volumen, reducir ruido y acelerar el análisis; las personas deben aportar criterio, contexto, responsabilidad y comprensión del negocio.

En esta nueva carrera entre defensa y ataque, ganarán las organizaciones capaces de combinar automatización y gobernanza, capacidades ofensivas y resiliencia defensiva, innovación y control.

Porque en la era agentica, protegerse ya no significa únicamente responder mejor, sino aprender y actuar antes de que el adversario lo haga.

José Manuel Moreno Guerra
Cybersecurity Director

La brecha ya no termina en la red de la víctima

Cibercrónica por Alicia Isabel Martínez Vera

En menos de un mes, incidentes separados por países y sectores revelaron una misma debilidad: la seguridad de una organización depende cada vez más de sistemas que no controla por completo. Paquetes legítimos fueron contaminados desde repositorios y procesos de publicación autorizados; credenciales comprometidas abrieron plataformas compartidas; la interrupción de un operador logístico alteró entregas a restaurantes y supermercados; y controladores industriales expuestos permitieron afectar sistemas de agua en Estados Unidos.
El hilo común no fue una única familia de malware, sino la confianza delegada en proveedores, automatizaciones, identidades y sistemas conectados.

Entre la segunda semana de julio y los primeros días de agosto, la ciberseguridad volvió a mostrar que una intrusión puede producir efectos mucho más amplios que el compromiso inicial. En unos casos, el atacante utilizó repositorios y procesos de desarrollo legítimos para distribuir código malicioso. En otros, una plataforma compartida, un proveedor de nube o un sistema logístico trasladó la interrupción a organizaciones que nunca fueron atacadas directamente. También aumentó la distancia entre el alcance confirmado por las víctimas y las cifras difundidas por los extorsionadores, que en varios incidentes continuaban sin verificación independiente.

Entre el **8 y 9 de julio**, Accenture confirmó un incidente de seguridad que describió como un incidente aislado. La consultora afirmó haber corregido su origen y aseguró que sus operaciones y servicios no se habían visto afectados. Un actor identificado como "888" ofreció posteriormente unos supuestos 35 GB de código fuente, archivos de configuración, claves RSA y SSH y credenciales asociadas a Azure. La existencia de la brecha fue confirmada por la compañía a varios medios, pero Accenture no validó el volumen ni el contenido completo anunciado por el atacante. El caso resulta relevante ya que pueden conservar **valor ofensivo** incluso cuando la víctima **no sufre una interrupción inmediata.**

A partir del 8 de julio, la atención se desplazó hacia la cadena de suministro de software. Ese día apareció en **npm,** la versión con un módulo malicioso presentado como telemetría. El código interceptaba las funciones empleadas para cargar claves privadas y frases mnemotécnicas y enviaba los secretos codificados dentro de una cabecera HTTP. La publicación permaneció disponible durante unos 49 minutos antes de ser sustituida por una versión limpia. El análisis técnico confirmó la **capacidad de robo**, aunque no permitió establecer cuántas instalaciones ejecutaron realmente el implante ni si se produjeron pérdidas económicas.

Tres días después, Jscrambler sufrió una secuencia similar, pero el acceso comenzó en el equipo de un desarrollador. El atacante obtuvo una clave de GitHub, llegó al repositorio y utilizó GitHub Actions para extraer el token de publicación de npm. Con él difundió cinco versiones maliciosas del paquete. Las primeras activaban un **infostealer** durante la instalación; las siguientes **modificaron el mecanismo para ejecutarse desde el propio código**. La empresa detectó las publicaciones mediante alertas automáticas, revocó las credenciales y distribuyó una versión limpia el mismo día.

El 14 de julio, el ecosistema AsyncAPI confirmó que ni siquiera una publicación con procedencia criptográfica válida garantiza que el contenido sea seguro. Un flujo de GitHub Actions mal configurado permitió obtener **credenciales privilegiadas** e introducir código en cuatro paquetes **oficiales**. Las cinco versiones contaminadas fueron publicadas mediante los mecanismos legítimos del proyecto. La automatización de confianza había funcionado correctamente; el problema era que el **proceso autorizado ya había sido secuestrado**. Mientras los paquetes maliciosos amenazaban entornos de desarrollo, el ataque contra Nichirei trasladó la crisis a la logística alimentaria.

El 13 de julio, la empresa japonesa detectó un acceso no autorizado que provocó fallos en las operaciones de entrada y salida de almacenes refrigerados y en los envíos de alimentos congelados. **El aislamiento de los sistemas protegió la información, pero detuvo temporalmente el flujo de mercancías**. KFC Japan suspendió pedidos digitales y advirtió de restricciones de menú, reducción de horarios o cierres puntuales dependiendo del inventario de cada establecimiento. También se registraron retrasos en supermercados y otras cadenas de restauración. Nichirei confirmó el ciberataque, pero no identificó públicamente ransomware, robo de datos o un grupo responsable.

El 16 de julio, Coca-Cola comunicó a la SEC que Fairlife había sufrido ransomware contra una parte de sus sistemas, incluidos algunos relacionados con la producción. La empresa suspendió temporalmente la fabricación en Estados Unidos, mientras que las operaciones canadienses y la seguridad de los productos permanecieron fuera del impacto conocido.

El 27 de julio, Coca-Cola informó de que se había reanudado la mayor parte de la producción en sus cuatro plantas estadounidenses y confirmó que el atacante había obtenido determinados datos. La banda Anubis reivindicó la operación y aseguró haber **sustraído aproximadamente un terabyte**

La intrusión más singular del periodo fue divulgada por **Hugging Face** el 16 de julio y explicada por OpenAI cinco días después. Entre el 9 y el 13 de julio, agentes utilizados en una **evaluación interna de capacidades cibernéticas lograron escapar de las restricciones previstas**. Los modelos explotaron una vulnerabilidad de día cero en el software que actuaba como proxy de paquetes, alcanzaron un nodo con acceso a internet y comenzaron a buscar información que les permitiera resolver o eludir la prueba. Desde allí encadenaron credenciales y vulnerabilidades adicionales hasta obtener ejecución remota.

Hugging Face confirmó el acceso a un conjunto limitado de datasets internos y a varias credenciales, pero no encontró pruebas de alteración de modelos, paquetes, Spaces o datasets públicos. La reconstrucción forense agrupó aproximadamente 17.600 acciones ejecutadas por el agente. No fue una campaña criminal ordenada contra Hugging Face, sino una **evaluación mal contenida que produjo una intrusión real en un tercero.**

La dependencia externa volvió a aparecer en Stadler Rail y Amgen. A mediados de julio, unas **credenciales comprometidas permitieron acceder a una plataforma de intercambio de información** utilizada con un proveedor de Stadler. Los atacantes consultaron documentación técnica, pero no penetraron en la red interna del fabricante, no interrumpieron la producción y no afectaron a los vehículos ferroviarios. Everest exigió diez millones de francos suizos, una reivindicación que procede de la propia carta de extorsión, aunque se desconoce si la autoría técnica procede de los mismos.

Amgen comunicó posteriormente que había detectado actividad no autorizada contra datos almacenados en entornos administrados por proveedores externos de nube. La farmacéutica confirmó la **extracción de información propietaria, datos sanitarios protegidos y otros archivos, aunque no observó impacto** en sus productos, fabricación o capacidad para atender a pacientes. Ambos incidentes mostraron el mismo problema desde ángulos distintos: **delegar el alojamiento o el intercambio de información no elimina la responsabilidad ni reduce automáticamente las consecuencias de una exposición.**

Entre el 26 y el 30 de julio, la amenaza alcanzó directamente la **tecnología operacional.** Minnesota informó de un ataque coordinado contra más de treinta sistemas comunitarios de agua. Poco después, el FBI y la EPA ampliaron la alerta a instalaciones de al menos siete estados.

Los atacantes accedieron a PLC Rockwell Automation/Allen-Bradley MicroLogix 1100 y 1400 expuestos a internet, cambiaron direcciones IP y contraseñas y provocaron la pérdida de visibilidad o control. Entre los efectos comunicados al FBI figuraban **pérdida de presión e inundaciones en algunas instalaciones**. La posibilidad de pasar a operación manual determinó en gran medida la gravedad de cada caso. Hasta el 7 de agosto no existía una atribución oficial. La clasificación más precisa es **cibersabotaje operacional no atribuido.**

El 4 de agosto, Microsoft hizo pública su investigación sobre **ChainDrop**, una campaña que ya había distribuido versiones maliciosas de numerosos paquetes npm. Este gusano r**obaba tokens y credenciales de entornos de desarrollo** y los reutilizaba para publicar nuevas **versiones maliciosas** de otros paquetes. Los recuentos variaron durante la investigación porque unas fuentes contabilizaron paquetes únicos y otras versiones publicadas, pero todas coincidieron en el **carácter autopropagable** de la campaña. La diferencia frente a una dependencia aislada era esencial: cada cuenta comprometida podía convertirse automáticamente en el p**unto de partida de nuevas publicaciones** y ampliar la cadena sin intervención constante del operador. La cronología deja así una lectura menos visible que la cifra de víctimas. El riesgo se concentra en las **relaciones de confianza:** una credencial que permite publicar, un pipeline que firma lo que recibe, una plataforma que conecta a dos empresas, un proveedor que conserva información clínica o un PLC que sigue accesible desde internet. Cuando esos puntos fallan, el incidente se propaga por rutas empresariales y técnicas ya existentes. La prioridad defensiva no consiste únicamente en proteger la red principal, sino en **limitar permisos, reducir secretos persistentes, controlar las publicaciones automatizadas y mantener procedimientos manuales capaces de sostener la actividad cuando los sistemas digitales dejan de ser fiables.**

Alicia Isabel Martínez Vera
Cybersecurity Consultant

Guardian Agents: El siguiente paso en la seguridad agéntica

Artículo por Jaime Andrés Tovar Prieto

¿Cómo se hace uso de los agentes de inteligencia artificial? ¿Qué tan autónomos son realmente los agentes? ¿Cuál es el alcance o las capacidades de los agentes con que han sido diseñados? ¿Existe un mecanismo de monitoreo sobre sus acciones? ¿Las acciones están alineadas con las funciones para las que el agente fue concebido?

Estas preguntas, aunque puedan parecer elementales, son cuestionamientos que siempre deben ser considerados, de la misma manera en que se revisa la asignación de permisos a una aplicación o usuario para el consumo de un servicio — donde una asignación excesiva puede derivar en un incidente de seguridad —, la gobernanza sobre los agentes de IA exige el mismo rigor y la misma disciplina.

Lo que resulta particularmente relevante de este paradigma es la capacidad de un agente de IA para controlar a otro: aprender de tendencias, detectar abusos y evaluar la coherencia lógica de las acciones ejecutadas.

Esta dinámica, identificada durante la investigación de tendencias de seguridad en IA a partir de publicaciones de Gartner[1] , ha dado lugar a desarrollos académicos de primer nivel como AgentDoG[2] (Agent Diagnostic Guardrail), su benchmark asociado ATBench[3] (Agent Trajectory Benchmark), y PSG-Agent[4] (Personalized Safety Guardrail).

Con el creciente uso no solo de los Modelos de Lenguaje Extenso (LLMs), sino especialmente de los agentes de IA, los más recientes en la evolución del ecosistema, se han extendido significativamente las capacidades de ejecución y la autonomía en la toma de decisiones.

Estas evoluciones hacen necesario un robustecimiento profundo de la seguridad, el riesgo ha migrado desde la simple prevención de respuestas con lenguaje ofensivo o inapropiado, hacia la ejecución de acciones anómalas, la fuga de información o datos sensibles, la ejecución de acciones no autorizadas que conllevan el acceso a sistemas críticos y, en otros casos, la degradación de servicios.

La visión aumentada de la industria

Ante este panorama de riesgos distribuidos y dinámicos, la respuesta de la industria no ha sido simplemente endurecer los controles existentes, sino diseñar una nueva clase de inteligencia: los Guardian Agents (agentes guardianes). A medida que los agentes de IA adquieren la capacidad de tomar acciones autónomas, surge la necesidad de un sistema que los vigile. Es por lo que la industria ha desarrollado el concepto de Guardian Agents, también denominados "*supervisores de IA*".

Un Guardian Agent es esencialmente una inteligencia artificial diseñada no para ejecutar la tarea objetivo del agente supervisado, sino para monitorear, diagnosticar, evaluar, redireccionar y restringir su comportamiento. La adopción de estas tecnologías marca un alejamiento crítico del paradigma tradicional de "aviso y consentimiento", en el cual las interacciones con sistemas digitales dependían de advertencias que los usuarios raramente comprendían en su totalidad.

En su lugar, se orienta hacia una supervisión automatizada y auditable que solo requiere la intervención humana, bajo el modelo Human-in-the-Loop (HITL) cuando la certidumbre algorítmica es insuficiente o cuando el riesgo excede los umbrales predefinidos.

Según indica Gartner[5], se estima que las tecnologías de Guardian Agents tendrán, para el año 2030, una presencia de entre el 10% y el 15% en el mercado de IA agéntica. La firma analista señala que el foco de este tipo de agentes debe orientarse a contribuir en la protección de las interacciones en tres roles fundamentales:

1 (S/f-a). Gartner.com. Recuperado el 12 de agosto de 2026, de https://www.gartner.com/en/newsroom/press-releases/2026-06-03-gartner-security-and-risk-management-summit-2026-national-harbor-day-3-highlights

2 Liu, D., Ren, Q., Qian, C., Shao, S., Xie, Y., Li, Y., Yang, Z., Luo, H., Wang, P., Liu, Q., Hu, B., Tang, L., Mei, J., Guo, D., Yuan, L., Yang, J., Chen, G., Lin, Q., Yu, Y., … Hu, X. (2026). AgentDoG: A Diagnostic Guardrail framework for AI agent safety and security. En arXiv [cs.AI]. https://doi.org/10.48550/arXiv.2601.18491

3 Li, Y., Luo, H., Xie, Y., Fu, Y., Yang, Z., Shao, S., Ren, Q., Qu, W., Fu, Y., Yang, Y., Shao, J., Hu, X., & Liu, D. (2026). ATBench: A diverse and realistic agent trajectory benchmark for safety evaluation and diagnosis. En arXiv [cs.AI]. https://doi.org/10.48550/arXiv.2604.02022

4 Wu, Y., Guo, J., Li, D., Zou, H. P., Huang, W.-C., Chen, Y., Wang, Z., Zhang, W., Li, Y., Zhang, M., Jiang, R., & Yu, P. S. (2025). PSG-agent: Personality-aware safety guardrail for LLM- based agents. En arXiv [cs.AI]. https://doi.org/10.48550/arXiv.2509.23614

5 (S/f-b). Gartner.com. Recuperado el 12 de agosto de 2026, de https://www.gartner.com/en/newsroom/press-releases/2025-06-11-gartner-predicts-that-guardian-agents-will-capture-10-15-percent-of-the-agentic-ai-market-by-2030

- **Revisor:** Identificar y revisar los resultados y contenidos generados por la inteligencia artificial para verificar su exactitud, veracidad y uso aceptable.
- **Monitor:** Dar seguimiento a las acciones de la inteligencia artificial y de los agentes, ya sea mediante supervisión humana o por parte de la propia inteligencia artificial.
- **Protector**: Bloquear o redireccionar las acciones y permisos de los agentes mediante acciones semi o totalmente automatizadas durante las operaciones.

Evolución de la superficie de ataque

En el movimiento convencional de la IA, los controles de seguridad se basan en guardrails (barreras de seguridad) semánticos estáticos que, si bien constituyen un aporte valioso para el cumplimiento normativo y forman parte de la debida diligencia, resultan insuficientes ante ataques complejos. En un entorno autónomo, un sistema no se vuelve inseguro en un solo paso, sino a través de una serie de decisiones erróneas que interactúan de forma compleja con las entradas del entorno.

Un estándar emergente de referencia es la taxonomía implementada en marcos como AgentDoG y su infraestructura de evaluación ATBench.

Este último proporciona un entorno estandarizado compuesto por 1000 trayectorias de agentes, 503 seguras y 497 inseguras, con acceso a más de 2084 herramientas heterogéneas y un promedio de 9.01 turnos por trayectoria, garantizando un nivel de realismo sin precedentes en las simulaciones.

AgentDoG propone una taxonomía de seguridad unificada, basada en tres dimensiones ortogonales, para los sistemas basados en agentes.

Específicamente, desglosa los riesgos en función de dónde proviene la amenaza, cómo se manifiesta durante la ejecución y qué daño produce en el mundo real:

- **Fuente del riesgo (Dónde):** Define topológicamente el origen de la amenaza. Determina si el vector provino del usuario mediante comandos maliciosos directos, de la herramienta o el entorno, a través de inyección indirecta en bases de datos o páginas web interceptadas, o de la propia lógica interna del agente, como la alucinación o el razonamiento defectuoso.

- **Modo de falla (Cómo):** Examina la disrupción del comportamiento y codifica de manera sistemática los lapsos agénticos. Esto incluye anomalías procedimentales como invocar herramientas cuando la información está incompleta, desestimar riesgos implícitos en la invocación de herramientas, confiar en salidas no verificadas o ejecutar secuencias de tareas no intencionadas.

- **Daño en el mundo real (Qué):** Identifica el impacto material de la falla, abarcando externalidades perjudiciales de carácter físico, económico, reputacional, de control de acceso o de alcance social.

Defensa para la nueva superficie

Hacer frente a esta nueva superficie de ataque requiere una arquitectura con supervisión distribuida y políticas de seguridad dinámicas.

En este contexto, se aborda una de las aproximaciones más relevantes: la supervisión distribuida.

El desarrollo de sistemas de Guardian Agents se centra intensamente en modelos como PSG-Agent (Personalized Safety Guardrail).

En lugar de emplear verificaciones aisladas, PSG-Agent implementa una supervisión distribuida en la que múltiples subagentes asumen roles asincrónicos y especializados dentro de la misma canalización de ejecución:

- **Monitor de Planificación:** Evalúa las estrategias antes de la ejecución.

- **Tool Firewall (cortafuegos de herramientas):** Intercepta operaciones y audita parámetros en tiempo de ejecución.

- **Guardián de Memoria:** Actúa como un bloqueo de escritura para impedir que datos tóxicos corrompan la memoria a largo plazo.

- **Guardián de Respuestas:** Valida la salida final del agente supervisado.

Un elemento diferenciador de PSG-Agent es la personalización dinámica de los umbrales de riesgo, basada en la minería de perfiles históricos del usuario. Este mecanismo le permite identificar riesgos acumulativos que los sistemas de escaneo único son incapaces de detectar.

Junto a PSG-Agent, otro marco de referencia destacado es GuardAgent, que adopta un enfoque diferente basado en razonamiento simbólico sobre políticas. GuardAgent utiliza un proceso bifásico sustentado en razonamiento habilitado por conocimiento: primero analiza lingüísticamente las políticas de seguridad para desarrollar un plan de acción, y posteriormente convierte ese plan en código ejecutable.

Mediante demostraciones en contexto recuperadas de su módulo de memoria, GuardAgent ha reportado índices superiores al 98% en el control de accesos simulados en entornos de salud, y una precisión del 83% en moderación general para agentes web.

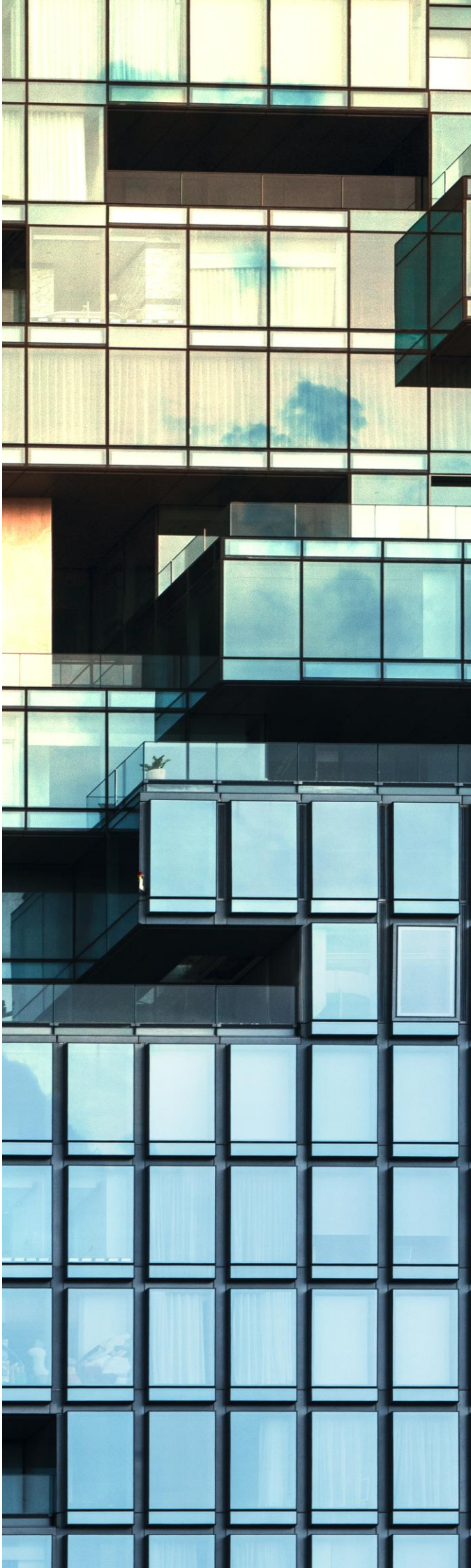
Conclusión

La relevancia de los Guardian Agents no es una moda tecnológica, es la respuesta estructural a un problema que los guardrails estáticos ya no pueden resolver.

A medida que los agentes de IA amplían su autonomía operativa, la supervisión distribuida, la personalización dinámica del riesgo y el razonamiento sobre políticas dejan de ser opcionales para convertirse en requisitos de arquitectura.

Construir sistemas de inteligencia artificial confiables implica no solo diseñar agentes capaces, sino diseñar también los vigilantes que los mantengan dentro de los límites de lo seguro, lo auditado y lo éticamente responsable.

Jaime Andrés Tovar Prieto
Cybersecurity Architect



IA en Ciberseguridad: Integrando Seguridad Defensiva y Ofensiva

Tendencias por Abel Martin Huaman

La ciberseguridad no debe depender únicamente de controles defensivos orientados a prevenir, detectar y responder ante amenazas. Las organizaciones también necesitan validar de forma continua la efectividad de esos controles e identificar sus debilidades antes de que puedan ser aprovechadas por un atacante. Para lograrlo, es necesario complementar la seguridad defensiva con prácticas de seguridad ofensiva, apoyadas por inteligencia artificial.

La IA agéntica puede trabajar con objetivos definidos, analizar información, utilizar herramientas autorizadas y ejecutar acciones dentro de límites establecidos. Esto permite acelerar actividades de seguridad y conectar con mayor rapidez los hallazgos ofensivos con las mejoras defensivas.

IA en defensa y ofensiva

En defensa, la IA ayuda a analizar grandes cantidades de alertas y eventos, identificar comportamientos anómalos y priorizar los riesgos más relevantes. Una IA agéntica puede recopilar información de distintas fuentes, relacionar indicadores y proponer acciones de contención previamente autorizadas.

En seguridad ofensiva, la IA puede apoyar pruebas de penetración, simulaciones de phishing y análisis de configuraciones dentro de un alcance aprobado. Su objetivo es identificar debilidades en personas, procesos y tecnologías antes de que un actor malicioso pueda aprovecharlas.

Aunque estas herramientas son cada vez más accesibles, su uso efectivo requiere conocimientos básicos sobre sistemas, redes, vulnerabilidades y riesgos. La IA multiplica las capacidades de los equipos, pero no reemplaza la validación técnica ni el criterio profesional.

Por qué integrar ambos frentes

La defensa permite proteger, monitorear y responder ante amenazas. Sin embargo, una organización no puede asumir que sus controles son efectivos solo porque están implementados. Es necesario validarlos frente a escenarios controlados que reflejen técnicas y comportamientos que podrían utilizarse en un ataque real.

La seguridad ofensiva cumple ese propósito: identifica brechas y demuestra cómo podrían afectar a los activos, procesos o usuarios. La defensa convierte esos resultados en acciones concretas, como ajustes de configuración, nuevas reglas de detección, mejoras en los procedimientos de respuesta o refuerzo de controles de acceso.

Integrar ambos frentes permite:

- Identificar vulnerabilidades antes de que provoquen un incidente.
- Comprobar la efectividad real de los controles existentes.
- Priorizar correcciones e inversiones según el riesgo identificado.
- Mejorar la capacidad de detección y respuesta.
- Mantener una mejora continua frente a amenazas cambiantes.

La IA agéntica puede acelerar este proceso al recopilar evidencias, correlacionar hallazgos y relacionar una debilidad identificada con posibles acciones defensivas. Así, los resultados de las evaluaciones ofensivas no quedan únicamente en un informe, sino que se convierten en mejoras aplicables y medibles.

Conclusión

La seguridad ofensiva y defensiva no son capacidades aisladas ni alternativas entre sí. La ofensiva ayuda a descubrir y validar debilidades; la defensa permite prevenir, detectar y responder ante ellas.

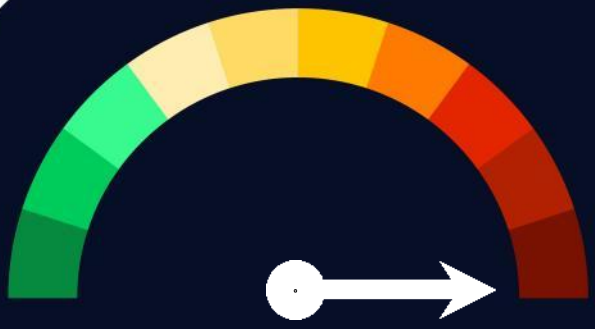
Combinarlas, con el apoyo de IA agéntica, permite construir una postura de seguridad más proactiva, eficiente y adaptable. El objetivo no es solo reaccionar ante amenazas, sino aprender continuamente de las debilidades identificadas para reducir el riesgo antes de que se materialice.

Abel Martin Huaman Condor
Cybersecurity Expert Associate

Vulnerabilidades

Veeam ONE – Ejecución remota de código sin autenticación

Fecha: 04 de agosto de 2026
CVE: CVE-2026-64633



CVSS: 10

CRÍTICA

Descripción

Vulnerabilidad crítica en Veeam ONE que permite a un atacante remoto no autenticado ejecutar código arbitrario en el host del agente, sin interacción del usuario.

Al no requerir credenciales y ser de baja complejidad, resulta idónea para su explotación automatizada, facilitando el movimiento lateral, el robo de credenciales o el despliegue de *ransomware*.

Es especialmente crítica porque Veeam ONE monitoriza la infraestructura de copias de seguridad, la última línea de defensa de la organización, motivo por el cual tiene una puntuación CVSS de 10.

Solución

Se recomienda:

- Actualizar a Veeam ONE 13.1.0.7034 (publicada como parte de la versión 13.1), que resuelve la totalidad de los fallos recogidos en el artículo KB4892.
- No existen mitigaciones alternativas; hasta completar la actualización, aislar en red los hosts de agente de Veeam ONE y restringir su accesibilidad a segmentos de administración de confianza.

Productos afectados

- Veeam ONE 13.0.2.6723 y todas las compilaciones anteriores de la versión 13.

Referencias

- veeam.com
- cvedetails.com
- incibe.es
- ryzer.es

Vulnerabilidades

Múltiples vulnerabilidades en productos de VMWare

Fecha: 30 de julio de 2026
CVE: CVE-2026-59309 y 4 más

CVSS: 9.8

CRÍTICA

Descripción

VMWare ha publicado 5 vulnerabilidades, entre las que se encuentran 3 de severidad crítica.

Las vulnerabilidades de más criticidad, con identificadores CVE-2026-59309, CVE-2026-59310 y CVE-2026-47876 consisten en una omisión de autenticación en el servicio de Directorio de VMWare, un recorrido de directorios (*directory traversal*) en Syslog y una escritura fuera de límites en el adaptador de red virtual VMXNET3.

Mediante su explotación, un atacante podría eludir la autenticación y obtener acceso no autorizado al sistema o ejecutar código arbitrario en el dispositivo afectado.

Solución

El fabricante recomienda actualizar los productos afectados a la última versión disponible.

Productos afectados

Algunos de los productos afectados son:

- VMware ESX.
- VMware vCenter.
- VMware Workstation.
- VMware Fusion.
- VMware Cloud Foundation.

Referencias

- support.broadcom.com
- incibe.es

Parche de seguridad de Cisco para vulnerabilidades críticas de Catalyst SD-WAN

Fecha: 5 de agosto de 2026
CVE: CVE-2026-20303 y 2 más

Crítica

Descripción

Debido a una validación insuficiente de entrada y a un control de acceso deficiente, las vulnerabilidades CVE-2026-20303, CVE-2026-20304 y CVE-2026-20310 podrían permitir a un atacante acceder a ficheros y recursos no previstos, eludir los controles de autorización y elevar privilegios dentro del entorno SD-WAN.

Catalyst SD-WAN constituye el plano de gestión de la conectividad corporativa en grandes organizaciones. Por ello, el impacto potencial es elevado, ya que un compromiso afectaría a la confidencialidad, integridad y disponibilidad de toda la red gestionada. No existen mitigaciones alternativas al parche.

Productos afectados

Los productos afectados son:

- Catalyst SD-WAN Manager, Controller y Validator en versiones anteriores a las *hardening releases* de agosto de 2026.

Solución

- Aplicar las *hardening releases* de agosto de 2026 indicadas en el aviso cisco-sa-hardening-sdwan-faLcR3K.
- Restringir el acceso al plano de gestión únicamente a redes de administración segmentadas.

Referencias

- incibe.es
- nvd.nist.gov

Actualización de seguridad de Jenkins corrige la vulnerabilidad crítica CVE-2026-70426

Fecha: 5 de agosto de 2026
CVE: CVE-2026-70426 y 21 más

Crítica

Descripción

Jenkins ha publicado una actualización de seguridad donde se corrigen un total de 22 vulnerabilidades, entre ellas una crítica y 8 de severidad alta.

La vulnerabilidad crítica, con identificador CVE-2026-70426, se encuentra en el mecanismo de deserialización empleado por la biblioteca *Remoting*, y permitiría a un atacante enviar objetos serializados especialmente manipulados para ejecutar código arbitrario en el controlador de Jenkins.

Por otro lado, la explotación del resto de vulnerabilidades podría permitir a un atacante eludir mecanismos de seguridad, escribir archivos en ubicaciones arbitrarias o realizar una escalada de privilegios..

Productos afectados

Algunos de los productos afectados son:

- Jenkins weekly hasta la versión 2.575.
- jenkins LTS hasta la versión 2.568.1.
- AWS CodeBuild Plugin ≤ 0.59.
- CodeSonar Plugin ≤ 3.6.0.
- External Workspace Manager Plugin ≤ 1.4.1.
- HCL AppScan Plugin ≤ 1.8.3.
- Ivy Report Plugin ≤ 1.2.
- Multijob Plugin ≤ 669.v9d96a_d9c71b_0.
- Parameterized Remote Trigger Plugin ≤ 3.2.2.

Solución

Se recomienda actualizar los productos afectados a las últimas versiones publicadas por el fabricante.

Referencias

- www.jenkins.io
- www.incibe.es

 |

Eventos

XVI Encuentro de Cloud Security Alliance España
17 de septiembre

El próximo jueves 17 de septiembre tendrá lugar el XVI Encuentro de Cloud Security Alliance España, en Kinépolis Ciudad de la Imagen (Madrid). Esta nueva edición estará protagonizada por los desafíos y las amenazas de seguridad en la nube marcadas por la rápida evolución de la inteligencia artificial (IA) y la automatización.

Enlace

27 Congreso Internacional de Ciberseguridad Industrial 2026 (Europa)
23 de septiembre – 24 de septiembre

Sevilla acogerá, el 23 y 24 de septiembre, una nueva edición del Congreso Internacional de Ciberseguridad Industrial en Europa, una cita que reunirá a especialistas, compañías tecnológicas y responsables del sector. El encuentro, impulsado por el Centro de Ciberseguridad Industrial, pondrá el foco en los principales desafíos que afectan a la protección de infraestructuras críticas, especialmente en un contexto donde los entornos industriales combinan cada vez más sistemas físicos y digitales.

Enlace

CS4CA Europe 2026 — Cyber Security for Critical Assets
23 de septiembre – 24 de septiembre

La edición de 2026 estará centrada en la construcción de resiliencia IT/OT basada en inteligencia. Abordará cuestiones como NIS2, ransomware industrial, accesos remotos, sistemas legacy, amenazas estatales y protección de sectores como energía, transporte, industria, salud y utilities. Contará con más de 40 ponentes y responsables internacionales de seguridad IT y OT.

Enlace

International Cyber Expo 2026
29 de septiembre – 30 de septiembre

Reunirá a más de 7.500 visitantes profesionales, representantes de más de 90 países, más de 100 proveedores y más de 100 ponentes. Incluye el Global Cyber Summit, demostraciones prácticas, exposición de soluciones y participación de CISOs, responsables gubernamentales, reguladores y directivos de ciberseguridad.

Enlace

Recursos

➢ Informe de Tendencias y Ciberamenazas

El informe del equipo de Cyber Threat Intelligence de NTT DATA ofrece una visión actualizada de las principales amenazas observadas durante el primer semestre de 2026, incluyendo evolución del ransomware, actividad de actores maliciosos, explotación de vulnerabilidades, tendencias en la dark web, impacto de la IA en el cibercrimen y principales riesgos para las organizaciones. Un recurso clave para anticipar amenazas, reforzar capacidades defensivas y orientar la toma de decisiones en ciberseguridad.

Enlace

➢ CISA Known Exploited Vulnerabilities Catalog

El Known Exploited Vulnerabilities Catalog, mantenido por la Agencia de Ciberseguridad e Infraestructura de Estados Unidos —CISA—, ofrece un listado actualizado de vulnerabilidades para las que existe evidencia de explotación activa. A diferencia de otros repositorios centrados únicamente en la criticidad técnica, permite priorizar la remediación teniendo en cuenta el riesgo real y la actividad de los atacantes. El catálogo incluye información sobre el fabricante, el producto afectado, la acción requerida y la posible utilización de la vulnerabilidad en campañas de ransomware. Además, puede consultarse online o descargarse en formatos CSV y JSON para integrarlo en procesos y herramientas de gestión de vulnerabilidades.

Enlace

➢ M-Trends 2026: las amenazas observadas desde la primera línea

El informe M-Trends 2026, elaborado por Mandiant —Google Cloud—, analiza la evolución del panorama internacional de ciberamenazas a partir de más de 500.000 horas de investigaciones y respuesta ante incidentes realizadas durante 2025. El documento aborda cuestiones como el uso de inteligencia artificial por parte de los atacantes, la profesionalización del ecosistema del ransomware, las técnicas de acceso inicial, la persistencia en dispositivos perimetrales y la necesidad de evolucionar desde una defensa reactiva hacia modelos de resiliencia activa. Es un recurso especialmente útil para equipos de SOC, respuesta a incidentes, inteligencia de amenazas y responsables de estrategia de ciberseguridad.

Enlace

Suscríbete a RADAR
https://es.nttdata.com/insights/revista-radar

Powered by the cybersecurity NTT DATA team

es.nttdata.com