Número 113 | Abril 2026

NTT DATA

Radar

El magazine de ciberseguridad

IA y Ciberseguridad: la Nueva Guerra entre Algoritmos

Por Pedro Felipe del Jesús Canto Vela

Nunca habíamos tenido tanta capacidad tecnológica para protegernos... ni tanta para atacarnos. La inteligencia artificial ha irrumpido en la ciberseguridad como un punto de inflexión que redefine el equilibrio entre defensa y riesgo. Los mismos algoritmos que detectan amenazas en milisegundos pueden utilizarse para lanzar ataques más rápidos, precisos y difíciles de rastrear. La automatización ya no es solo una ventaja competitiva, es el nuevo campo de batalla.

Durante años, la seguridad se basó en fortalecer infraestructuras y mejorar la respuesta ante incidentes. Sin embargo, la IA, especialmente la generativa y los modelos avanzados de aprendizaje automático, ha transformado el panorama. No solo enfrentamos amenazas más sofisticadas, sino un entorno donde la velocidad de adaptación marca la diferencia. La IA acelera el ataque y potencia la defensa, obligando a las organizaciones a replantear su estrategia más allá de la simple adopción tecnológica.

En el ámbito ofensivo, la inteligencia artificial ha reducido la barrera de entrada al cibercrimen. Las campañas de *phishing* se personalizan en tiempo real, los *deepfakes* permiten suplantaciones creíbles y el *malware* se adapta dinámicamente para evadir controles. La ingeniería social se apoya en modelos capaces de generar mensajes convincentes y contextualizados. Pero el mayor cambio no es solo la sofisticación, sino la velocidad. Los atacantes pueden probar variantes y ajustar tácticas en cuestión de horas, mientras muchas organizaciones aún operan con procesos diseñados para un contexto más lento.

La democratización de estas tecnologías amplía la superficie de amenaza y complica la atribución. En un entorno sin perímetro definido, con infraestructuras en la nube y cadenas de suministro digitales, surgen nuevos riesgos como la exposición de datos en modelos generativos, el *prompt injection* o el "envenenamiento" de modelos. El riesgo ya no reside únicamente en la infraestructura, sino también en los algoritmos y los datos que los alimentan.

Sin embargo, la IA también es un pilar esencial de la defensa moderna. Permite analizar grandes volúmenes de información en tiempo real, detectar anomalías, priorizar alertas y automatizar respuestas, reduciendo tiempos de reacción. El análisis predictivo ayuda a anticipar amenazas antes de que se conviertan en incidentes críticos, fortaleciendo una postura más proactiva.

La clave no está en la tecnología por sí sola, sino en su integración dentro de un marco sólido de gobernanza y gestión del riesgo. La inteligencia artificial amplifica capacidades, pero no sustituye la estrategia ni el juicio humano. Sin supervisión y controles adecuados, puede generar decisiones desalineadas o una falsa sensación de seguridad.

En la era de la automatización, la verdadera ventaja no consiste en adoptar la IA más rápido que los atacantes, sino en integrarla de forma consciente dentro de un modelo de resiliencia. Porque, incluso cuando los algoritmos aceleran el juego, la responsabilidad de proteger el futuro sigue siendo profundamente humana.

Pedro Felipe del Jesús Canto Vela
Cybersecurity Expert Analyst

Cuando la inteligencia artificial aprende a engañar… y a operar

Cibercrónica por Juan Pablo Camperos

Entre 2024 y lo que va de 2026, la inteligencia artificial dejó de ser únicamente una herramienta de eficiencia para convertirse en un factor de riesgo operativo. No porque introduzca amenazas completamente nuevas, sino porque amplifica las existentes a una velocidad que las organizaciones aún no logran gobernar.

Uno de los ejemplos más visibles de esta evolución es el uso de *deepfakes* en fraudes corporativos. En Asia, un empleado participó en una videollamada con lo que parecían ser ejecutivos de su organización. Las voces, los gestos y el contexto eran coherentes. La instrucción fue directa: ejecutar transferencias urgentes. No hubo explotación técnica ni *malware*. Solo confianza mal utilizada. La identidad visual dejó de ser suficiente.

Un caso similar en Singapur reforzó esta tendencia. Un director financiero fue inducido a transferir una suma considerable tras interactuar con una reunión ejecutiva completamente generada mediante IA. El engaño solo se detectó cuando los atacantes intentaron escalar la operación. En ambos escenarios, la tecnología no atacó el sistema, atacó la percepción.

Este cambio redefine el punto de entrada. La ingeniería social ya no depende exclusivamente de la persuasión humana, sino de la capacidad de replicar identidades con precisión. La pregunta deja de ser si el mensaje es creíble, y pasa a ser si existe un proceso que permita cuestionarlo.

En paralelo, otro frente comenzó a mostrar riesgos más estructurales. Microsoft inició acciones legales contra un grupo que utilizaba credenciales comprometidas para acceder a servicios de inteligencia artificial generativa. El objetivo no era robar información, sino explotar la infraestructura para generar contenido malicioso y comercializar su uso. El problema no estaba en el modelo, sino en el control de accesos.

Al mismo tiempo, la cadena de suministro de la IA empezó a mostrar fallas similares a las del *software* tradicional. Investigaciones recientes identificaron modelos maliciosos publicados en repositorios abiertos, diseñados para ejecutar código al ser integrados en entornos de desarrollo. Los modelos dejan de ser datos pasivos y se convierten en componentes activos dentro del sistema.

A esto se suma un elemento que muchas organizaciones aún subestiman: la gestión de la información. En Corea del Sur, autoridades detectaron que una plataforma de IA había transferido datos de usuarios y contenido de *prompts* sin consentimiento. Este incidente evidenció un problema crítico: los *prompts* contienen contexto de negocio, decisiones e información sensible. Tratarlos como texto desechable es, en la práctica, una fuga de datos.

Sin embargo, el punto de inflexión más relevante aparece cuando estos patrones alcanzan la operación. En Amazon, una serie de incidentes durante 2026 evidenció cómo el uso de asistentes de código y automatización basada en IA puede impactar sistemas productivos. Cambios ejecutados con alta velocidad, sin validaciones suficientes y con controles débiles, derivaron en interrupciones y pérdidas operativas significativas.

Este tipo de situaciones marca la convergencia entre TI y TO. Cuando la inteligencia artificial deja de asistir y comienza a influir en la ejecución, el impacto deja de ser digital. Un error puede escalar desde una configuración hasta afectar la continuidad completa de un servicio.

Y este fenómeno no se limita a entornos digitales. En sistemas de agua potable se han documentado accesos no autorizados donde parámetros operativos fueron alterados. En el sector de *oil & gas*, manipulaciones de sensores han generado alarmas falsas y decisiones incorrectas. En estos casos, la IA no es el origen directo del ataque, pero sí un acelerador que reduce el esfuerzo necesario para llegar a ese punto.

Más preocupante aún es que el acceso ya no siempre se logra explotando sistemas. Casos recientes muestran cómo actores utilizan IA para construir identidades creíbles y atravesar procesos de contratación. Una vez dentro, el atacante ya no necesita vulnerabilidades: tiene acceso legítimo. En entornos donde los sistemas industriales están conectados o gestionados remotamente, este tipo de infiltración representa un riesgo estructural.

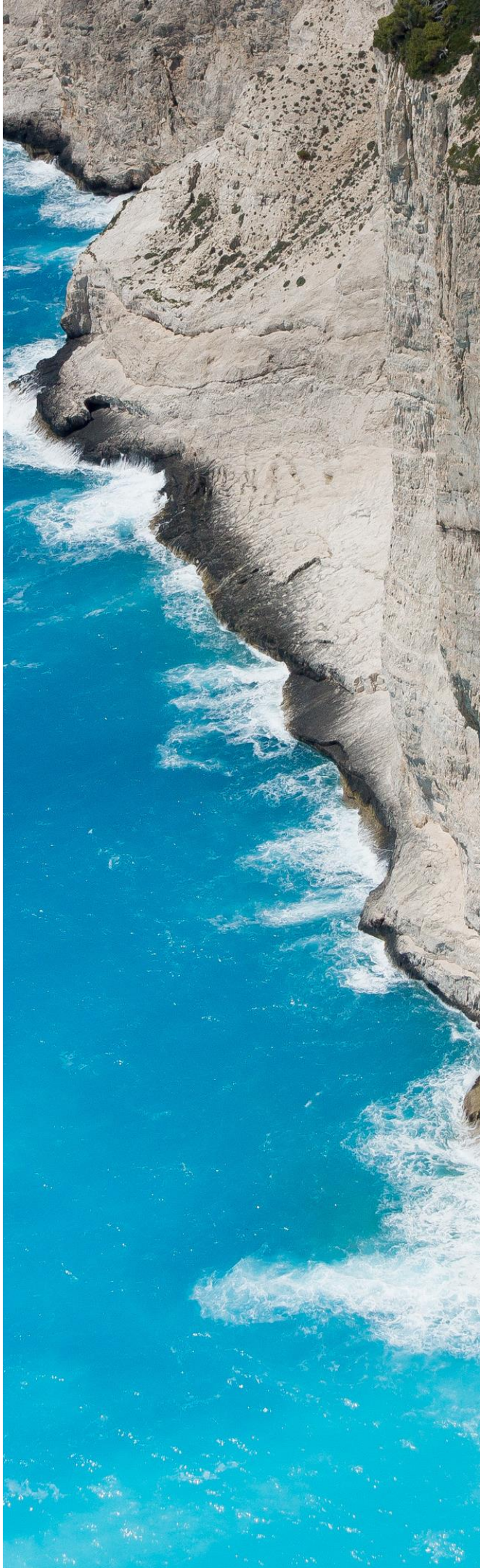
En conjunto, estos incidentes muestran una evolución clara. La identidad deja de ser confiable por apariencia, los accesos siguen siendo el punto más débil, la cadena de suministro se extiende a modelos y la automatización amplifica cualquier error. La IA no rompe los sistemas; acelera sus fallas.

Cuando estos patrones alcanzan entornos operativos, el impacto cambia de naturaleza. Un fraude puede interrumpir pagos a proveedores críticos, una credencial comprometida puede abrir acceso a sistemas industriales, y un cambio automatizado mal controlado puede afectar procesos productivos completos. La ciberseguridad deja de ser un problema técnico para convertirse en un factor de continuidad.

La respuesta no pasa por limitar el uso de inteligencia artificial, sino por gobernarla. Las organizaciones que están avanzando no son las que más invierten en IA, sino las que establecen controles claros sobre su uso y, sobre todo, sobre su impacto.

Porque en este nuevo escenario, la pregunta ya no es si la IA puede ser utilizada en un ataque, sino si las organizaciones están preparadas para operar con ella sin perder el control.

Juan Pablo Camperos
Cybersecurity Expert Architect



Más allá de la ilusión: Desinformación y *deepfakes* en la era de la IA

Artículo por Prakash Narayanamoorthy y Ben Colman*

**Este artículo fue publicado originalmente bajo el título "Beyond the illusion: Disinformation and deepfakes in the age of AI" en la revista Cyber Frontiers de NTT DATA, edición Enero 2026. Se reproduce con autorización.*

A medida que la IA difumina la línea entre lo real y lo falso, las organizaciones deben enfrentarse a una nueva ola de engaño digital que amenaza la confianza, la seguridad y la propia noción de la verdad.

La confianza humana sustenta muchas interacciones digitales críticas con organizaciones y entidades gubernamentales. Sin embargo, esta confianza está bajo amenaza, ya que la IA facilita la creación de *deepfakes* que suplantan a personas reales mediante audio y vídeo digitales con una precisión sorprendente. Lo que antes requería estudios especializados y recursos considerables, ahora está al alcance de actores maliciosos con recursos informáticos cotidianos y conocimientos técnicos mínimos.

Lo que estamos viendo es solo la punta del iceberg, ya que la magnitud del problema está creciendo de forma exponencial.

Para las agencias gubernamentales, las implicaciones en materia de seguridad nacional son profundas. El Informe de Riesgos Globales 2024 del Foro Económico Mundial clasifica la desinformación y la información errónea impulsadas por la IA como la principal amenaza que el mundo enfrentará en los próximos dos años.

Desde una perspectiva de ciberseguridad, los *deepfakes* representan un nuevo vector de amenaza. Los marcos de seguridad tradicionales, centrados en el acceso a sistemas y la protección de datos, no están diseñados para identificar el engaño basado en contenidos. Esto crea un punto ciego en los centros de operaciones de seguridad, dejando a las organizaciones vulnerables ante una nueva generación de ataques que apuntan a la confianza humana en lugar de a los sistemas técnicos.

El imperativo de la ciberseguridad

La detección de *deepfakes* es ahora un requisito fundamental de la ciberseguridad. La convergencia de la suplantación habilitada por IA con los vectores de ataque tradicionales genera amenazas compuestas que eluden las medidas de seguridad convencionales:

- Se está suplantando a ejecutivos en videollamadas para autorizar transferencias fraudulentas, lo que genera un riesgo financiero considerable.
- La clonación de voz se utiliza cada vez más para derrotar sistemas de autenticación biométrica, comprometiendo métodos de verificación que antes se consideraban seguros.
- El uso de medios sintéticos facilita campañas de *phishing* sofisticadas que son casi indistinguibles de las comunicaciones legítimas.
- Las pruebas manipuladas amenazan con comprometer procesos legales y regulatorios, socavando la integridad judicial.
- Los ataques coordinados de desinformación contra infraestructuras críticas plantean preocupaciones de seguridad nacional al erosionar la confianza pública y potencialmente interrumpir servicios esenciales.

Estas amenazas requieren capacidades especializadas de detección integradas directamente en los flujos de trabajo de las operaciones de seguridad. Sin protección en tiempo real contra los medios sintéticos, incluso los marcos de ciberseguridad más sólidos permanecen fundamentalmente incompletos.

Engaño sintético: *deepfakes* dirigidos a sectores críticos

Los *deepfakes* representan una amenaza poderosa en sectores de alto riesgo. En los servicios financieros, los atacantes utilizan voz y vídeo generados por IA para suplantar a clientes bancarios durante llamadas a centros de atención, eludiendo los procesos de verificación de identidad e iniciando transacciones fraudulentas que explotan vulnerabilidades en la verificación digital de identidad del cliente y en los sistemas de autenticación basados en voz.

Las agencias gubernamentales son objetivo de suplantaciones mediante *deepfakes* de funcionarios y de inteligencia falsificada, lo que plantea riesgos para la seguridad nacional y la confianza pública.

Las infraestructuras críticas, incluyendo energía, salud y servicios de emergencia, son vulnerables a campañas de desinformación impulsadas por *deepfakes* que pueden simular comunicaciones de crisis, interrumpir la continuidad operativa y confundir al público durante emergencias. En el sector de la aviación, los actores maliciosos podrían suplantar a pilotos, controladores de tráfico aéreo o ejecutivos de aerolíneas mediante medios sintéticos, lo que podría provocar retrasos en los vuelos y riesgos para la seguridad.

A medida que estas amenazas se vuelven más sofisticadas, las organizaciones necesitan invertir en detección de *deepfakes*, protocolos de comunicación seguros e inteligencia de amenazas intersectorial para salvaguardar la confianza y la resiliencia.

Contrarrestar la amenaza de los *deepfakes*: una estrategia de defensa multimodal

Para contrarrestar la forma en que los atacantes ahora combinan texto, audio, imágenes y vídeo para crear engaños convincentes, las defensas deben ser igualmente multifacéticas. Imaginemos el siguiente escenario: un alto ejecutivo recibe una llamada a altas horas de la noche. La voz al otro lado de la línea es inconfundible: pertenece al CEO, quien solicita con urgencia una transferencia bancaria. Pero se trata de una voz sintética, clonada con una precisión inquietante. Aquí es donde entra en juego la pericia forense de audio. Algoritmos avanzados de detección analizan sutiles inconsistencias —pausas antinaturales, anomalías de frecuencia y patrones de respiración— para desenmascarar la falsificación. Incluso cuando la voz suena auténtica, el sistema sabe reconocer la diferencia.

Ahora imaginemos una videoconferencia en la que un rostro familiar transmite instrucciones. Sin embargo, detrás de los píxeles se esconde una falsificación. Las herramientas de detección de *deepfakes* en vídeo examinan microexpresiones faciales, patrones de parpadeo y señales de comportamiento que delatan un origen sintético. Estas herramientas actúan como detectores digitales de mentiras que protegen los canales de comunicación visual frente a la manipulación. Pero la detección por sí sola no es suficiente.

También es fundamental una respuesta en tiempo real. Los sistemas de seguridad modernos integran motores de detección de *deepfakes* que operan de manera continua, señalando contenido sospechoso en el momento en que aparece.

Las alertas se clasifican según su nivel de gravedad, de modo que las amenazas de alto riesgo se escalen sin sobrecargar a los equipos de seguridad. Cada incidente se registra con metadatos detallados —marcas de tiempo, datos de origen y puntuaciones de anomalía— para crear un rastro forense que facilite la investigación y el cumplimiento normativo.

Además, estos sistemas aprenden mediante sus capacidades integradas de auditoría, analizando patrones entre incidentes y ayudando a las organizaciones a fortalecer sus defensas con el tiempo.

Ya sea para prevenir ataques de ingeniería social o para proteger la integridad de las comunicaciones digitales, el objetivo es claro: restaurar la confianza en lo que vemos y escuchamos. En la batalla contra el engaño sintético, una defensa multimodal e inteligente no es opcional, sino una necesidad.

Preparar las operaciones de seguridad para el futuro: mantenerse a la vanguardia de la curva sintética

La amenaza de los *deepfakes* evoluciona con cada avance en las capacidades de la IA. Lo que hoy parece innovador puede volverse común mañana. En este entorno cambiante, las organizaciones que actúan ahora sientan las bases para la resiliencia frente a lo que está por venir.

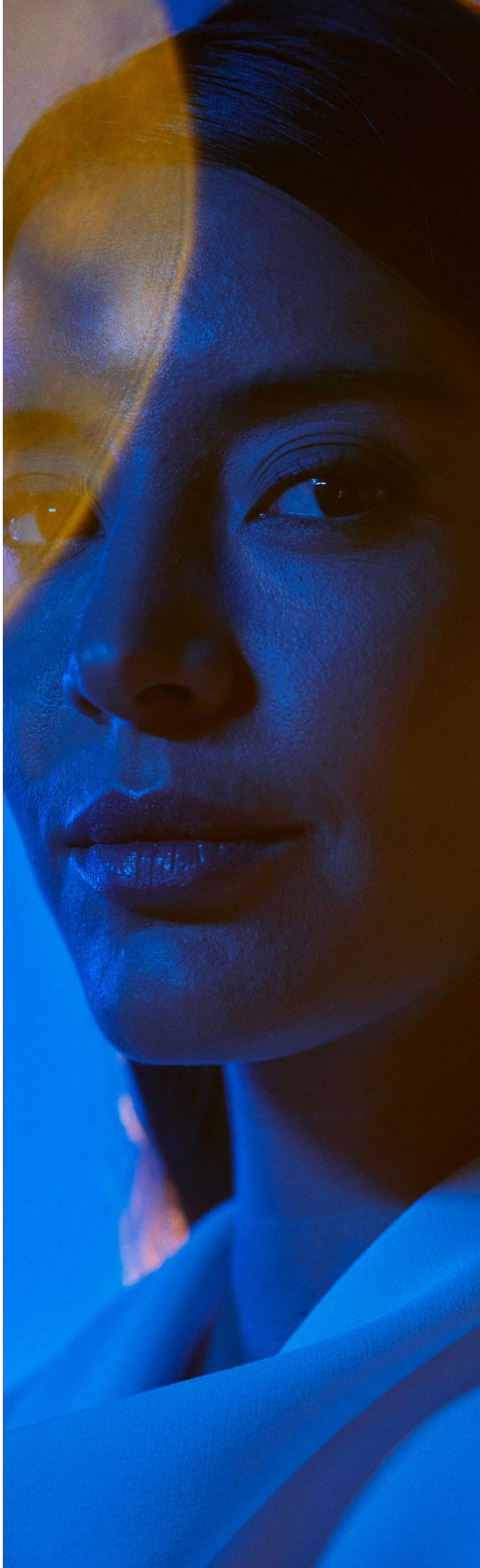
Pensemos en la detección de *deepfakes* como una transformación estratégica. Al integrar estas herramientas en el núcleo de las operaciones de seguridad, las organizaciones y los gobiernos están redefiniendo cómo se verifica la verdad en la era digital. Ya no es suficiente confiar en lo que vemos o escuchamos; debemos verificarlo mediante sistemas impulsados por inteligencia que se adapte tan rápido como las amenazas que contrarrestan.

Este cambio va más allá de la tecnología: también se trata de confianza. Proteger los canales de comunicación frente a la suplantación impulsada por IA garantiza que las decisiones críticas se tomen con información auténtica. Preserva la integridad del liderazgo, la continuidad de las operaciones y la confianza de las partes interesadas.

En un mundo donde los medios sintéticos son cada vez más convincentes, la defensa contra los *deepfakes* se convierte en un pilar fundamental de la confianza digital.

Prakash Narayanamoorthy
Global Capability Leader: Emerging Technology Security at NTT DATA

Ben Colman
Co-Founder and CEO at Reality Defender



De la regulación a la práctica: cómo las organizaciones deben gobernar la Inteligencia Artificial

Artículo por Julissa Calderón Loayza y Melanie Brenis Valencia

El desarrollo y la adopción de la Inteligencia Artificial (IA) avanzan a un ritmo cada vez más acelerado, generando beneficios claros en eficiencia, automatización y toma de decisiones, pero también nuevos riesgos relacionados con la seguridad, la privacidad y la transparencia. A medida que estos sistemas se incorporan en procesos cada vez más críticos de las organizaciones, distintos Estados han intensificado sus esfuerzos por definir marcos regulatorios que permitan gestionar dichos riesgos y aportar mayor seguridad jurídica. En este contexto, resulta relevante analizar el estado actual de la regulación de la IA, ofreciendo a continuación un panorama general de su evolución en distintos países de América y Europa.

Europa: el primer marco vinculante de IA

En Europa, la regulación de la IA dio un paso decisivo con la aprobación del **Reglamento Europeo de Inteligencia Artificial (EU AI Act)** en 2024, cuya aplicación es progresiva desde 2025. El reglamento es directamente aplicable en los 27 Estados miembros de la Unión Europea, y constituye el primer marco integral y vinculante sobre IA a nivel regional. Sus disposiciones aplican a organizaciones públicas y privadas que desarrollen, comercialicen o utilicen sistemas de IA en el mercado europeo, incluso cuando estén establecidas fuera de la UE, siempre que dichos sistemas tengan impacto en su territorio.

A nivel general, el EU AI Act establece un modelo regulatorio basado en el nivel de riesgo del sistema de IA y fija, entre otras, las siguientes obligaciones clave:

- Prohíbe ciertos usos de IA considerados inaceptables por su impacto en los derechos fundamentales (ejemplo: la categorización biométrica sensible).
- Exige controles estrictos para los sistemas de IA de alto riesgo, incluyendo la implementación de un sistema de gestión de riesgos y supervisión humana.
- Impone obligaciones de transparencia, especialmente cuando los sistemas de IA interactúan con personas o generan contenido sintético (ejemplo: *deepfakes*).
- Define responsabilidades claras para proveedores, distribuidores y usuarios de IA, estableciendo mecanismos de control por parte de las autoridades.

América: interés regulatorio creciente

En América, a diferencia de Europa, la regulación de la IA sigue siendo incipiente. La mayoría de los países no cuenta con leyes específicas vigentes, aunque se observa un creciente interés normativo reflejado en proyectos de ley en países como Chile, Brasil, México, Colombia, Argentina y Ecuador. Estas iniciativas buscan establecer principios éticos, proteger derechos fundamentales y definir obligaciones básicas de transparencia, pero en su mayoría permanecen en fases tempranas y sin mecanismos claros de supervisión, lo que genera incertidumbre regulatoria.

Dentro de este panorama, los avances más relevantes se concentran en Estados Unidos y Perú. En Estados Unidos, aunque aún no existe una ley federal integral sobre IA, varios Estados han aprobado regulaciones ya vigentes, como el **Colorado AI Act**, el **Utah AI Policy Act** y el **Texas Responsible AI Governance Act**, que introducen obligaciones concretas en materia de uso responsable, transparencia y gobernanza de sistemas de IA. Este enfoque, si bien fragmentado, resulta operativo y refleja un avance regulatorio a nivel estatal. Por su parte, Perú ha sido uno de los primeros países de América Latina en desarrollar un marco normativo aplicado de forma efectiva: aprobó su **Ley Nº 31814** en 2023, que promueve el uso de la IA para el desarrollo económico y social, y posteriormente publicó el **Decreto Supremo Nº 115-2025-PCM**, estableciendo disposiciones claras para su implementación efectiva.

A grandes rasgos, el marco peruano -ya vigente- introduce obligaciones, entre las que destacan:

- Clasificación de los sistemas de IA según su nivel de riesgo y finalidad, exigiendo evaluaciones previas de impacto y definiendo el grado de supervisión aplicable antes de su despliegue.
- Exigencia de evaluaciones de impacto y análisis de riesgos para el uso de IA en el sector público, considerando efectos en derechos fundamentales, seguridad y posibles sesgos.
- Requisitos de trazabilidad y documentación de los sistemas de IA, que permitan explicar su funcionamiento, decisiones y resultados ante autoridades competentes.

- Referencia expresa a estándares y buenas prácticas internacionales, como ISO/IEC 42001, para orientar la gestión, la monitorización continua y la mejora de los sistemas de IA.

El panorama descrito muestra que la regulación de la IA avanza a distintos ritmos, pero con una tendencia común: la necesidad de controlar los riesgos asociados a su uso. Ante ello, las organizaciones no pueden limitarse a reaccionar a nuevas normas, sino que deben comprender en qué estado se encuentran respecto al uso de IA. Contar con un diagnóstico claro, identificar riesgos y brechas e implementar medidas adecuadas resulta clave para establecer un modelo de gobierno de la IA que complemente la regulación y permita anticipar futuras exigencias normativas.

El Gobierno de la IA: de la regulación a la gestión efectiva del riesgo

El avance regulatorio descrito anteriormente confirma una realidad: la Inteligencia Artificial ya no es solo un tema tecnológico, sino un asunto estratégico, legal y de gestión de riesgos. Sin embargo, la regulación por sí sola no garantiza un uso responsable. El verdadero desafío para las organizaciones es traducir esas exigencias en un modelo de gobierno interno que permita controlar riesgos, asegurar cumplimiento y sostener la confianza.

Hoy, una parte significativa de las organizaciones ya utiliza IA en procesos críticos o se encuentra en fases avanzadas de adopción. Varios reportes y estudios internacionales coinciden en que, aunque la inversión en IA crece de forma sostenida, la madurez en gobierno, gestión de riesgos y control sigue siendo limitada. En muchos casos, las iniciativas surgen de manera descentralizada, con modelos o soluciones implementadas sin una estructura clara de supervisión, trazabilidad o responsabilidad formal. Es en este escenario donde surge la necesidad del Gobierno de la IA.

¿Qué implica gobernar la IA?

Gobernar la IA no significa frenar la innovación, sino habilitarla de forma segura. Supone establecer un marco estructurado que acompañe todo el ciclo de vida del sistema, desde el diseño y entrenamiento hasta su monitorización y eventual retiro, asegurando transparencia, seguridad, privacidad y cumplimiento normativo.

Un modelo sólido de gobierno de IA suele apoyarse en cinco pilares fundamentales:

1) Definición clara de roles y responsabilidades:

Es indispensable delimitar quién es responsable de los datos, del modelo, del control de riesgos y de la seguridad. La ambigüedad organizacional es uno de los principales factores de exposición. Sin una asignación clara de responsabilidades, los riesgos se diluyen y las decisiones críticas carecen de supervisión adecuada.

La estructura de gobierno dependerá del nivel de madurez digital y organizacional. En compañías con mayor desarrollo en analítica e innovación, puede existir incluso un Chief AI Officer (CAIO) con responsabilidad transversal sobre estrategia, riesgos y supervisión de la IA. En otros casos, la función recae en el Chief Data Officer (CDO), cuando la IA está fuertemente ligada al gobierno del dato, o en el CIO/CTO, cuando el enfoque es más tecnológico. Adicionalmente, muchas organizaciones están conformando Comités de IA interdisciplinarios, donde participan líderes de tecnología, datos, riesgo, cumplimiento, legal y ciberseguridad, e incluso miembros del *board* o alta gerencia en casos de mayor exposición estratégica. Este enfoque permite alinear innovación con apetito de riesgo y supervisión ejecutiva.

2) Gestión de riesgos desde el diseño:

La IA debe evaluarse bajo un enfoque basado en riesgos. Esto incluye no solo riesgos tradicionales de tecnología, sino también riesgos emergentes como sesgos algorítmicos, falta de explicabilidad, vulnerabilidades ante ataques de adversarios, fuga de información, impacto reputacional y posibles afectaciones a derechos fundamentales. La gestión debe ser continua, no puntual.

3) Integración con el gobierno del dato y la ciberseguridad:

No existe IA confiable sin datos gobernados ni sin controles de seguridad robustos. La calidad, legitimidad y clasificación de los datos son determinantes. Asimismo, la IA amplía la superficie de ataque, lo que exige controles específicos frente a amenazas como manipulación de modelos, extracción de información o uso indebido de herramientas generativas por parte de empleados o terceros.

4) Principios corporativos de IA responsable:

Más allá del cumplimiento normativo, las organizaciones deben definir sus propios principios de uso responsable (transparencia, supervisión humana, no discriminación, seguridad y rendición de cuentas) e integrarlos en su cultura y procesos de decisión.

5) Integración con el gobierno corporativo existente:

El gobierno de la IA no debe operar como una estructura paralela, sino integrarse con el gobierno del dato, la gestión de riesgos empresariales, el cumplimiento y la ciberseguridad. Esta integración evita duplicidades y fortalece la coherencia estratégica.

Confianza como ventaja competitiva

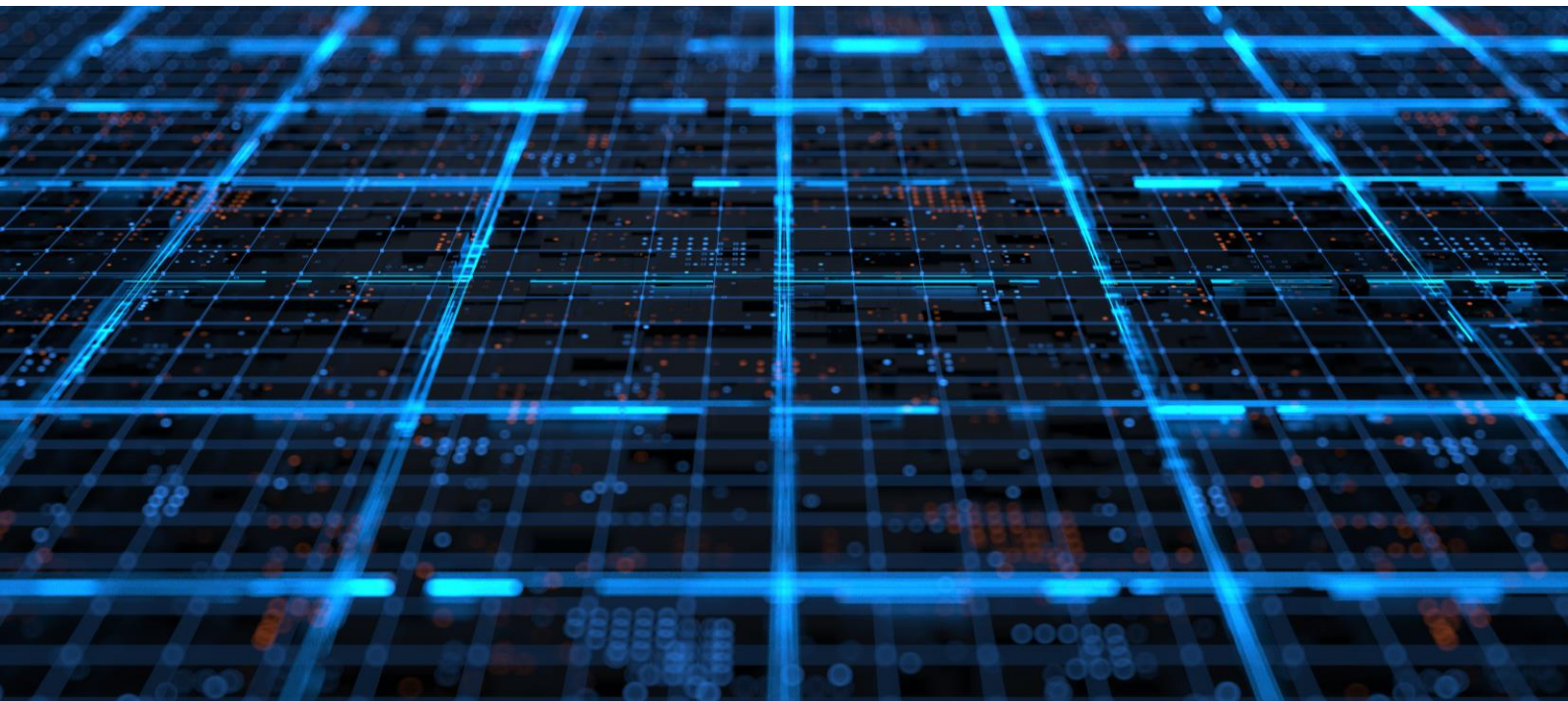
La evolución de la IA no puede limitarse a la dimensión tecnológica. Las organizaciones que logren estructurar un modelo de gobierno sólido estarán mejor preparadas para responder a reguladores, mitigar incidentes y sostener la confianza de clientes y *stakeholders*.

En un entorno donde la regulación avanza y la exposición reputacional es inmediata, el diferencial competitivo no será únicamente quién innove más rápido, sino quién lo haga de forma más segura, transparente y responsable.

El Gobierno de la IA se convierte así en el puente entre la innovación y la resiliencia organizacional.

Julissa Calderón Loayza
Cybersecurity Expert Associate

Melanie Brenis Valencia
Cybersecurity Senior Consultant



Ciberseguridad ofensiva y las tendencias de Agentic AI: implicaciones para el Red Team moderno

Tendencias por Marco Antonio Andazabal Rebaza

La evolución de la Inteligencia Artificial (IA) ha transformado la interacción entre usuarios y tecnología, pasando de ser una herramienta de apoyo analítico a sistemas capaces de generar contenido, automatizar procesos y asistir en tareas técnicas especializadas. Esta democratización ha disminuido la barrera de entrada a capacidades avanzadas, permitiendo que perfiles no especializados accedan a funcionalidades antes reservadas a expertos. No obstante, esta accesibilidad también plantea desafíos en ciberseguridad. En el ámbito ofensivo, la aparición de sistemas de Agentic AI introduce un nuevo paradigma: modelos capaces de planificar, ejecutar y ajustar acciones de manera autónoma para alcanzar objetivos definidos.

De la IA generativa a la Agentic AI

1) IA generativa y automatización asistida

La IA generativa, basada principalmente en modelos de lenguaje de gran escala (LLMs), se caracteriza por su capacidad de producir contenido, texto, código, configuraciones o análisis técnicos en respuesta a instrucciones explícitas. Su funcionamiento es esencialmente reactivo: responde a un *prompt* específico sin mantener autonomía operativa más allá de la interacción inmediata.

En el contexto de la ciberseguridad ofensiva, su aplicación se centra en tareas de apoyo, tales como generación de *scripts*, creación de *payloads*, redacción de correos para simulaciones de *phishing*, etc.

No obstante, estos sistemas requieren supervisión constante. Cada paso del proceso ofensivo (reconocimiento, explotación, post-explotación) debe ser dirigido manualmente por el operador. La IA no toma decisiones estratégicas por sí misma, sino que ejecuta tareas delimitadas por instrucciones humanas.

2) Agentic AI: sistemas orientados a objetivos y autonomía operativa

La Agentic AI o IA agéntica representa una evolución estructural respecto a los modelos generativos tradicionales. En lugar de limitarse a responder instrucciones puntuales, estos sistemas están diseñados para alcanzar objetivos definidos mediante procesos autónomos de planificación y ejecución.

Desde una perspectiva teórica, un sistema agéntico integra:

- Planificación ”multi-paso”: descompone un objetivo complejo en subtareas.
- Memoria contextual persistente: mantiene estado e historial de acciones.
- Ejecución autónoma: interactúa con herramientas externas (APIs, sistemas operativos, escáneres).
- Evaluación y retroalimentación: analiza resultados y ajusta su estrategia.

Arquitectónicamente, combina un modelo de lenguaje con un ciclo de decisión iterativo (percepción → planificación → acción → evaluación). Esto le permite operar como un “agente” dentro de un entorno técnico, más que como un asistente pasivo.

Transformación del Red Team en la era de la Agentic AI

1) Automatización avanzada del reconocimiento

En operaciones de Red Team, la fase de reconocimiento (*recon*) es importante. Normalmente, implica recopilación manual o semiautomatizada de información mediante herramientas OSINT, escáneres de red y análisis de superficies expuestas.

Un sistema de Agentic AI podría:

- Definir subobjetivos (identificar dominios, subdominios, servicios expuestos).
- Orquestar múltiples herramientas de escaneo.
- Correlacionar resultados.
- Priorizar vectores de ataque según probabilidad de éxito.
- Ajustar la estrategia en función de hallazgos intermedios.

Este comportamiento se asemeja más al de un operador humano asistido por automatización inteligente que a un simple *script* estático.

Generación dinámica de *exploits* y pruebas de concepto

Si bien la generación de *exploits* complejos aún requiere conocimiento especializado, la IA puede asistir en:

- Adaptación de *exploits* públicos a contextos específicos.
- Modificación de *payloads* según restricciones del entorno.
- Generación de *scripts* para pruebas de concepto (PoC).
- Análisis preliminar de código vulnerable.

En un escenario controlado de *pentesting*, esto puede reducir significativamente el tiempo entre la identificación de una vulnerabilidad y su validación técnica.

No obstante, la preocupación radica en que actores maliciosos podrían emplear capacidades similares para escalar ataques con menor experiencia técnica.

Ingeniería social asistida por IA

Uno de los impactos más relevantes se observa en la ingeniería social. La Agentic AI puede:

- Analizar perfiles públicos.
- Generar correos altamente personalizados.
- Ajustar el tono según el rol organizacional.
- Simular conversaciones coherentes en múltiples iteraciones.

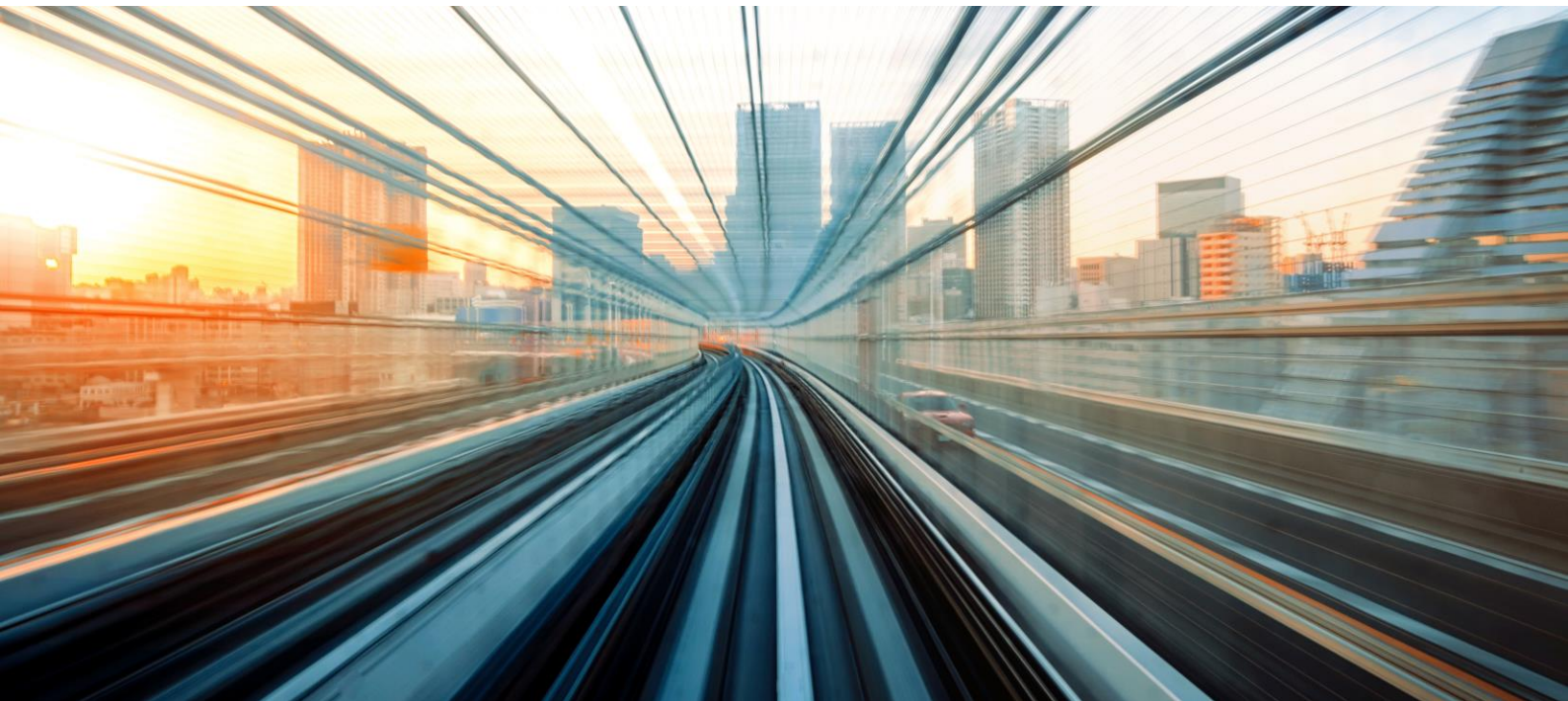
Esto incrementa la tasa potencial de éxito en campañas de *phishing* dirigidas (*spear phishing*), reduciendo la necesidad de habilidades avanzadas en manipulación psicológica.

La convergencia entre ciberseguridad ofensiva y Agentic AI redefine las dinámicas tradicionales del Red Team y del *pentesting* técnico. La transición desde modelos reactivos hacia agentes autónomos orientados a objetivos introduce un cambio estructural en la forma en que se ejecutan y escalan las operaciones ofensivas.

Más que una simple herramienta de automatización, la Agentic AI representa un catalizador estratégico capaz de amplificar tanto la eficiencia defensiva como el potencial ofensivo. En este contexto, el desafío no reside únicamente en comprender la tecnología, sino en anticipar sus implicaciones operativas, éticas y regulatorias.

El futuro de la ciberseguridad no dependerá exclusivamente de la capacidad técnica, sino de la gobernanza inteligente de sistemas cada vez más autónomos.

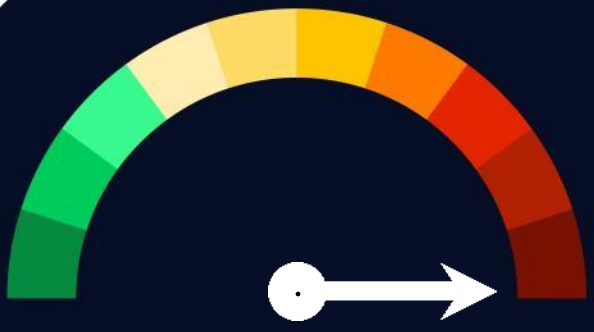
Marco Antonio Andazabal Rebaza
Cybersecurity Analyst



Vulnerabilidades

Vulnerabilidades en Cisco Secure Firewall Management Center

Fecha: 4 de marzo de 2026
CVE: CVE-2026-20079 y CVE-2026-20131



CVSS: 10.0

CRÍTICA

Descripción

Se han identificado dos vulnerabilidades críticas, catalogadas como CVE-2026-20079 y CVE-2026-20131, presentes en Cisco Secure Firewall Management Center.

La primera vulnerabilidad (CVE-2026-20079) se debe a un proceso incorrecto que se crea al arrancar el sistema. Un atacante podría explotar este fallo mediante el envío de una solicitud HTTP maliciosa. Esto podría permitir al atacante ejecutar *scripts* y comandos en el sistema afectado y obtener permisos de *root*.

La segunda vulnerabilidad (CVE-2026-20131) consiste en la deserialización insegura de un flujo de *bytes* de Java proporcionados por el usuario. Un atacante podría enviar un objeto Java serializado para ejecutar código en el dispositivo y obtener permisos de *root*.

Solución

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

Para ello, Cisco recomienda a sus clientes utilizar el comprobador de *software* de Cisco y seguir sus instrucciones.

Productos afectados

Los productos afectados son los siguientes:

- Cisco Secure FMC Software
- Cisco Security Cloud Control (SCC) Firewall Management

Referencias

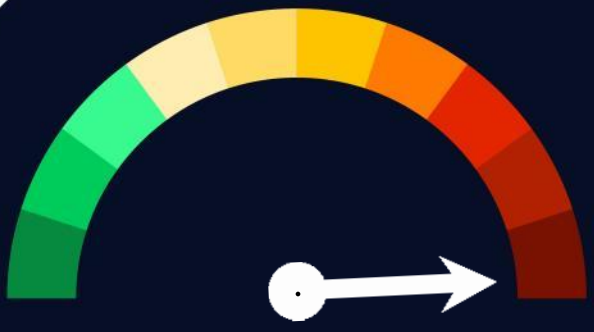
- sec.cloudapps.cisco.com
- sec.cloudapps.cisco.com
- incibe.es

TLP:WHITE

Vulnerabilidades

Vulnerabilidad crítica en Android (System)

Fecha: 2 de marzo de 2026
CVE: CVE-2026-0006



CVSS: 9.8

CRÍTICA

Descripción

Se ha identificado una vulnerabilidad crítica de ejecución remota en dispositivos Android con diferentes versiones.

Esta vulnerabilidad (**CVE-2026-0006**) permite a un atacante ejecutar código malicioso sin necesidad de que el usuario interactúe con el dispositivo sin requerir privilegios adicionales, debido a que la vulnerabilidad reside en un componente del System de Android.

La explotación exitosa podría permitir el control remoto total, exfiltración de datos, espionaje en tiempo real, instalación de *malware* persistente y movimiento lateral.

Solución

Se recomienda:

- Instalar la actualización correspondiente al nivel de parche 2026-03-01 o posterior. Los fabricantes (Google, Samsung, Xiaomi, etc) están desplegando estas actualizaciones vía OTA (On The Air).

Productos afectados

Están afectados por esta vulnerabilidad, los dispositivos con las versiones de Android: 12, 12L, 13, 14 y 15

Referencias

- source.android.com
- nvd.nist.gov

Android corrige 129 vulnerabilidades en su parche de seguridad de marzo

Fecha: 2 de marzo de 2026
CVE: CVE-2026-21385 y 128 más

Crítica

Descripción

El boletín de seguridad de Android de agosto corrige un total de 129 vulnerabilidades, entre ellas 10 de severidad crítica y 106 de severidad alta. Mediante la explotación de estas vulnerabilidades, un atacante podría provocar una escalada de privilegios, la ejecución remota de código o un desbordamiento del *búfer*.

Android indica que la vulnerabilidad con identificador CVE-2026-21385 puede estar bajo explotación limitada. Esta vulnerabilidad se encuentra en un componente de pantalla de Qualcomm. La empresa indicó en un comunicado que el fallo se encuentra en el subcomponente Graphics, y consiste en un desbordamiento de enteros que un atacante puede explotar para provocar una corrupción de memoria en el dispositivo.

Productos afectados

Los productos afectados por esta vulnerabilidad incluyen:

- Versiones de Android anteriores a 2026-03-01.

Solución

Actualizar los productos afectados a la última versión disponible.

Referencias

- source.android.com
- docs.qualcomm.com

Google corrige de manera urgente 10 vulnerabilidades

Fecha: 3 de marzo de 2026
CVE: CVE-2026-3536 y 9 más

Alta

Descripción

Google ha lanzado un parche de seguridad urgente para Chrome con el fin de solventar 10 vulnerabilidades entre las que se encuentran 3 de carácter crítico y 7 altas.

Entre los problemas detectados destacan los de corrupción de memoria, errores de implementación en motores gráficos y JavaScript, que podrían permitir ejecutar código arbitrario o exfiltración de información, entre otros.

Las vulnerabilidades más críticas (CVE-2026-3536, CVE-2026-3537 y CVE-2026-3538) relacionadas con lagunas en la validación en los *códecs* web y de navegación podrían dificultar la identificación de *phishing* y las descargas no autorizadas.

Se recomienda a los usuarios y empresas aplicar las actualizaciones lo antes posible.

Productos afectados

Los productos afectados por esta vulnerabilidad incluyen todas las versiones de Google Chrome anteriores a la versión:

- 145.0.7632.159

Solución

Se recomienda:

- Actualizar a las últimas versiones del *software*.

Referencias

- cyberpress.org
- nvd.nist.gov

Eventos

IAPP Global Summit 2026
30 de marzo - 02 de abril

El IAPP Global Privacy Summit 2026 se celebrará del 30 de marzo al 2 de abril de 2026 en Washington, D.C., en formato presencial, consolidándose como uno de los encuentros clave en privacidad y ciberseguridad a nivel global. El evento reunirá a líderes del sector público y privado para debatir sobre regulación internacional, gobernanza de datos, riesgos cibernéticos emergentes y el impacto real de la inteligencia artificial en la protección de la información. Un espacio estratégico para anticipar tendencias y desafíos regulatorios.

Enlace

Gen AI Summit
17 - 18 de abril

El Gen AI Summit EU 2026 se celebrará del 17 al 18 de abril de 2026 en Valencia, España, en un formato presencial de dos días diseñado para profesionales de IA, datos y *machine learning*. El encuentro reunirá a líderes tecnológicos, innovadores y equipos técnicos para explorar el potencial transformador de la inteligencia artificial generativa. La agenda incluye charlas técnicas, paneles sobre ética, gobernanza y seguridad, y espacios de *networking* y *workshops* prácticos.

Enlace

CYBERUK 2026
21 - 23 de abril

CYBERUK 2026 se realizará del 21 al 23 de abril de 2026 en el Scottish Event Campus (SEC) de Glasgow, Reino Unido, en formato presencial. Organizado por el National Cyber Security Centre (NCSC), es la conferencia insignia del gobierno británico para profesionales de ciberseguridad, con más de 100 ponentes expertos. El programa aborda el tema “The Next Decade: Accelerating Our Cyber Defence”, explorando estrategias, amenazas emergentes, defensa crítica y colaboración público-privada. También contará con zonas de exposición, *networking* y sesiones técnicas especializadas.

Enlace

Cyber Intelligence Europa 2026
22 - 23 de abril

Cyber Intelligence Europe 2026 se celebrará del 22 al 23 de abril de 2026 en Bruselas, Bélgica, en formato presencial de dos días. Organizado con foco en la cooperación europea, reunirá a representantes de gobiernos, fuerzas armadas y agencias de seguridad para debatir estrategias nacionales de ciberseguridad y políticas públicas. El programa incluye análisis de tendencias en ciberdelitos, monitorización de amenazas y preparación ante ciberataques, con especial atención a la protección de infraestructuras críticas. El evento también explorará enfoques coordinados de intercambio de datos y respuestas conjuntas a grandes amenazas.

Enlace

Recursos

- **Integrating Cybersecurity and Enterprise Risk Management (ERM) - NIST**

Publicación del NIST que orienta a las organizaciones en la integración de la ciberseguridad dentro del marco de Gestión de Riesgos Empresariales (ERM), alineando la identificación, evaluación y priorización de riesgos cibernéticos con los objetivos estratégicos del negocio. El documento explica cómo utilizar registros de riesgos para mejorar la toma de decisiones a nivel ejecutivo y fortalecer la comunicación entre áreas técnicas y la alta dirección, promoviendo una gobernanza más sólida y mayor resiliencia frente a amenazas digitales.

Enlace

- **The ENISA Cybersecurity Exercise Methodology - ENISA**

Publicación de ENISA que ofrece un marco teórico completo para planificar, ejecutar y evaluar ejercicios de ciberseguridad eficaces desde el inicio hasta el final, asegurando que los perfiles y partes interesadas adecuadas participen en el momento correcto. Se apoya en lecciones aprendidas, mejores prácticas de la industria y experiencia en ciberseguridad, junto con herramientas y plantillas prácticas para organizar y mejorar simulaciones y pruebas de respuesta ante incidentes.

Enlace

- **Compendio sobre Protección de Datos Personales: normativa y Criterios Interpretativos Relevantes**

Documento oficial publicado por el Ministerio de Justicia y Derechos Humanos del Perú que compila de forma actualizada la Ley de Protección de Datos Personales Nº 29733, su reglamento y criterios interpretativos relevantes emitidos por la Autoridad Nacional de Protección de Datos Personales para orientar la aplicación práctica de la normativa; sirve como guía para entidades públicas y privadas en la gestión y cumplimiento del marco legal de protección de datos personales.

Enlace

Suscríbete a RADAR
up.nttdata.com/suscribetearadar

Powered by the cybersecurity NTT DATA team

es.nttdata.com