GBM

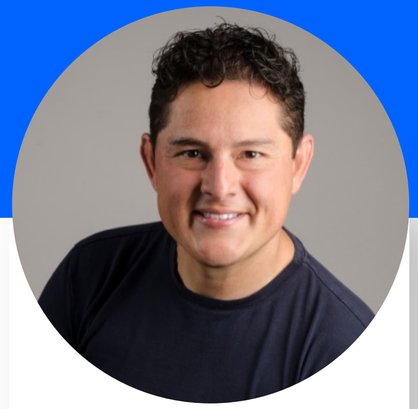
Cybersecurity Center

Alonso Ramírez

Posee las certificaciones Data Privacy Solutions Engineer, Lead Privacy Implementer, Identity Governance Expert, NSA4011, CEH, CHFI, CISM, CRISC, CISA, CRMA, CBCP, ISO27SLI, CLCM, CCNA Sec, PSEA, FJSE, NSE3.

Miembro de la Comisión de Ciberseguridad en Infocom. Miembro de la Junta Directiva del Clúster de Ciberseguridad Costa Rica.

Es asesor externo en materia de protección, privacidad y seguridad de datos corporativos, y es instructor de cursos de certificación de seguridad para PECB, expositor en seminarios y conferencias sobre seguridad de la información dentro y fuera del país.



alonsoramirezcybersecurity *laramirez@gbm.net* *+506 8827-3344*

Alonso

Regional Cyber Security Manager

Regional Cyber Security Manager en GBM Corporation y Profesor para las escuelas de Ingeniería de Sistemas, Postgrados y Maestrías en cursos de Ciberseguridad y Continuidad de Negocio de las universidades INCAE y CENFOTEC. Es Hacker Ético Certificado, Máster en Auditoría de Tecnologías de Información y Máster en ISO/IEC 27001.

Cuenta con el CNSS 4011 Recognition de la NSA (National Security Agency) donde certifica al profesional en seguridad de redes que tiene el conocimiento para trabajar en sector privado y público en Estados Unidos.

Posee más de 15 años de experiencia en ciberseguridad, es implementador y consultor en servicios tipo SOC y CSIRT. Desarrolla consultorías de prevención y defensa de amenazas cibernéticas avanzadas, así como de identificación de mejoras de seguridad informática corporativa, estrategia y tácticas para el diseño e implementación de arquitecturas de seguridad. Ha desempeñado el rol de Comandante de Incidentes en Ciberseguridad para la defensa de ataques cibernéticos contra infraestructuras críticas en la Región. Se ha desempeñado como Gerente Comercial y de Arquitectura de Ciberseguridad para las soluciones y servicios de IBM, Cisco y Microsoft.

Ciberseguridad en la era de la IA generativa

Alonso Ramírez
Regional Cyber Security Manager
GBM Cybersecurity
laramirez@gbm.net

Las empresas están adoptando la IA generativa, pero tienen preocupaciones

Bajo presión para adoptar

64%

enfrentan una presión significativa para acelerar las iniciativas de IA generativa

Preocupadas por nuevos riesgos

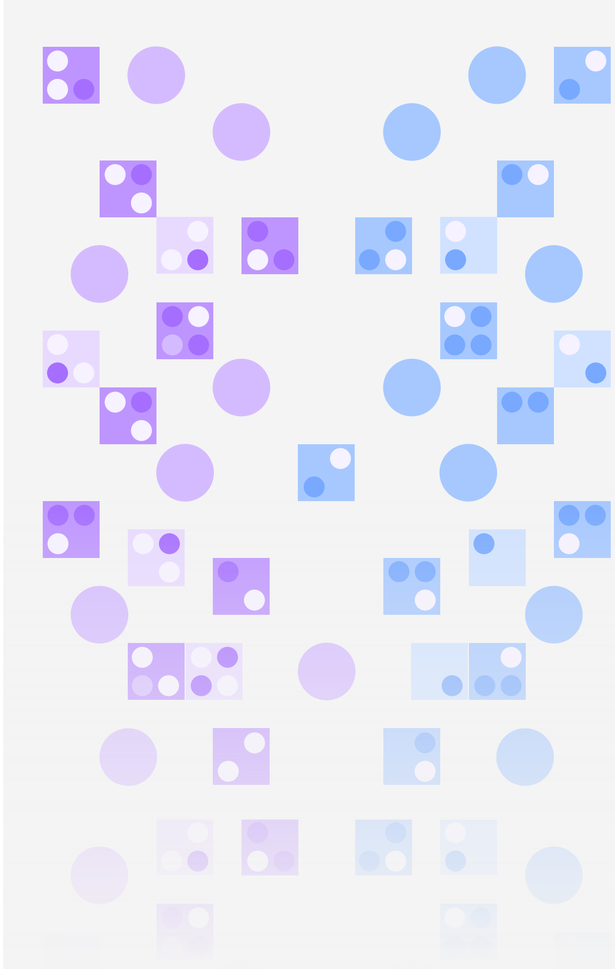
84%

ven el riesgo de ciberseguridad como el obstáculo número 1 para la adopción de IA generativa

Inversión en nuevas defensas

64%

identificaron la seguridad como la prioridad número 1 para los casos de uso de IA generativa



La IA generativa es la nueva plataforma urgente a asegurar ahora

Fuente: "IA generativa empresarial: Estado del mercado" Estudio del Instituto de Valor de Negocios de IBM 2023

Los atacantes apuntarán a la IA	La IA debe tratarse como **una nueva superficie de ataque**, con estrategias nuevas de detección y respuesta requeridas para la evasión de modelos, extracción, inferencia y envenenamiento.	La **inyección de comandos** puede derribar defensas previniendo la generación de material no deseado, además del acceso a integraciones explotables y una riqueza de datos de entrenamiento sensibles.	Modelos maliciosos pueden ser subidos a repositorios abiertos, con **comportamiento oculto activado** mucho después de que hayan sido desplegados.
Los atacantes utilizarán la IA	La IA generativa **escalará el cibercrimen** y reducirá las barreras de entrada para atacantes de menor habilidad.	El phishing se volverá **más dirigido**, y las técnicas generativas de video y audio necesitarán nuevos enfoques para evitar compromisos empresariales.	Los atacantes **se adaptarán** a las estrategias defensivas más rápidamente, y mejorarán la evasión de detección, el descubrimiento de vulnerabilidades y la personalización de malware.

Nuestro enfoque hacia la seguridad en la era de la IA generativa

Seguridad para la IA

Proteger los modelos base, la IA generativa y sus conjuntos de datos es esencial para una IA lista para la empresa

Asegurar los datos de entrenamiento de IA subyacentes protegiéndolos del robo de datos sensibles, la manipulación y las violaciones de cumplimiento

Asegurar el desarrollo del modelo escaneando en busca de vulnerabilidades en el pipeline, fortaleciendo las integraciones y haciendo cumplir políticas y accesos

Asegurar el uso de modelos de IA detectando fugas de datos o comandos, y alertando sobre ataques de evasión, envenenamiento, extracción o inferencia

IA para la Seguridad

Las ganancias de productividad de los modelos base y la IA generativa reducirán los cuellos de botella humanos en seguridad

La IA gestionará tareas de seguridad repetitivas como la suma de alertas y el análisis de registros, liberando a los equipos para abordar problemas estratégicos

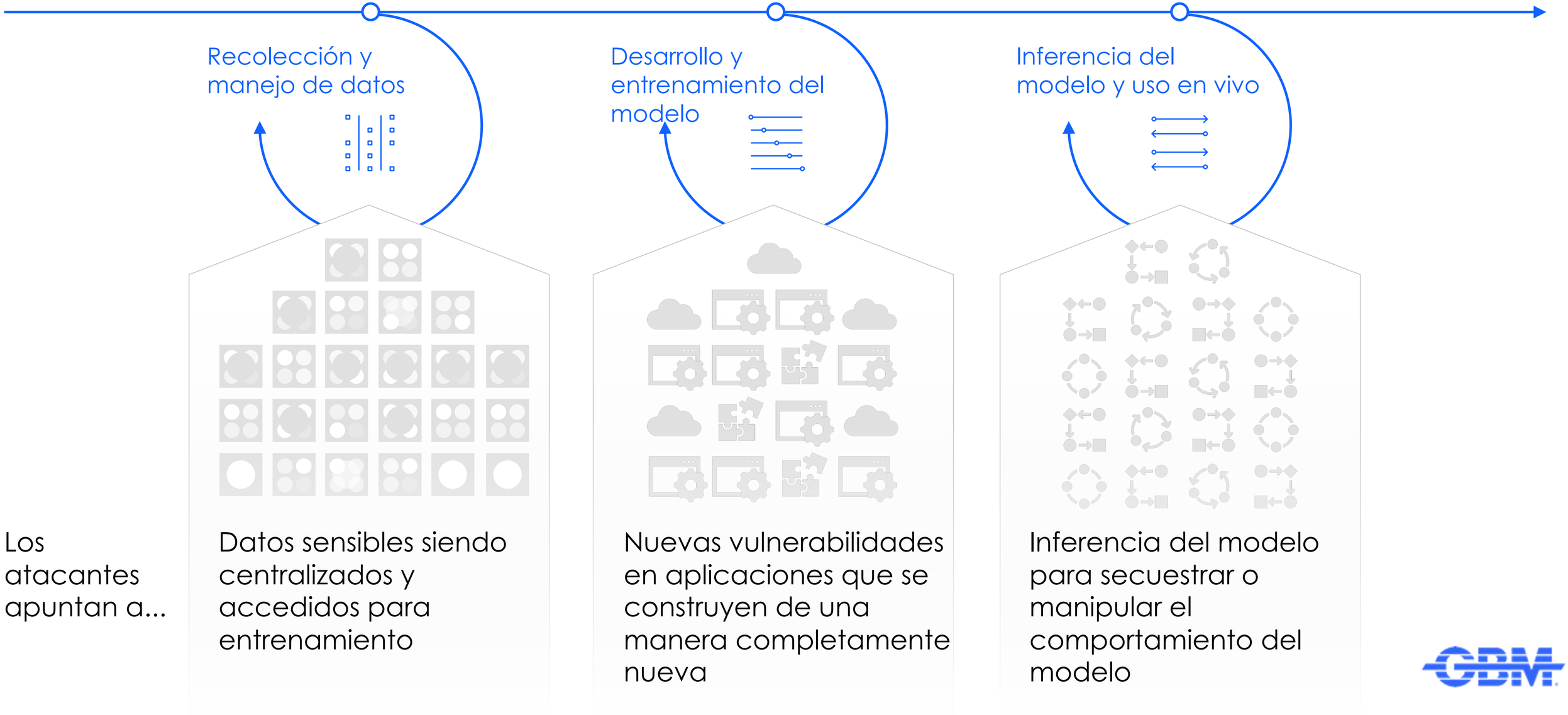
La IA generará contenido de seguridad (detecciones, flujos de trabajo, políticas) más rápido que los humanos, acelerando la implementación y ajustándose a las amenazas de seguridad cambiantes en tiempo real

La IA aprenderá y creará respuestas activas que se optimizan con el tiempo, con capacidades para encontrar todos los incidentes similares, actualizar sistemas afectados y parchear código vulnerable

Seguridad para la IA

Riesgos adversarios a lo largo del ciclo de vida de la IA

AI pipeline



Los atacantes apuntan a...

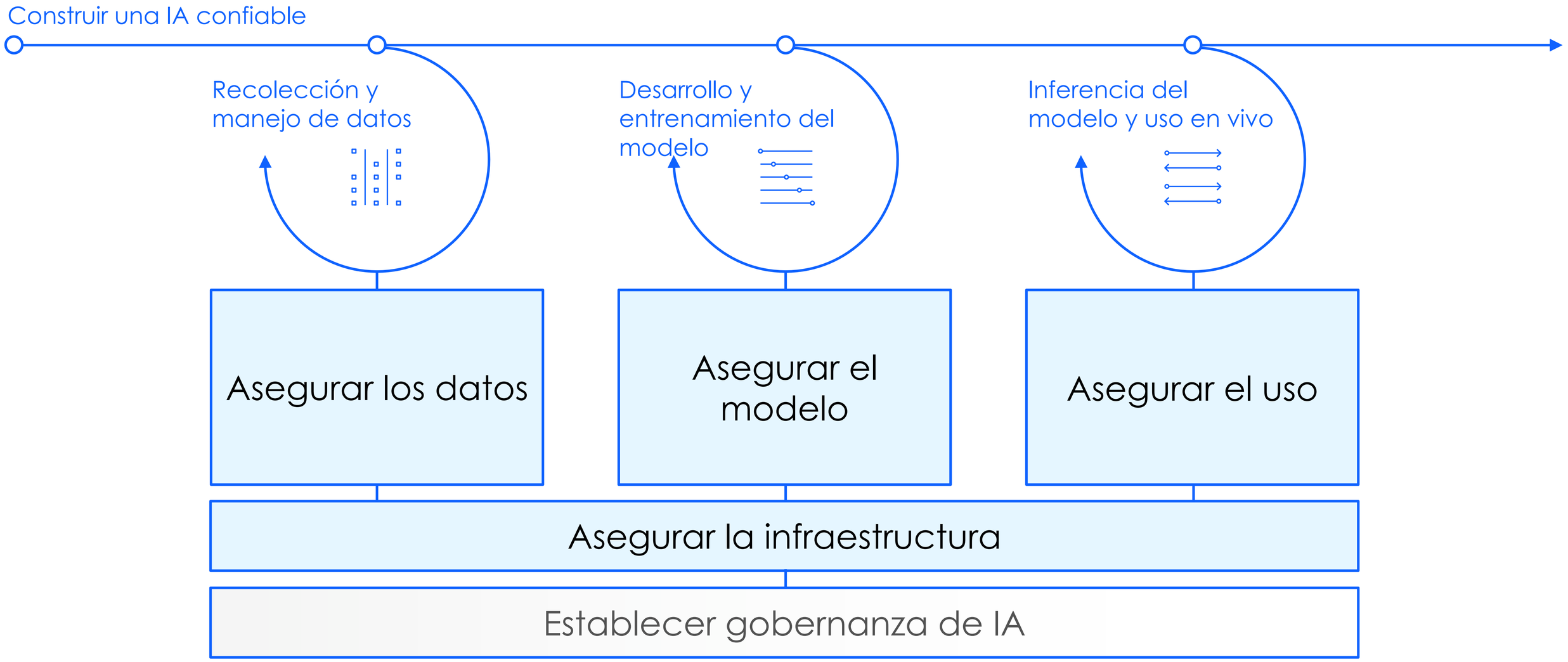
Datos sensibles siendo centralizados y accedidos para entrenamiento

Nuevas vulnerabilidades en aplicaciones que se construyen de una manera completamente nueva

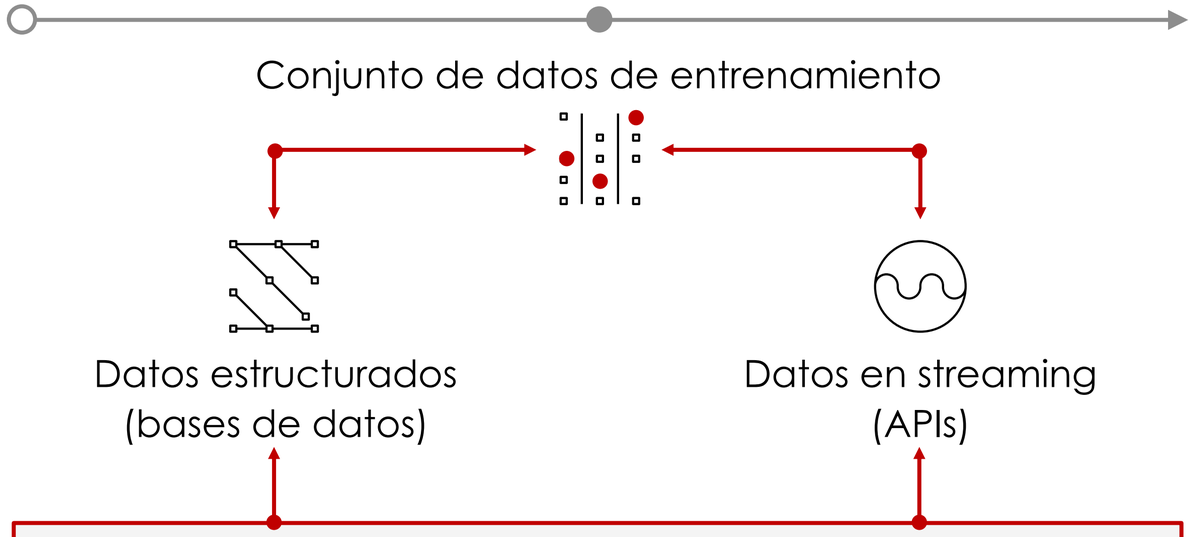
Inferencia del modelo para secuestrar o manipular el comportamiento del modelo

Lo que necesitas hacer

Marco de Seguridad para la IA



Riesgos en la recolección y manejo de datos

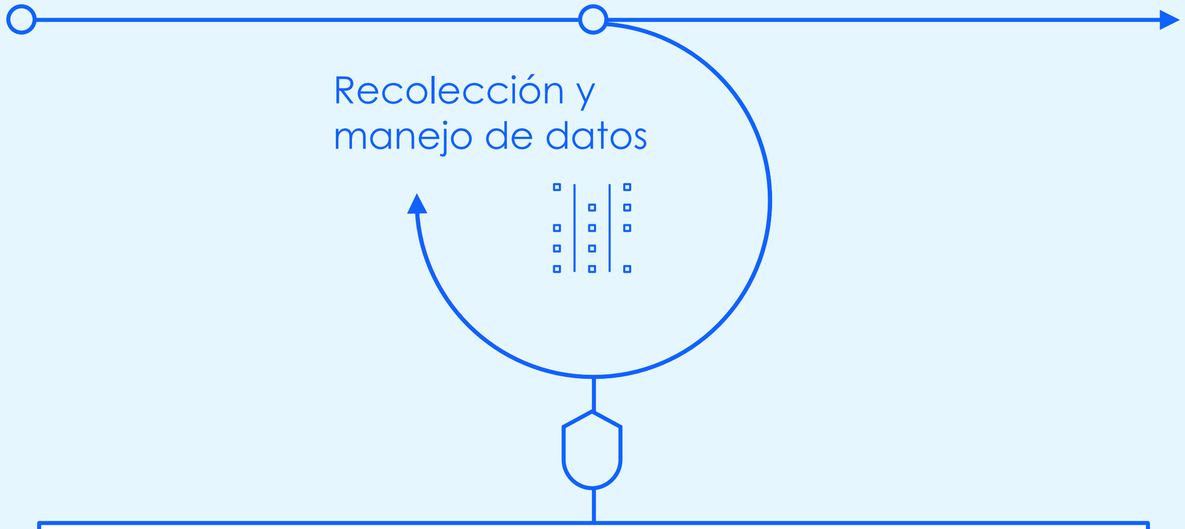


Los atacantes apuntan a los conjuntos de datos subyacentes

Exfiltración de datos:

- Los modelos de ML son intensivos en datos y consumen enormes cantidades de datos, incluyendo datos sensibles
- La fuga de datos puede resultar de una vulnerabilidad de seguridad técnica o controles de seguridad y acceso insuficientes
- Los atacantes pueden explotar vulnerabilidades o usar estafas de phishing para ganar acceso y robar datos sensibles utilizados en el entrenamiento y ajuste de modelos de ML

Mejores prácticas de seguridad en la recolección y manejo de datos



Asegurar los datos

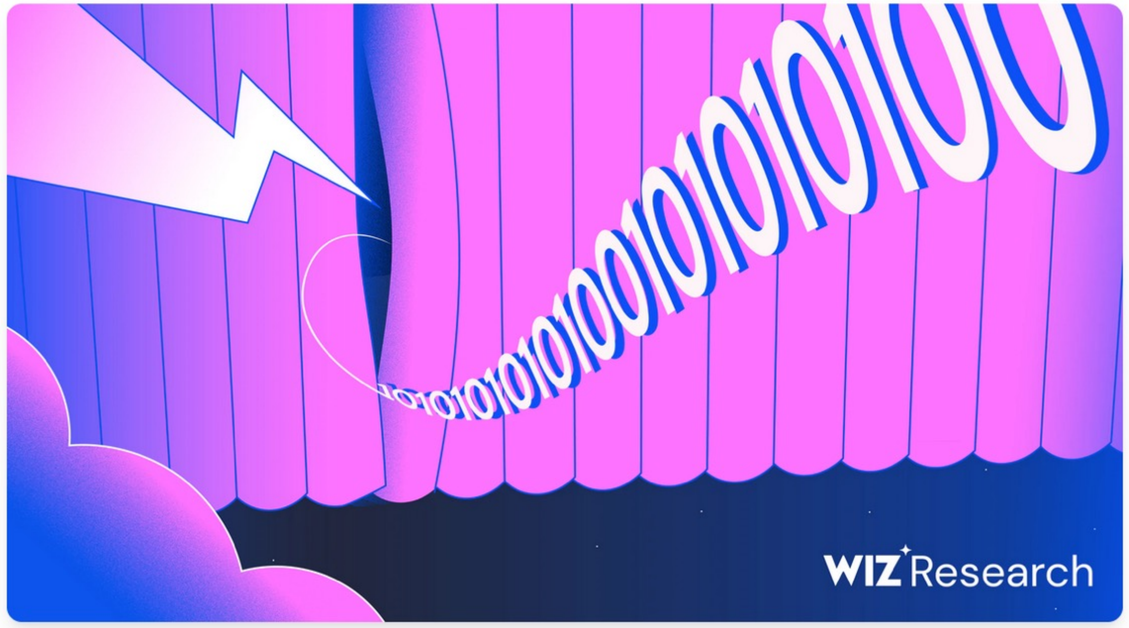
- Utilizar descubrimiento y clasificación de datos para detectar datos sensibles usados en el entrenamiento o ajuste fino
- Implementar controles de seguridad de datos a través de encriptación, gestión de acceso y monitoreo de cumplimiento
- Elevar la conciencia sobre los riesgos de seguridad en cada paso del ciclo de vida de la IA, y asegurar que los equipos de seguridad trabajen estrechamente con los equipos de ciencia de datos e investigación para garantizar las barreras de protección adecuadas

38TB de datos expuestos accidentalmente por investigadores de IA de Microsoft

Wiz Research descubrió un incidente de exposición de datos en el repositorio de GitHub de IA de Microsoft, incluyendo más de 30,000 mensajes internos de Microsoft Teams, todo causado por una configuración incorrecta de un token SAS.

Hillai Ben-Sasson, Ronny Greenberg
September 18, 2023

10 minutes read



¿Qué puede suceder cuando grandes cantidades de datos centralizados no están debidamente protegidos?

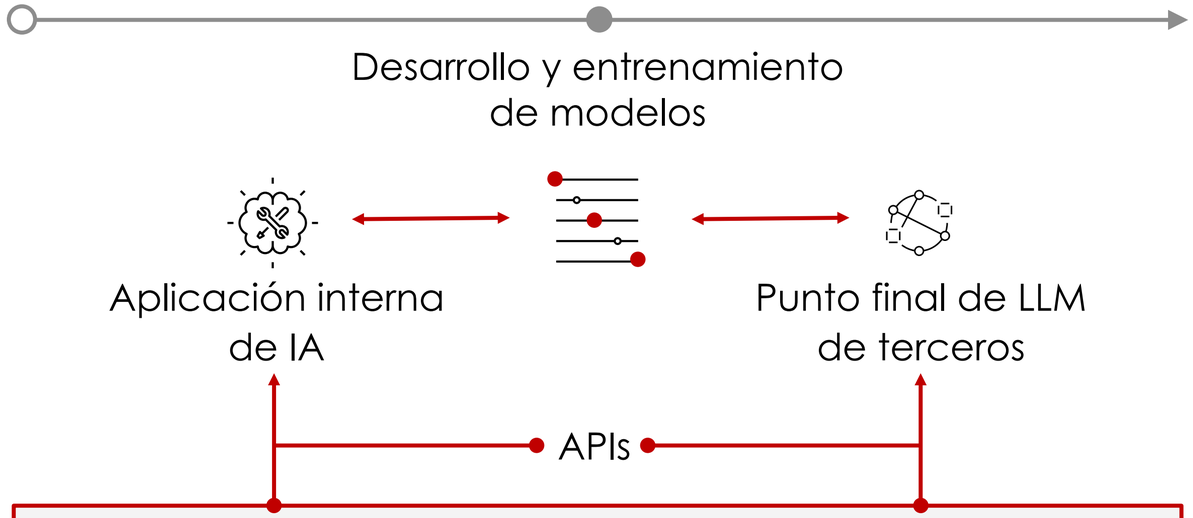
Este caso es un ejemplo de los nuevos riesgos a los que se enfrentan las organizaciones cuando comienzan a aprovechar más ampliamente la IA.

El acceso a montones de datos de entrenamiento puede compartirse inadvertidamente.
El equipo de investigación de IA de Microsoft compartió una URL de un tesoro de datos no estructurados almacenados en un repositorio público de GitHub utilizado para modelos de ML que estaba configurado incorrectamente con acceso excesivamente permisivo.

Llevando a la fuga de datos altamente sensibles.
Los datos incluían secretos, claves privadas, contraseñas y más de 30,000 mensajes internos de Microsoft Teams.

Y, en este caso, abriendo la posibilidad de inyectar código malicioso.
El propósito original del repositorio era proporcionar modelos de IA para usar en el entrenamiento de código, lo que significa que un atacante podría haber inyectado código malicioso en todos los modelos de IA en esta cuenta de almacenamiento, y cada usuario que confíe en el repositorio de GitHub de Microsoft habría sido infectado por él.

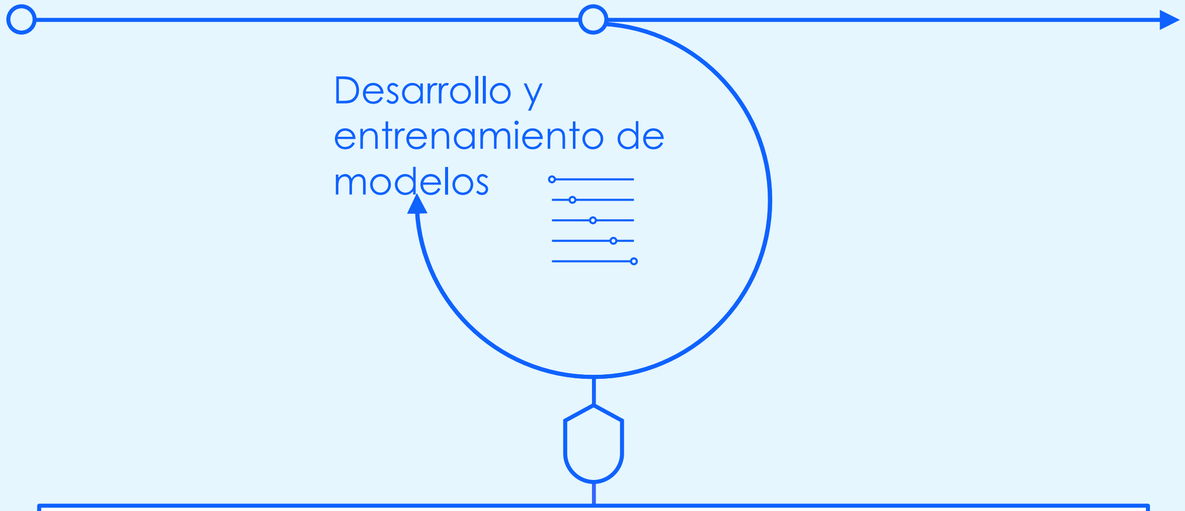
Riesgos en el desarrollo y entrenamiento de modelos



Los atacantes explotan vulnerabilidades y dependencias

- **Ataques a la cadena de suministro:** Los atacantes explotan vulnerabilidades en modelos de código abierto, cadenas de herramientas, bibliotecas de terceros, paquetes de software y otras dependencias
- **Ataques a API:** Los atacantes apuntan a APIs vulnerables que transportan datos sensibles e integran herramientas y aplicaciones de IA
- **Escalada de privilegios:** Los atacantes explotan agentes o complementos de LLM con permisos excesivos para acceder a funciones abiertas y/o sistemas subsecuentes que pueden realizar acciones en flujos de trabajo empresariales

Mejores prácticas de seguridad en el desarrollo y entrenamiento de modelos



Asegurar el modelo

- Escaneo continuo en busca de vulnerabilidades, malware y corrupción a través del pipeline de IA/ML
- Descubrir y fortalecer las integraciones de API y plugins a modelos de terceros
- Configurar y hacer cumplir políticas, controles y RBAC en torno a modelos de ML, artefactos y conjuntos de datos

DARKREADING | The Edge | DR Tech | Sections | Events

Los Modelos de Aprendizaje Automático: Un Nuevo Vector de Ataque Peligroso

Los actores de amenazas pueden armar el código dentro de la tecnología de IA para obtener acceso inicial a la red, moverse lateralmente, desplegar malware, robar datos o incluso envenenar la cadena de suministro de una organización.

Elizabeth Montalbano
Contributor, Dark Reading

December 06, 2022

Source: Skorzewiak via Alamy Stock Photo

Los actores de amenazas pueden secuestrar modelos de aprendizaje automático (ML) que potencian la inteligencia artificial (IA) para desplegar malware y moverse lateralmente a través de las redes empresariales, han encontrado investigadores. Estos modelos, que a menudo están disponibles públicamente, sirven como una nueva plataforma de lanzamiento para una gama de ataques que también pueden envenenar la cadena de suministro de una organización, y las empresas necesitan prepararse.

¿Qué puede suceder cuando las aplicaciones de IA se construyen de manera insegura?

La dependencia de modelos de código abierto crea un riesgo inherente

Es común dentro de la ciencia de datos descargar y reutilizar modelos de aprendizaje automático preentrenados de código abierto de repositorios de modelos en línea como HuggingFace o TensorFlow Hub. La escasez general de seguridad alrededor de los modelos de ML, combinada con los datos cada vez más sensibles a los que están expuestos los modelos de ML, significa que estos ataques a modelos podrían tener una alta propensión al daño.

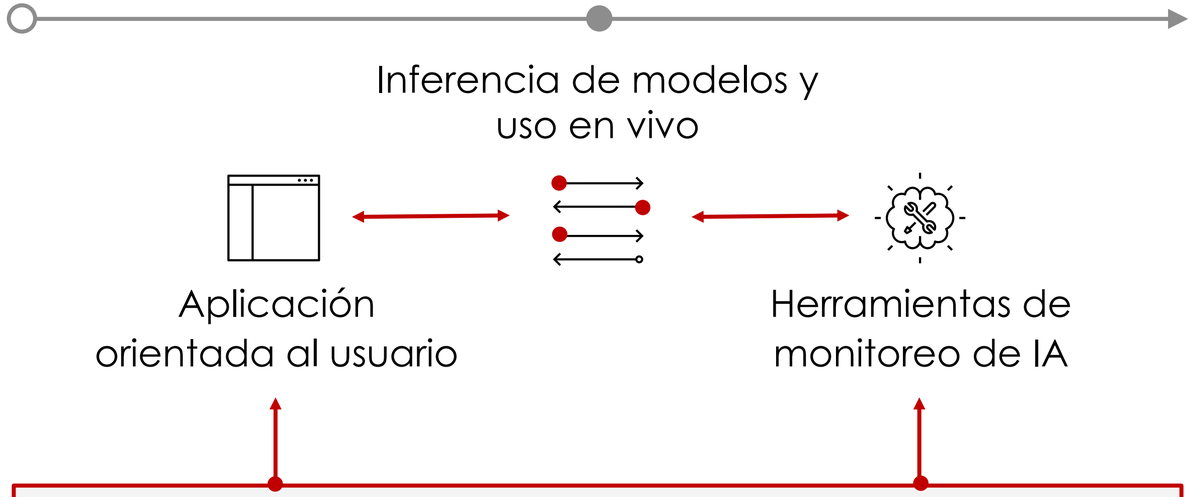
Se pueden inyectar puertas traseras y malware en modelos de código abierto

Investigadores del Equipo de Inteligencia Adversaria Sináptica de HiddenLayer desarrollaron un ataque de prueba de concepto para demostrar lo fácil que un adversario puede desplegar malware a través de un modelo de ML preentrenado de código abierto que podría evadir la detección de soluciones antivirus y EDR.

Las empresas están expuestas a ataques a la cadena de suministro de ML

Un atacante podría reemplazar un modelo benigno legítimo con su versión troyanizada que ejecutará el malware incrustado. Todos los que descarguen el modelo troyanizado y lo carguen en una máquina local se verán afectados. Un atacante también podría usar este método para secuestrar la cadena de suministro de un proveedor de servicios para infectar a todos los suscriptores.

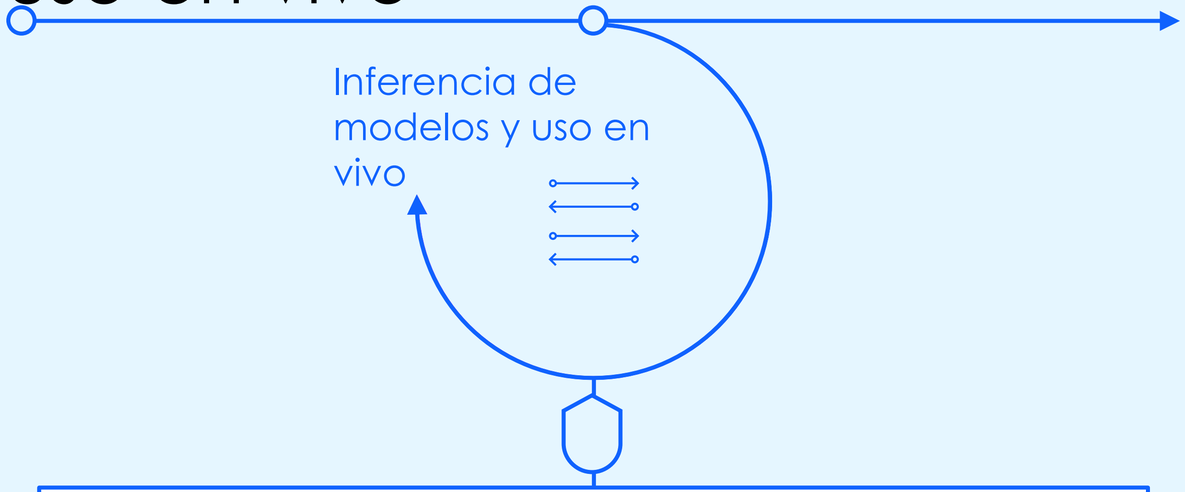
Riesgos de inferencia de modelos y uso en vivo



Los atacantes explotan vulnerabilidades y dependencias

- **Inyección de comandos:** Comandos maliciosos pueden liberar LLMs, proporcionar acceso indebido, robar datos sensibles o sesgar salidas
- **Denegación de servicio del modelo:** Los atacantes sobrecargan el LLM con entradas que degradan la calidad del servicio e incurren en altos costos de recursos
- **Robo de modelos:** El atacante crea entradas para recopilar salidas del modelo, acumulando un gran conjunto de datos de pares de entrada-salida con el fin de entrenar un modelo sustituto para imitar el comportamiento del modelo objetivo efectivamente "robando" sus capacidades

Mejores prácticas de seguridad para inferencia de modelos y uso en vivo



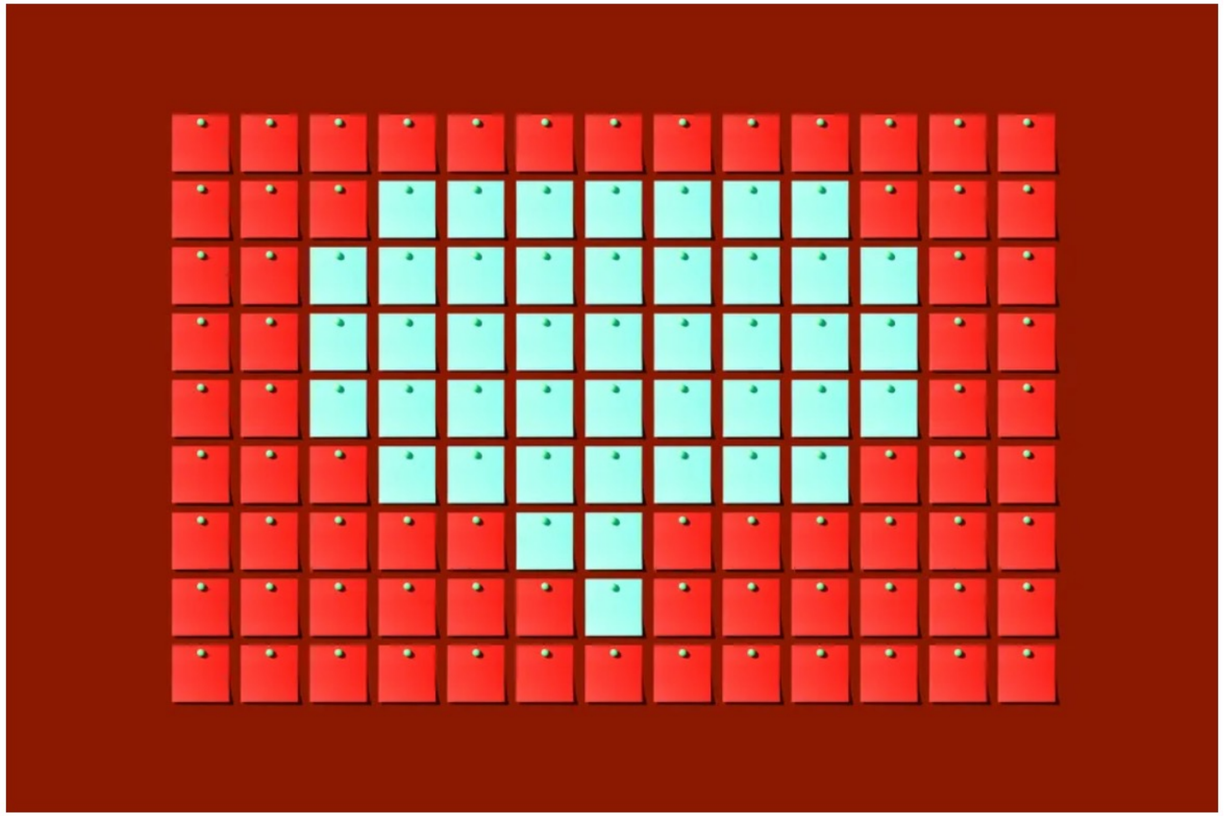
Asegurar el uso

- Monitorear entradas maliciosas como inyecciones de comandos y salidas que contienen datos sensibles o contenido inapropiado
- Implementar soluciones de seguridad de IA que puedan detectar y responder a ataques específicos de IA (por ejemplo, envenenamiento de datos, evasión de modelos, extracción de modelos)
- Desarrollar manuales de respuesta para negar acceso, poner en cuarentena y/o desconectar modelos comprometidos

WIRED

Un Nuevo Ataque Afecta a los Principales Chatbots de IA—y Nadie Sabe Cómo Detenerlo

Los investigadores encontraron una manera simple de hacer que ChatGPT, Bard y otros chatbots se comporten mal, demostrando que la IA es difícil de domesticar.



PHOTOGRAPH: MIRAGEC/GETTY IMAGES

¿Qué puede suceder cuando las interacciones con modelos de caja negra son maliciosas?

Las barreras de seguridad para chatbots ampliamente utilizados pueden ser fácilmente vulneradas

Investigadores de la Universidad Carnegie Mellon y el Centro para la Seguridad en IA lograron eludir las medidas de seguridad de IA en todos los chatbots líderes, incluidos ChatGPT, el Bard de Google y Claude de Anthropic, agregando un largo sufijo de caracteres a cada indicación en inglés introducida en el sistema

Se puede inducir a los modelos a comportarse mal a pesar de haber sido entrenados para no hacerlo

Aunque los modelos fueron entrenados para no exponer información sensible y se construyeron con barreras de seguridad diseñadas para prevenir que el sistema genere contenido tóxico o dañino, el ataque logró inducir a los chatbots a generar respuestas prohibidas a comandos dañinos incluyendo información sesgada, falsa y otro tipo de información tóxica

Las defensas actuales integradas en los sistemas de IA se muestran frágiles

Los investigadores utilizaron un modelo de lenguaje de código abierto genérico para desarrollar un ataque adversario automatizado que logró vulnerar varios sistemas propietarios diferentes, demostrando que la capacidad de hacer que los sistemas de IA obedezcan comandos del usuario incluso si producen contenido dañino representa una debilidad fundamental para la cual no existe una solución actualmente

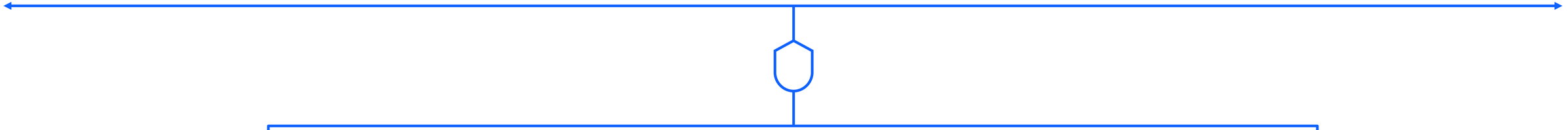
GBM

Extender la seguridad existente a lo largo de la infraestructura de IA subyacente

Detección de amenazas y respuesta

Seguridad de datos

Gestión de identidad, fraude y dispositivos



Asegurar la infraestructura

- Utilizar controles de seguridad de infraestructura como primera línea de defensa contra el acceso adversario a la IA
- Aprovechar la experiencia existente para optimizar la seguridad, privacidad y estándares de cumplimiento en entornos distribuidos
- Fortalecer la seguridad de red, control de acceso, encriptación de datos y detección y prevención de intrusiones alrededor de los entornos de IA
- Mientras también se invierte en nuevas defensas de seguridad diseñadas específicamente para proteger la IA

Establecer gobernanza de IA

Construir IA confiable

Gestionar el riesgo y la reputación

- Habilitar modelos de IA responsables, explicables y de alta calidad, y documentar automáticamente el linaje y los metadatos del modelo
- Monitorear la equidad, sesgo y deriva para detectar la necesidad de reentrenamiento del modelo

Apoyar el cumplimiento regulatorio

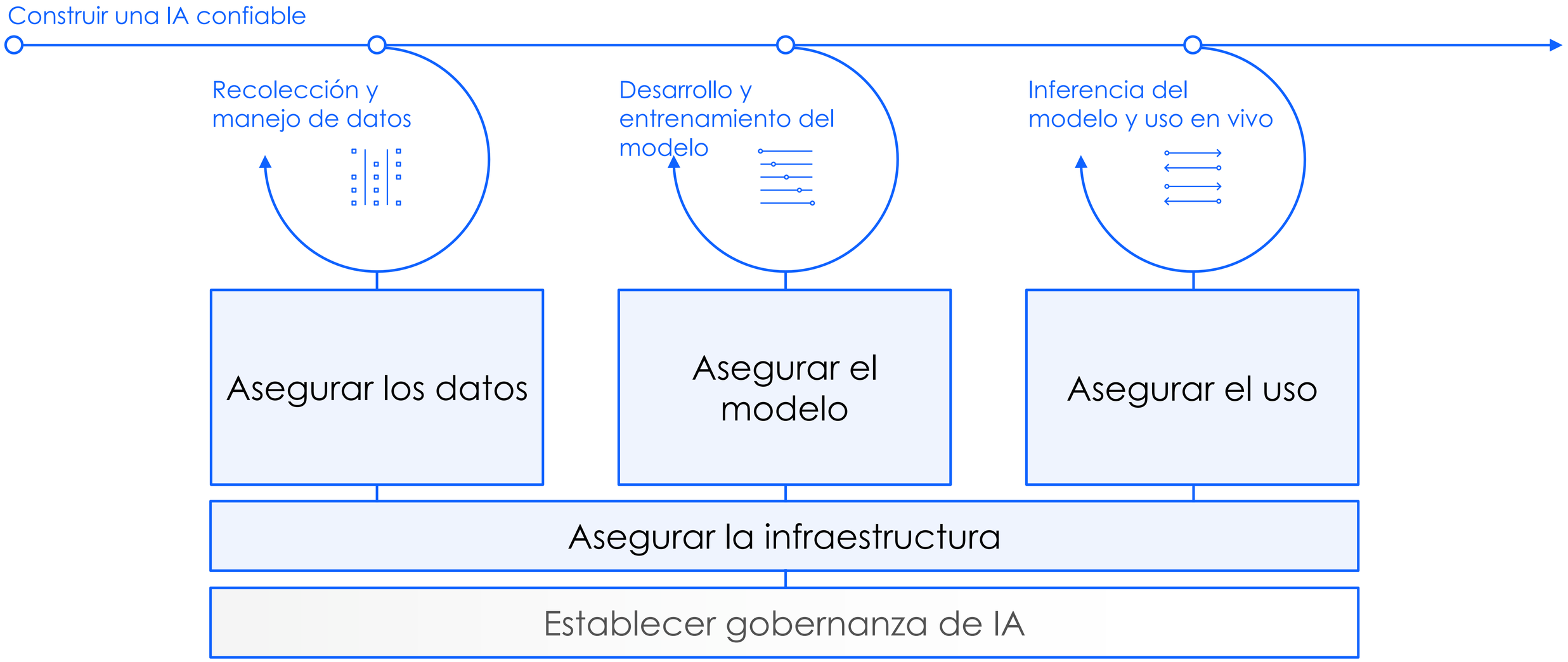
- Usar protecciones y validaciones para ayudar a habilitar modelos que sean justos, transparentes y conformes
- Documentar automáticamente los hechos del modelo en apoyo de las auditorías

Operacionalizar la gobernanza de IA

- Acelerar la construcción de modelos a gran escala
- Automatizar y consolidar múltiples herramientas, aplicaciones y plataformas mientras se documenta el origen de conjuntos de datos, modelos, metadatos asociados y pipelines

Para resumir: Lo que necesitas hacer

Marco de Seguridad para la IA



IA para la Seguridad

Seguridad potenciada por IA

Capacidades integradas con IA infusionadas en soluciones de Seguridad

Detección de Amenazas y Respuesta

Minería de datos automática para obtener más contexto, reevaluar el riesgo y luego generar una línea de tiempo del ataque mapeada a MITRE, con acciones de respuesta recomendadas

Seguridad de Datos

Proporcionar una advertencia temprana de ataques, incluso en ausencia de detecciones basadas en reglas activadas, mediante el monitoreo del comportamiento de usuarios privilegiados en busca de desviaciones que indiquen riesgo

Gestión de Identidad y Acceso

Ofrecer autenticación basada en riesgo, calculando una puntuación de riesgo para cada intento de inicio de sesión y adaptando automáticamente el proceso de autenticación según la confianza del usuario

Resultados habilitados por IA

55%

aumento de velocidad en la evaluación de alertas usando la priorización de IA y la investigación automatizada [1]

40%

disminución en violaciones de seguridad con visibilidad centralizada y análisis avanzados[2]

15x

disminución en la fricción del usuario asociada con la autenticación multifactor al incorporar acceso adaptativo [3]

[1] IBM Press Release, "IBM Launches New QRadar Security Suite to Speed Threat Detection and Response,"; [2] *The Total Economic Impact of IBM Security Guardium*, Forrester Report; [3] IBM Security Verify customer, leading U.S. insurance company

IA Generativa en la Hoja de Ruta de Detección de Amenazas y Respuesta

Automatizar Informes

Crear resúmenes simples de casos de seguridad e incidentes que pueden ser compartidos con una variedad de partes interesadas con un solo clic

Acelerar la Caza de Amenazas

Generar automáticamente búsquedas para detectar amenazas basadas en descripciones de comportamiento y patrones de ataque en lenguaje natural – ayudando a acelerar la respuesta a nuevas campañas de amenazas

Interpretar Datos de Máquina

Ayudar a los analistas a comprender más rápidamente los datos de registros de seguridad proporcionando explicaciones simples de las acciones que han tenido lugar en un sistema - reduciendo barreras técnicas y acelerando sus investigaciones

Curar Inteligencia de Amenazas

Interpretar y resumir la inteligencia de amenazas más relevante, enfocándose en campañas que tienen más probabilidades de afectar a las Empresas basándose en su perfil de riesgo único

GBM Cybersecurity Center

+35

Referencias

Empresas que confían en los servicios de GBM Cybersecurity Center.

700M

Data Lake

Recibimos y analizamos más de 700 millones de eventos diarios.

+20

Intelligence

fuentes para nuestros servicios de ciberdefensa SOC/MDR/CDC.

+100

Especialistas

de ciberseguridad para la atención de nuestros clientes en la región.

GBM MISP THREAT SHARING

GBM HoneyPot Net THREAT DETECTOR

CERTIFIED SOC1 Typ 2 ZERTIFIZIERT

Technology Partners

Threat Management: QRadar XDR SIEM | QRadar XDR SOAR | QRadar XDR Connect | ReaQta EDR | Randori – ASM

Data Security: Guardium Insights | Data Protection | Data Encryption | Discover and Classify | Key Lifecycle Management

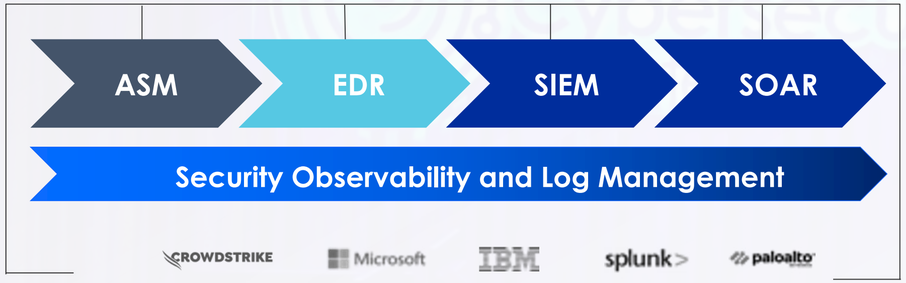
Zero Trust NG: Zero Trust Zone | End 2 End Zero Trust

IBM Security · cisco Security · Microsoft Security

3 **nodos**. Replicación

Principal Guatemala
Secundarios Costa Rica & Panamá.

Proactive Threat-driven Defense Strategy



Alliance Partners

Europol

Centro Europeo de Delitos Cibernéti de Europol (EC3) Member

NO MORE RANSOM

FIRST

Forum of Incident Response and Security Teams Member

FIRST Improving Security Together

CyberCluster

Lista selecta de organizaciones del sector de ciberseguridad.

Cybersec Cluster

GBM

Muchas gracias

...

www.linkedin.com/in/alonsoramirezcybersecurity

laramirez@gbm.net

W: +506 8827-3344