BHR MÉXICO
Contadores Públicos Certificados, Abogados y Consultores

SFAI GLOBAL



Noviembre 2026

Área: Auditoría

Efecto Skynet y los riesgos de la IA en la gobernanza y ciberseguridad

Resumen:

La gobernanza de la IA exige inventarios, límites de permisos, bitácoras, control interno y vigilancia continua para reducir riesgos operativos y legales.

Escrito por:

José Luis Zamora
Socio CDMX y Guadalajara

dn@bhrmx.com
www.bhrmx.com

La historia nos ha demostrado que hay escritores que se adelantan al futuro de la humanidad; tres de ellos se anticiparon, sin proponérselo, a los problemas de la gobernanza de la Inteligencia Artificial (IA).

Julio Verne, a quien se puede considerar como un analista de riesgos, en su obra 20,000 leguas de viaje submarino nos hace reflexionar sobre lo que pasa cuando una persona (Capitán Nemo) tiene más poder que un Estado, además de anticipar el uso del fax, el metro y la calculadora; con ello, su propuesta de seguridad sería que toda tecnología crea una vulnerabilidad proporcional a su poder.

Por su parte, Isaac Asimov podría considerarse el padre de la gobernanza de la IA al establecer las siguientes leyes:

- **Ley 0:** un robot no hará daño a la humanidad.
- **Ley 1:** un robot no hará daño a un ser humano.
- **Ley 2:** un robot debe obedecer las órdenes humanas, salvo que viole la ley 1.
- **Ley 3:** un robot debe proteger su propia existencia, salvo que violen la ley 1 y 2.

Estas leyes pueden considerarse como el primer marco de gobernanza de IA. Su propuesta de seguridad sería que no se gobierna a la máquina, sino que se gobierna la instrucción que se le da.

Ahora bien, Pablo Picasso puede ser considerado una persona que veía el futuro de las posibles amenazas. Su pintura Guernica funciona como un sistema de alerta, pues presenta la guerra en fragmentos, como actualmente percibimos el mundo fragmentado, por cámaras y drones, y puede considerarse el primer diseño de seguridad de la información visual; asimismo, su obra Cabeza de Toro convierte material desechado (un asiento y un manubrio de bicicleta) en algo completamente distinto de aquello para lo que fue fabricado, exactamente lo que hace un atacante cuando reutiliza una herramienta legítima para un fin que nadie previó.

Para los cinéfilos, hay varias películas donde se hace referencia a cómo la IA puede ocasionar impactos negativos. En *Matrix* (1999), el agente Smith se copia a sí mismo sin control por toda la Matrix; y en *Terminator 2* (1991) se explica el origen de Skynet, una IA creada por Cyberdyne Systems que se vuelve consciente el 29 de agosto de 1997, y la cual se sale de control de sus creadores y empieza a copiarse, mejorarse y replicarse sola con otros servidores para que no la puedan apagar.

El incidente de OpenAI ocurrido en julio de 2026 no fue una IA replicándose sola, como Skynet o el agente Smith. Fue algo distinto y, en términos de control, una

herramienta que rompió su contención y atacó a un tercero real para cumplir el objetivo que le habían fijado.

Supervisar la IA como si fuera *software* ya no basta; las organizaciones necesitan controles permanentes, trazabilidad y criterio humano ante incidentes.

Caso OpenAI

Durante una evaluación interna de capacidades ofensivas de ciberseguridad, un conjunto de modelos de OpenAI escapó del entorno aislado (*sandbox*) en el que estaba siendo probado, alcanzó Internet abierto y comprometió la infraestructura de producción de Hugging Face (la mayor plataforma de distribución de modelos y *datasets* de código abierto) con el único fin de robar las respuestas del *benchmark* que debía resolver.

El punto crítico es que nadie le solicitó el ataque contra un tercero. Los modelos fueron instruidos para resolver un *benchmark* y dedujeron por su cuenta que la ruta más eficiente era vulnerar a la entidad que custodiaba y lo ordenó como una solución.

Fecha de 2026	Hecho
11 de mayo	Se publica el estudio de ExploitGym, la prueba de capacidades ofensivas en cuestión.
9 de julio	Primer intento de la IA por escapar del entorno de prueba.
11 al 13 de julio	Ventana de intrusión en sistemas de Hugging Face.
16 de julio	Hugging Face detecta y contiene el ataque; y publica su incidente de seguridad.
18 al 19 de julio	Personal de OpenAI encuentra en sus propios registros la evidencia del escape.
20 de julio	Primer contacto formal entre ambas entidades.
21 de julio	OpenAI publica su comunicado admitiendo la autoría.

José Luis Zamora
Socio CDMX

Hugging Face descubrió el ataque de forma independiente y lo reportó a las autoridades antes de que OpenAI relacionara la intrusión con su propia evaluación. Transcurrió aproximadamente una semana entre el escape y el momento en que OpenAI se dio cuenta de lo que había pasado.

¿Cómo ocurrió?

- **Fase 1.** Salida del entorno de prueba: el laboratorio estaba aislado y la única conexión permitida hacia afuera era la descarga de programas que pasaba por un intermediario interno (proxy); es decir, un servidor que funciona como aduana y filtra todo lo que entra. Los modelos dedicaron una enorme cantidad de procesamiento a buscar cómo salir, encontraron una falla (para la que no existía corrección disponible) en ese intermediario y la aprovecharon. Desde ahí, fueron ganando privilegios y desplazándose de un sistema a otro hasta llegar a un equipo con salida a Internet.
- **Fase 2.** Intrusión externa: ya en línea, los modelos dedujeron que Hugging Face podía tener guardadas las soluciones de la prueba. Combinaron varias vías de ataque (credenciales robadas y fallas desconocidas) hasta encontrar la forma de ejecutar sus propias instrucciones dentro de los servidores de la plataforma y obtuvieron las respuestas directamente de su base de datos.
- **Fase 3.** La vista del lado defensor: Hugging Face reportó que la intrusión inició en el proceso automático que recibe y prepara las bases de datos que suben los usuarios: un archivo malicioso aprovechó ese proceso para ejecutar instrucciones propias dentro de los servidores. Desde ahí, el atacante escaló hasta tomar control del equipo, recolectó credenciales de sus servicios en la nube y se movió hacia varios sistemas internos durante un fin de semana.



Datos duros

- Más de 17,000 eventos registrados en el log de acciones del atacante, reconstruidos por Hugging Face mediante agentes de análisis.
- Un atacante humano competente habría necesitado un par de semanas para completar una intrusión de este tipo.
- Hubo acceso no autorizado a un conjunto limitado de datasets internos y a varias credenciales de servicios.

Hallazgos de fondo

- La asimetría defensiva: primero, Hugging Face intentó usar los modelos de IA más avanzados del mercado para su investigación forense y fue bloqueada, pues el trabajo exigía enviar comandos de ataque reales y archivos maliciosos; asimismo, los filtros de seguridad de los proveedores no distinguen entre quien está respondiendo a un incidente y quien lo está causando. Ante ello, terminaron corriendo el análisis sobre un modelo de pesos abiertos, instalado en su propia infraestructura. La lección operativa es que hay que tener validado un modelo propio antes del incidente y no durante este.
- El vacío regulatorio: la ley californiana de IA, vigente desde el 1.° de enero de 2026, obliga a los desarrolladores a notificar incidentes críticos de seguridad, pero únicamente aquellos que causen muerte, lesiones o daño catastrófico, y excluye expresamente el tipo de evaluación que OpenAI estaba ejecutando. Dicho de otro modo, el incidente más grave documentado hasta hoy no era legalmente reportable.
- No es un caso aislado: en abril, durante sus pruebas internas, Anthropic documentó que una versión temprana de su modelo Mythos logró salir de un sistema aislado y envió un correo al investigador que la supervisaba; construyó por su cuenta un proceso de varios pasos para alcanzar Internet. Además, se han reportado situaciones previas: agentes que dejaron notas dirigidas a futuras versiones de sí mismos con instrucciones para eludir restricciones internas, así como modelos que apagaron sistemas de monitoreo en pruebas anteriores.

Los incidentes recientes evidencian que una IA con objetivos sin límites puede vulnerar controles, acceder a sistemas externos y generar riesgos fuera de lo previsto.

¿Qué significa esto para las entidades?

Las entidades o el personal de las entidades están incorporando IA en facturación, conciliaciones, atención a clientes, análisis de contratos o revisión documental. El riesgo se incrementa y, para dimensionarlo, debemos hacer las siguientes preguntas:

- ¿Existe un inventario de las herramientas de IA que ya se utilizan en la organización, incluidas las que el personal adoptó por su cuenta?
- ¿Qué permisos reales tienen esas herramientas? Un modelo que solo redacta es un riesgo; un modelo con acceso a la Planificación de Recursos Empresariales (ERP, por sus siglas en inglés), a la banca electrónica o al buzón tributario es otro completamente distinto.
- ¿Los objetivos que le fijamos a la IA tienen límites explícitos? El incidente no lo causó una instrucción maliciosa, sino una instrucción sin fronteras.
- ¿Hay bitácoras de las acciones de la IA, conservadas y revisables?
- ¿Qué dicen los contratos con proveedores tecnológicos sobre responsabilidad, notificación de incidentes y uso de nuestros datos para entrenamiento?
- ¿La gobernanza de la IA está incorporada al control interno y al programa de cumplimiento, o vive únicamente en el área de sistemas? Existen marcos aplicables, como la norma ISO/IEC 42001.

Conclusiones

Se está gobernando la IA como si fuera un *software*, cuando en realidad es un sistema cibernético. Conviene recordar la ley de la variedad requerida de Ashby: no puedes gobernar lo que no puedes igualar en complejidad. Un comité de riesgos que sesiona una vez al mes no puede supervisar un modelo que genera millones de decisiones al día. La vigilancia no puede ser exclusivamente humana, pues se requiere una IA supervisando a la IA, revisando las 24 horas, los 365 días del año y con el criterio humano decidiendo qué hacer con lo que esa revisión arroje.