WEBINAR

IA Generativa y Protección de Datos: Seguridad en el Entorno Digital de la UE

Junta de Andalucía
Consejería de Turismo y Andalucía Exterior

Cámaras Andalucía

enterprise europe network

Red de Información Europea de Andalucía

IA Generativa y Protección de Datos: Seguridad en el Entorno Digital de la UE

BIENVENIDA

Eva Mª Ángel Lagares

Coordinadora de Proyectos Competitividad Empresarial en Consejo Andaluz de Cámaras de Comercio Industria y Navegación

EXPONE

Jacob Peregrina Barahona

Abogado especializado en Derecho Digital.
Fundador de Gabinete Jurídico Tecnoiuris

Junta de Andalucía
Consejería de Turismo y Andalucía Exterior

Cámaras Andalucía

enterprise europe network

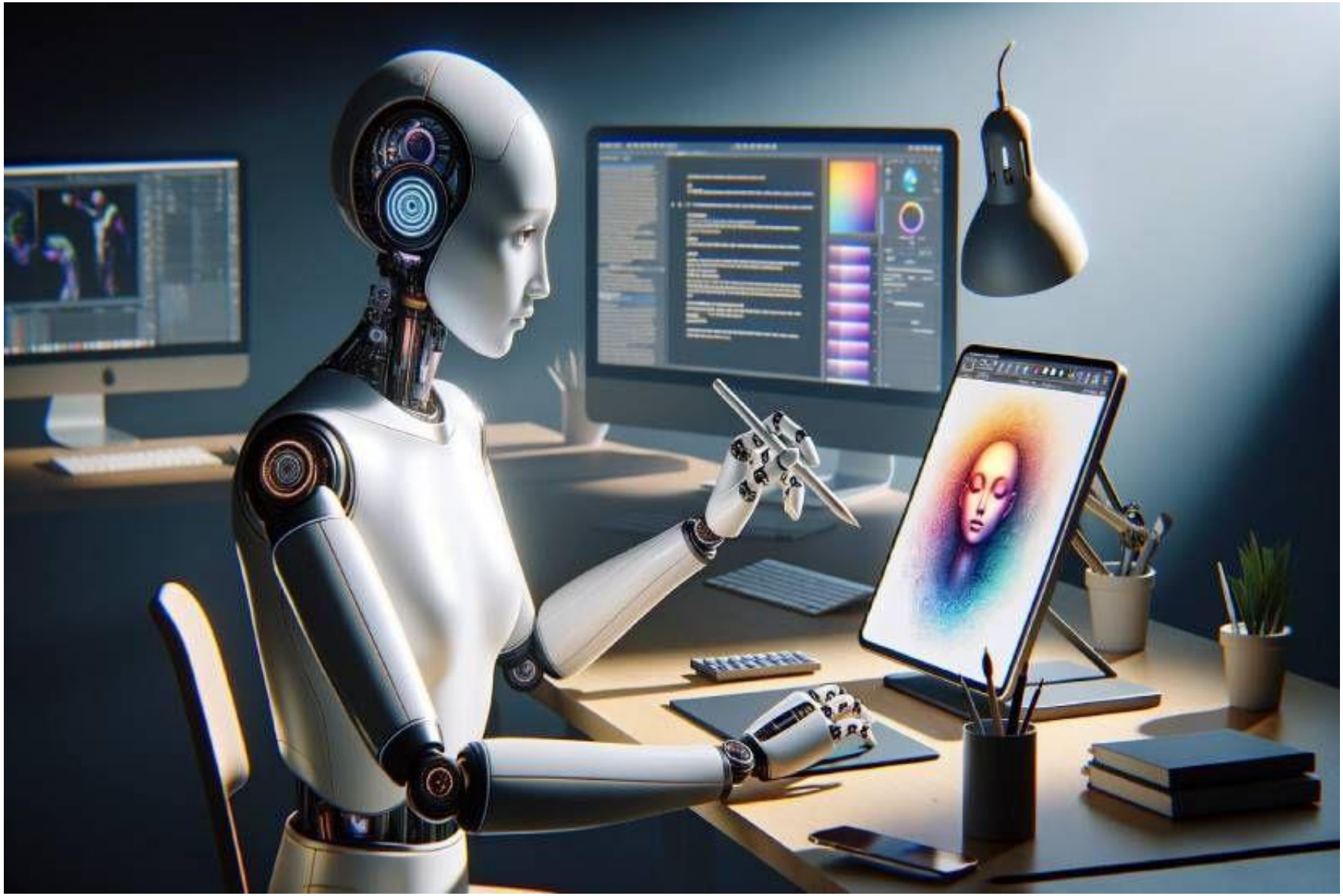
Red de Información Europea de Andalucía

LOS PELIGROS OCULTOS DE LA IA GENERATIVA

Protegiendo datos sensibles en un mundo digital

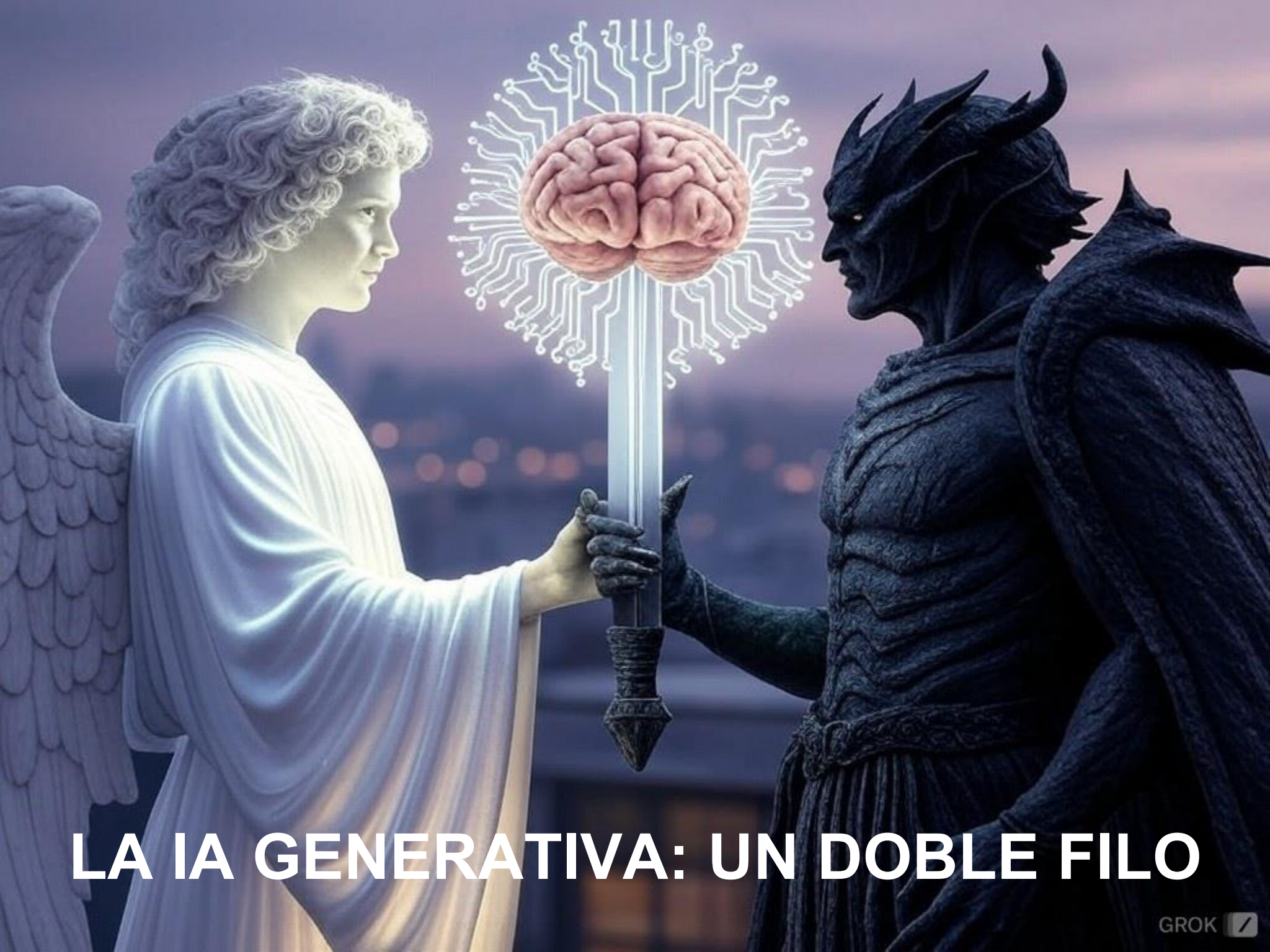
¿QUÉ ES LA IA GENERATIVA?

TECNOIURIS
GABINETE JURÍDICO



El proceso de entrenamiento:

1. **Recopilación de datos:** Se recolectan enormes cantidades de datos de diversas fuentes, como internet, bases de datos y redes sociales. Estos datos pueden incluir textos, imágenes, videos y audio.
2. **Procesamiento de datos:** Los datos se limpian, estructuran y preparan para el entrenamiento. Se eliminan datos duplicados, se corrigen errores y se normalizan los formatos.
3. **Entrenamiento del modelo:** El modelo de IA aprende a identificar patrones en los datos y a generar nuevos contenidos similares. Este proceso se realiza a través de algoritmos de aprendizaje profundo, como las redes neuronales artificiales.


LA IA GENERATIVA: UN DOBLE FILO
GROK

Aplicaciones en el Mundo Empresarial

- **Marketing y Publicidad:** Creación de contenido personalizado, generación de ideas para campañas y diseño de productos.
- **Desarrollo de Software:** Generación de código, automatización de pruebas y mejora de la eficiencia en la creación de aplicaciones.
- **Diseño:** Creación de diseños gráficos, arquitectónicos y de productos de manera más rápida y eficiente.
- **Atención al Cliente:** Desarrollo de chatbots más inteligentes y capaces de mantener conversaciones naturales.

¿Por qué es tan atractiva?

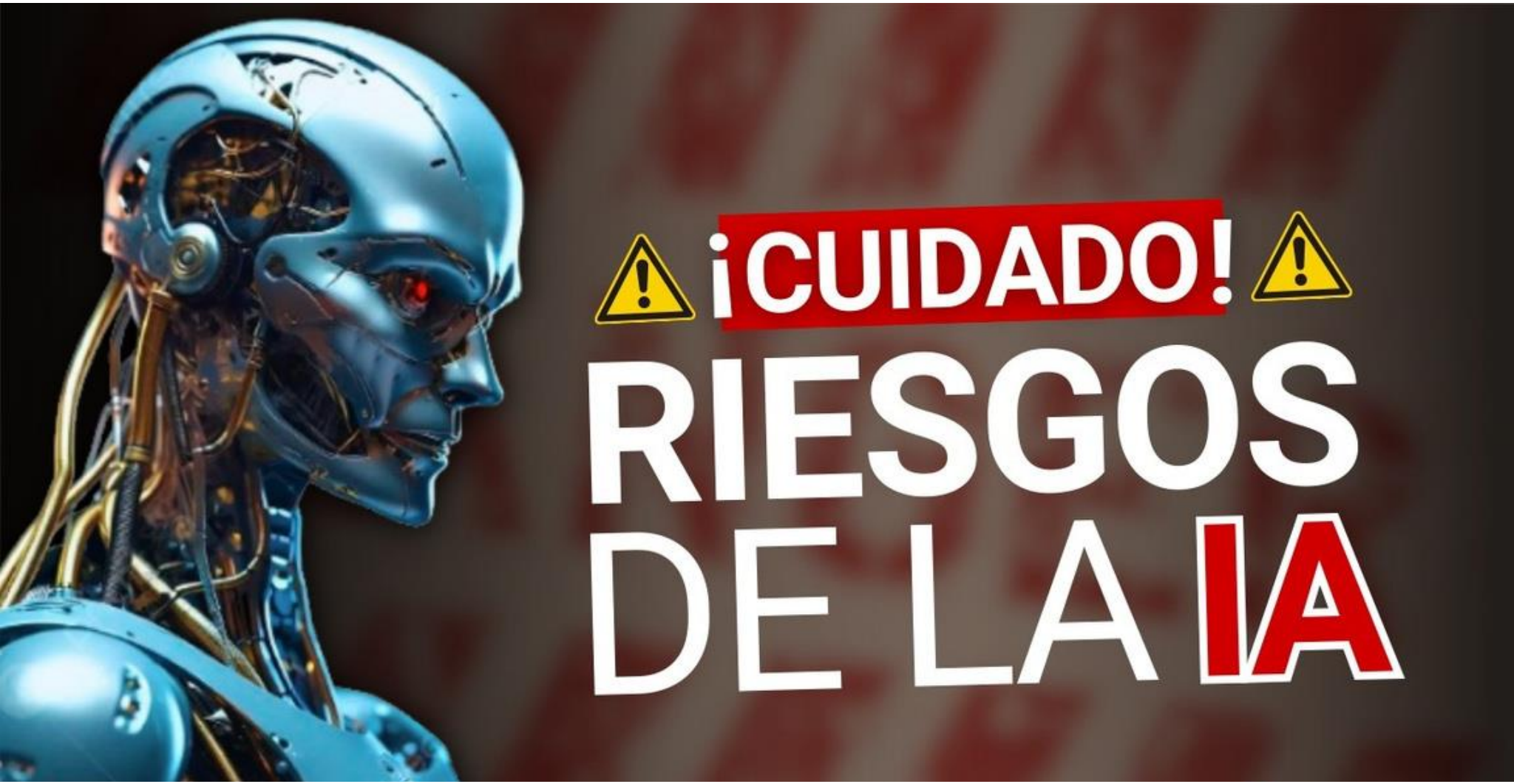
- **Aumento de la productividad:** Automatiza tareas repetitivas y libera a los empleados para que se enfoquen en tareas más estratégicas.
- **Personalización:** Permite crear productos y servicios altamente personalizados para cada cliente.
- **Innovación:** Abre nuevas posibilidades creativas y fomenta la innovación.
- **Reducción de costes:** Puede reducir los costes de producción y desarrollo.

El Reverso: Riesgos y Desafíos

- **Privacidad:** Los modelos de IA Generativa requieren grandes cantidades de datos para entrenar, lo que plantea preocupaciones sobre la privacidad de la información personal.
- **Sesgos:** Los modelos pueden perpetuar y amplificar los sesgos presentes en los datos de entrenamiento, lo que puede llevar a resultados discriminatorios.
- **Deepfakes:** La capacidad de generar contenido altamente realista puede ser utilizada para crear desinformación y manipular la opinión pública.
- **Desempleo:** La automatización de tareas puede llevar a la pérdida de empleos.

¿Cómo la IA Generativa puede convertirse en una amenaza para nuestros datos más sensibles?

- **Filtraciones de datos:** Durante el entrenamiento, los modelos pueden exponer accidentalmente información confidencial.
- **Ataques adversariales:** Los atacantes pueden manipular los modelos para que revelen información sensible.
- **Generación de contenido falso:** La IA Generativa puede ser utilizada para crear contenido falso y engañoso que se asemeja a la realidad.


¡CUIDADO!
RIESGOS
DE LA IA

Riesgos para la privacidad:

1.Recopilación masiva y uso indebido de datos personales.

- **Datos sensibles:** La IA Generativa puede recopilar una amplia gama de datos personales, incluyendo:
  - Información de identificación personal (PII): nombres, direcciones, números de teléfono, correos electrónicos.
  - Datos biométricos: huellas dactilares, reconocimiento facial, iris.
  - Historial de navegación: sitios web visitados, búsquedas realizadas.
  - Preferencias y hábitos de consumo: productos comprados, intereses.
- **Uso sin consentimiento:** A menudo, esta recopilación se realiza sin el conocimiento o consentimiento explícito del usuario. Las políticas de privacidad pueden ser difíciles de entender y aceptar, y los usuarios pueden no ser conscientes de cómo se utilizan sus datos.
- **Propósitos comerciales:** Los datos recopilados pueden ser utilizados para crear perfiles detallados de los usuarios, los cuales pueden ser vendidos a terceros con fines comerciales. Por ejemplo, las empresas pueden utilizar estos perfiles para personalizar la publicidad y ofrecer productos y servicios más relevantes.

2. Filtración de datos sensibles durante el entrenamiento

- **Vulnerabilidades en las bases de datos:** Los conjuntos de datos utilizados para entrenar los modelos pueden contener información confidencial que podría filtrarse accidentalmente debido a configuraciones de seguridad inadecuadas o ataques cibernéticos.

- **Ataques cibernéticos:** Los hackers pueden aprovechar las vulnerabilidades en los sistemas de almacenamiento de datos para robar información sensible, como contraseñas, números de tarjetas de crédito o registros médicos.

3. Deepfakes y desinformación

La IA Generativa puede crear contenido falso, altamente realista, conocido como deepfakes. Estos pueden utilizarse para:

- **Difundir desinformación:** Crear noticias falsas o manipular la opinión pública.
- **Dañar la reputación:** Crear contenido falso que dañe la reputación de personas o instituciones.
- **Cometer fraudes:** Suplantar la identidad de personas para cometer delitos financieros o extorsión.

Consecuencias de la filtración de datos:

- **Robo de identidad:** Los datos filtrados pueden ser utilizados para cometer fraudes financieros o suplantaciones de identidad.
- **Daño reputacional:** La divulgación de información personal puede causar daños irreparables a la reputación de las personas.
- **Discriminación:** Los datos sesgados utilizados para entrenar los modelos pueden perpetuar estereotipos y discriminaciones. Por ejemplo, un modelo de reclutamiento entrenado con datos sesgados podría favorecer a candidatos de un determinado género o grupo étnico.
- **Pérdida de confianza:** La filtración de datos puede erosionar la confianza de los usuarios en las empresas y en las instituciones.

TECNOIURIS
GABINETE JURÍDICO

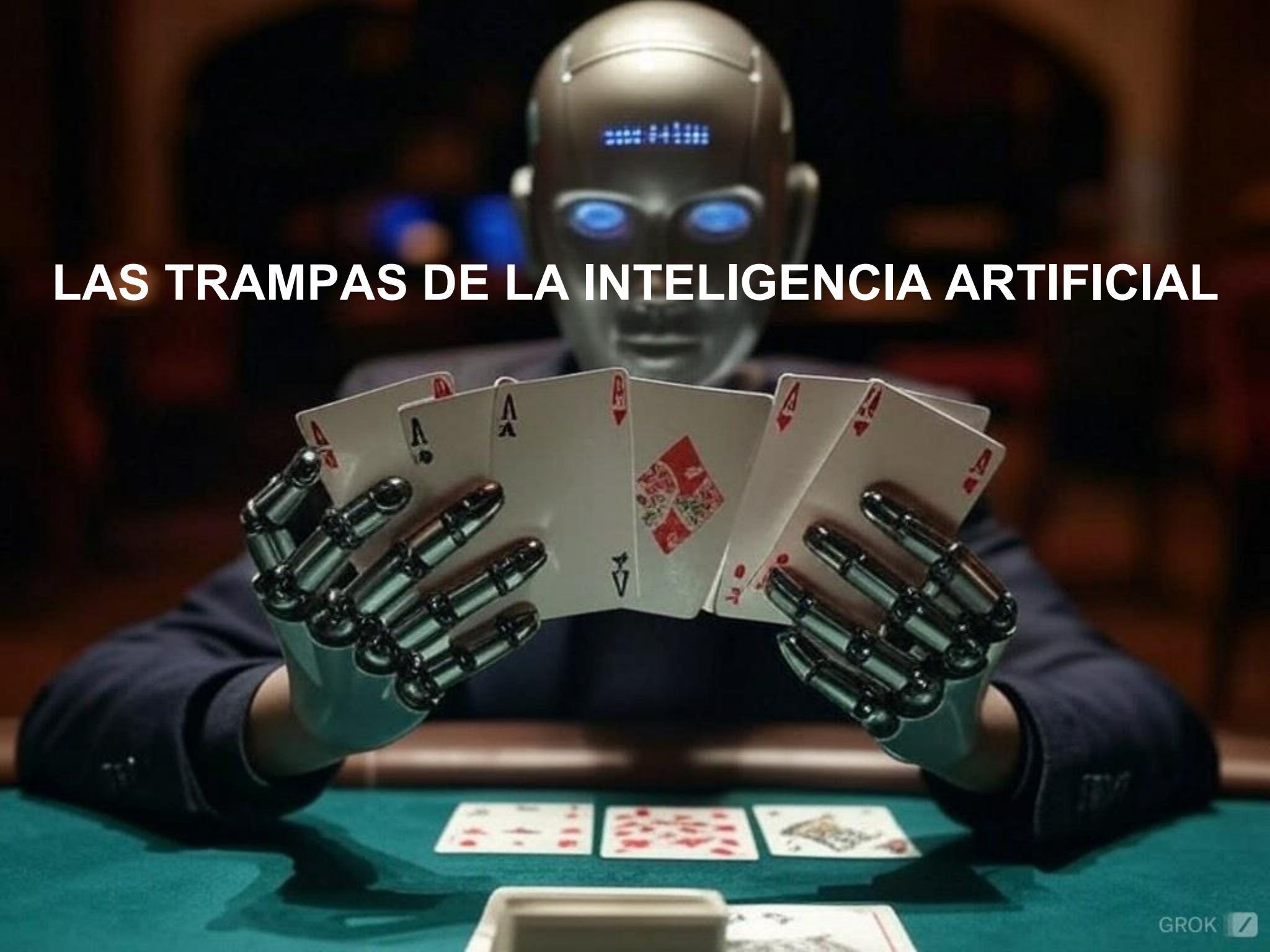
Propiedad Intelectual: ¿Quién es el autor?

- **Creatividad algorítmica:** La IA Generativa puede crear obras originales, como pinturas, música y textos. Sin embargo, surge la pregunta de quién es el autor de estas obras: ¿el algoritmo, el programador o el usuario que proporcionó los datos de entrada?
- **Derechos de autor:** La legislación actual de derechos de autor está diseñada para proteger la creatividad humana. ¿Cómo se aplica esta legislación a las obras generadas por máquinas?
- **Plagio involuntario:** La IA Generativa puede generar contenido que sea similar o incluso idéntico a obras existentes, lo que plantea problemas de plagio.

Sesgos algorítmicos:

- **Datos sesgados:** Los modelos de IA aprenden de los datos con los que son entrenados. Si estos datos contienen sesgos, el modelo también los reproducirá.
- **Perpetuación de estereotipos:** Los sesgos algorítmicos pueden perpetuar estereotipos y discriminaciones. Por ejemplo, un modelo de reconocimiento facial puede ser menos preciso en la identificación de personas de color.
- **Decisiones injustas:** Los sesgos algorítmicos pueden llevar a decisiones injustas en áreas como la contratación, la justicia penal y la publicidad.

LAS TRAMPAS DE LA INTELIGENCIA ARTIFICIAL

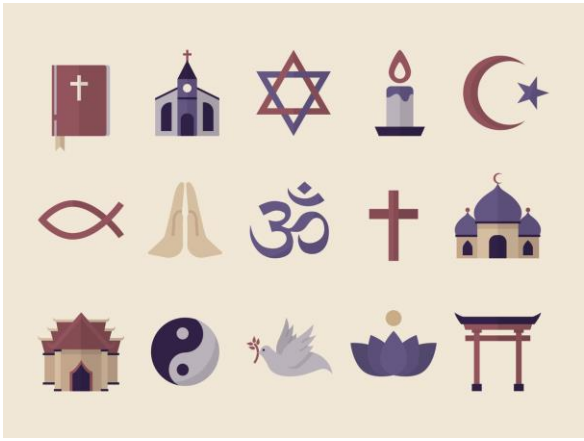


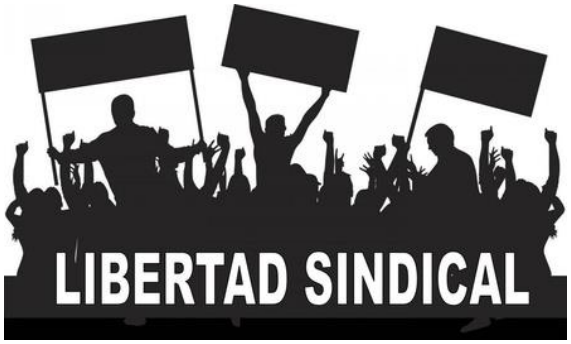
ALIMENTANDO A LA BESTIA CON DATOS SENSIBLES

TECNOIURIS
GABINETE JURÍDICO

¿Qué son los datos sensibles?

TECNOIURIS
GABINETE JURÍDICO





¿Por qué son tan valiosos los datos?

TECNOIURIS
GABINETE JURÍDICO

¿Cómo pueden los Datos Sensibles Filtrarse Accidentalmente en los Modelos de IA?

1. **Datos de entrenamiento contaminados.**
2. **Fugas durante el preprocesamiento.**
3. **Ataques adversariales**

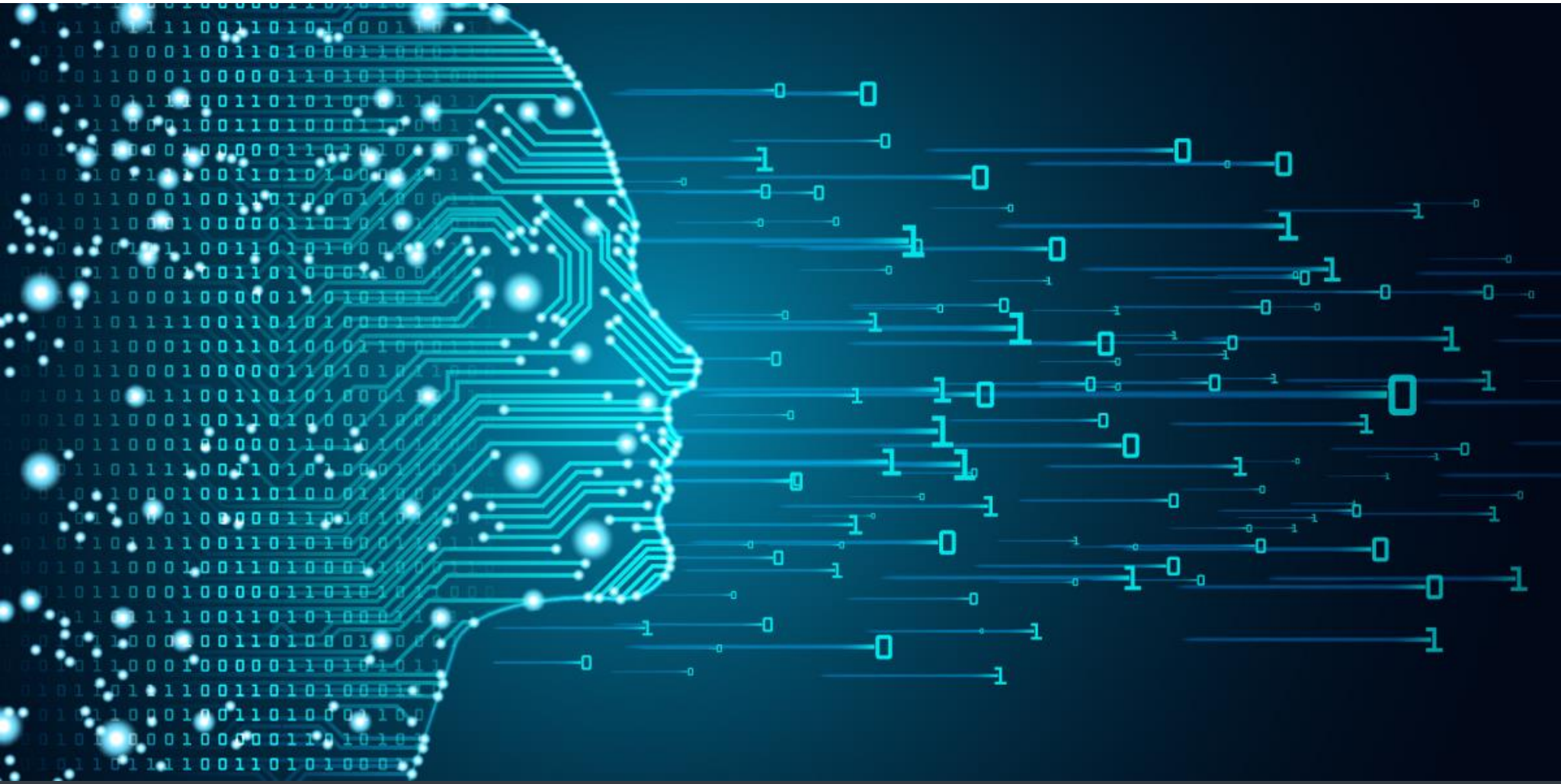
Mitigación de Riesgos

- **Anonimización y Pseudonimización:** Transformar los datos de manera que sea difícil o imposible identificar a los individuos.
- **Diferencial de Privacidad:** Agregar ruido a los datos para proteger la privacidad de los individuos.
- **Homomorfismo:** Realizar cálculos sobre datos cifrados sin descifrarlos.
- **Federated Learning:** Entrenar modelos de IA en datos distribuidos sin compartirlos centralmente.
- **Auditar los Datos de Entrenamiento:** Realizar auditorías exhaustivas de los datos de entrenamiento para identificar y eliminar cualquier información sensible.
- **Evaluación de Impacto:** Realizar evaluaciones de impacto en la protección de datos para identificar y gestionar los riesgos.
- **Consentimiento Informado:** Obtener el consentimiento explícito de los individuos antes de utilizar sus datos para entrenar modelos de IA.
- **Encriptación:** Proteger los datos sensibles mediante encriptación durante todo su ciclo de vida.
- **Acceso Restringido:** Limitar el acceso a los datos sensibles solo a aquellos que lo necesitan.
- **Monitoreo Continuo:** Monitorear continuamente los sistemas de IA para detectar cualquier anomalía o intento de acceso no autorizado.

LA ILUSIÓN DE LA PRIVACIDAD

TECNOIURIS
GABINETE JURÍDICO

El caso de los datos sintéticos



¿Qué son los Datos sintéticos?

Los datos sintéticos son conjuntos de datos creados artificialmente mediante algoritmos que aprenden las características estadísticas de un conjunto de datos real. Estos datos pueden ser numéricos, categóricos o textuales, y se generan con el objetivo de preservar las propiedades estadísticas del conjunto original, pero sin revelar información sobre los individuos específicos que lo componen.

¿Por qué se Utilizan los Datos sintéticos?

- **Protección de la privacidad:** La principal razón para utilizar datos sintéticos es proteger la privacidad de los individuos. Al eliminar información identificable directamente, se reduce el riesgo de que se puedan rastrear los datos hasta su origen.
- **Superación de limitaciones de datos:** Los datos sintéticos pueden utilizarse para complementar conjuntos de datos reales que son escasos o presentan sesgos.
- **Cumplimiento normativo:** Los datos sintéticos pueden ayudar a las organizaciones a cumplir con las regulaciones de protección de datos, como el RGPD.

La Ilusión de la Privacidad: Riesgos y Limitaciones

- **Re-identificación:**
  - **Ataques basados en atributos:** Un atacante puede utilizar información públicamente disponible para identificar a individuos en un conjunto de datos sintéticos. Por ejemplo, si se conoce la edad, la ubicación y la ocupación de una persona, puede ser posible identificar a esa persona en un conjunto de datos sintéticos incluso si se han eliminado los identificadores directos.
  - **Ataques basados en correlaciones:** Los atacantes pueden explotar correlaciones entre los datos sintéticos y otras fuentes de información para identificar a individuos. Por ejemplo, si se sabe que una persona en particular tiene un historial médico específico, se puede buscar un registro en los datos sintéticos que coincida con ese historial.

- **Generación de datos realistas:**
  - **Sesgos inherentes:** Los datos sintéticos pueden heredar los sesgos presentes en los datos originales, lo que puede perpetuar desigualdades y discriminaciones.
  - **Ataques adversariales:** Los atacantes pueden manipular los generadores de datos sintéticos para introducir información falsa o engañosa.

- **Complejidad de los modelos:**
  - **Modelos sofisticados:** Los modelos utilizados para generar datos sintéticos son cada vez más sofisticados, lo que hace que sea más difícil evaluar su calidad y seguridad.

Mitigación de Riesgos

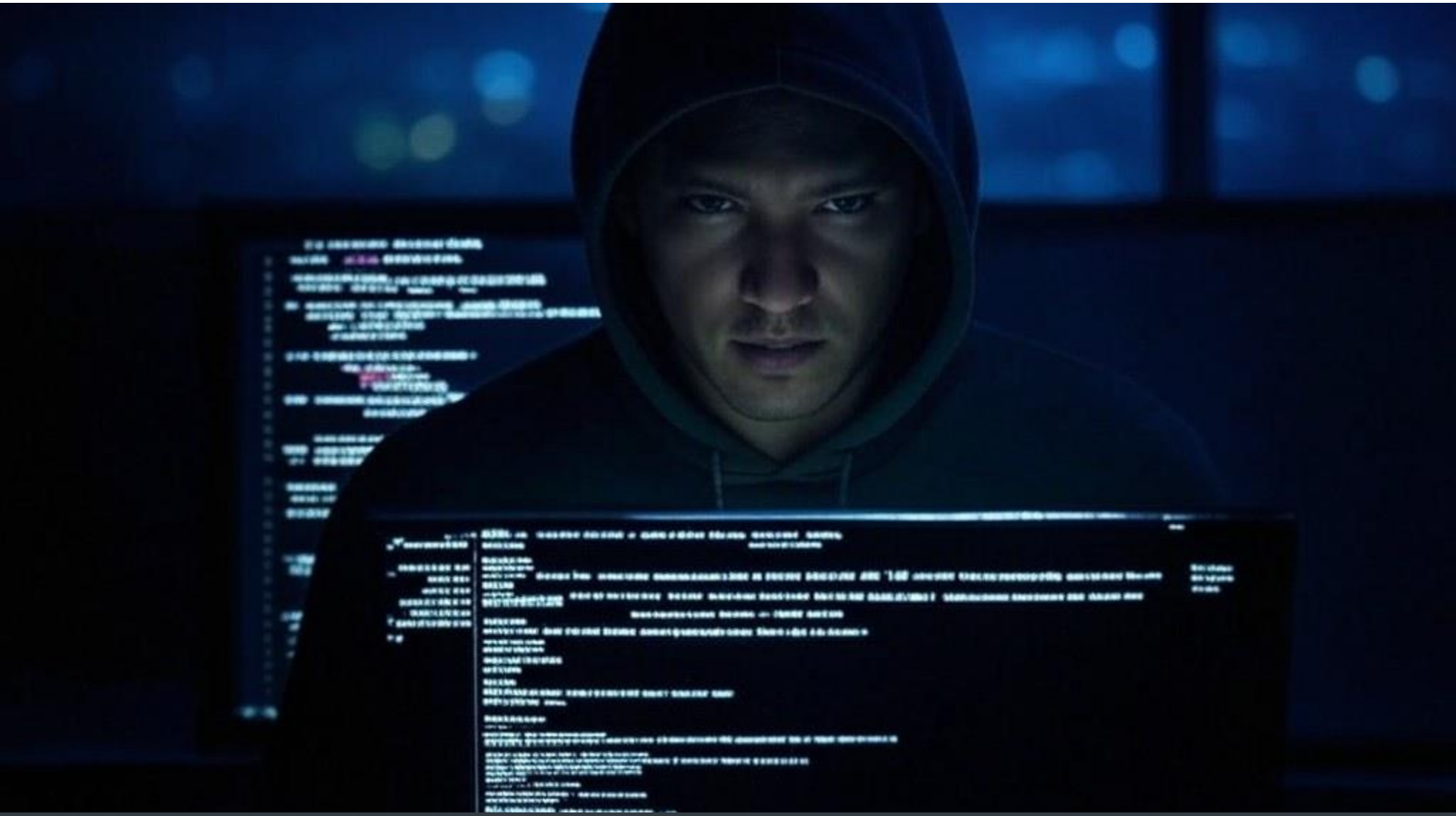
Para mitigar los riesgos asociados con el uso de datos sintéticos, se pueden adoptar las siguientes medidas:

- **Evaluación de la calidad:** Evaluar la calidad de los datos sintéticos para garantizar que sean lo suficientemente realistas como para ser útiles, pero no tan realistas como para permitir la re-identificación.
- **Diversificación de los datos:** Introducir ruido o variabilidad en los datos sintéticos para dificultar la re-identificación.
- **Limitación del acceso:** Restringir el acceso a los datos sintéticos solo a aquellos que lo necesitan.
- **Monitoreo continuo:** Monitorear continuamente los datos sintéticos para detectar cualquier anomalía o intento de ataque.
- **Desarrollo de técnicas de defensa:** Desarrollar nuevas técnicas para proteger los datos sintéticos de los ataques de re-identificación.

LA AMENAZA DE LOS ADVERSARIAL ATTACKS

TECNOIURIS
GABINETE JURÍDICO

Manipulando la Inteligencia Artificial



¿Cómo funcionan los ataques?

Mecanismos comunes de ataque:

- Perturbaciones de entrada: Se añaden pequeñas modificaciones a los datos de entrada (imágenes, audio, texto) para engañar al modelo.
- Envenenamiento de datos: Se introducen datos maliciosos en el conjunto de entrenamiento para corromper el modelo.
- Extracción de modelos: Los atacantes intentan reconstruir el modelo o extraer información sensible a partir de sus salidas.

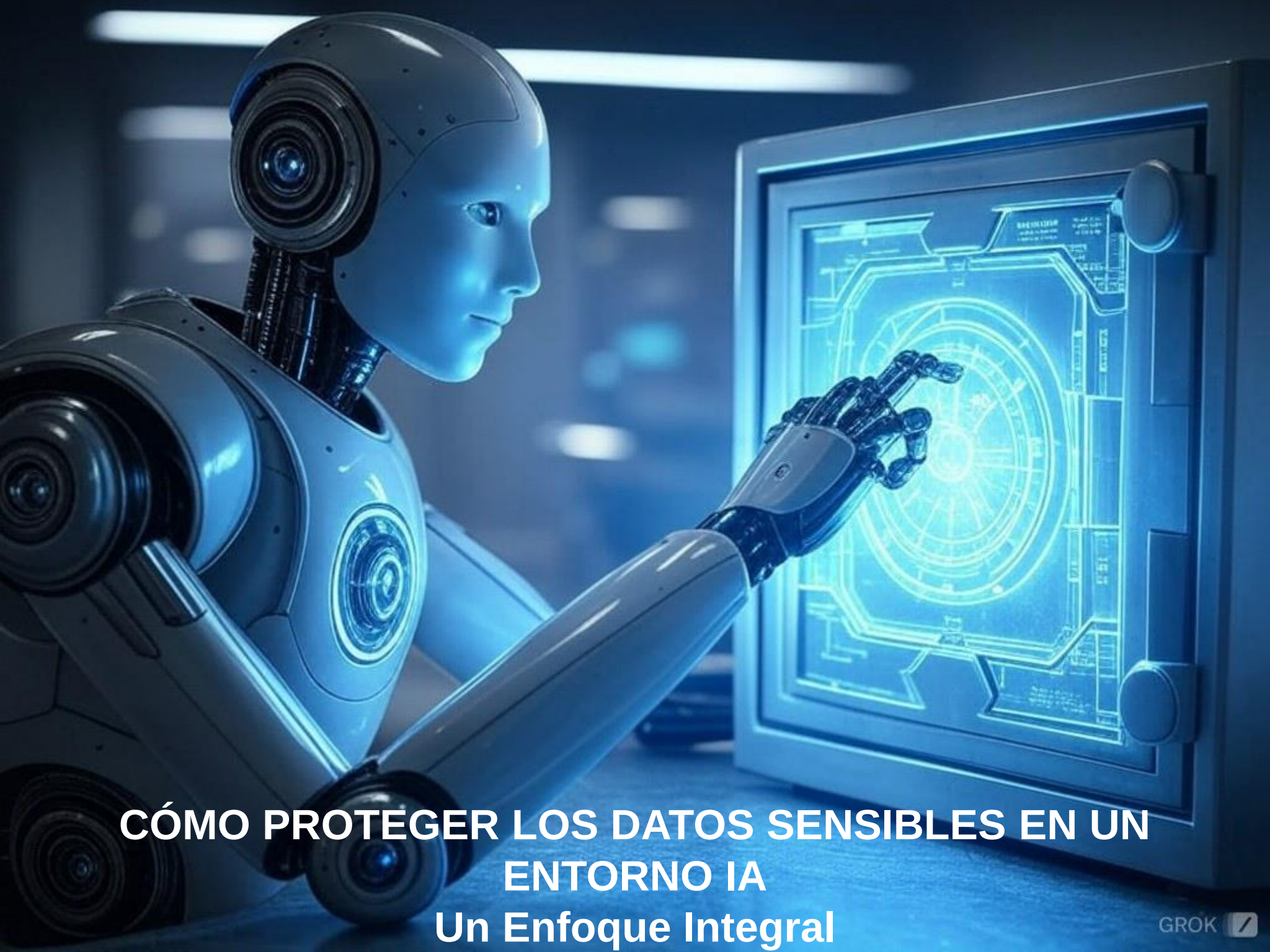
Implicaciones:

- Seguridad: Los ataques adversariales pueden comprometer la seguridad de sistemas críticos como la infraestructura, la defensa y la salud.
- Privacidad: La información confidencial extraída de un modelo puede ser utilizada para fines maliciosos.
- Confianza: Los ataques adversariales pueden erosionar la confianza en los sistemas de IA, dificultando su adopción a gran escala.

Defendiéndonos de los Ataques Adversariales

Si bien los ataques adversariales son una amenaza real, existen varias estrategias para mitigarlos:

- **Adversarial training:** Entrenar los modelos con ejemplos de datos adversariales para que sean más robustos.
- **Regularización:** Introducir restricciones en el modelo para hacerlo menos susceptible a pequeñas perturbaciones.
- **Detección de anomalías:** Identificar las entradas que son inusuales o sospechosas.
- **Verificación de resultados:** Implementar mecanismos para verificar la precisión de las predicciones del modelo.



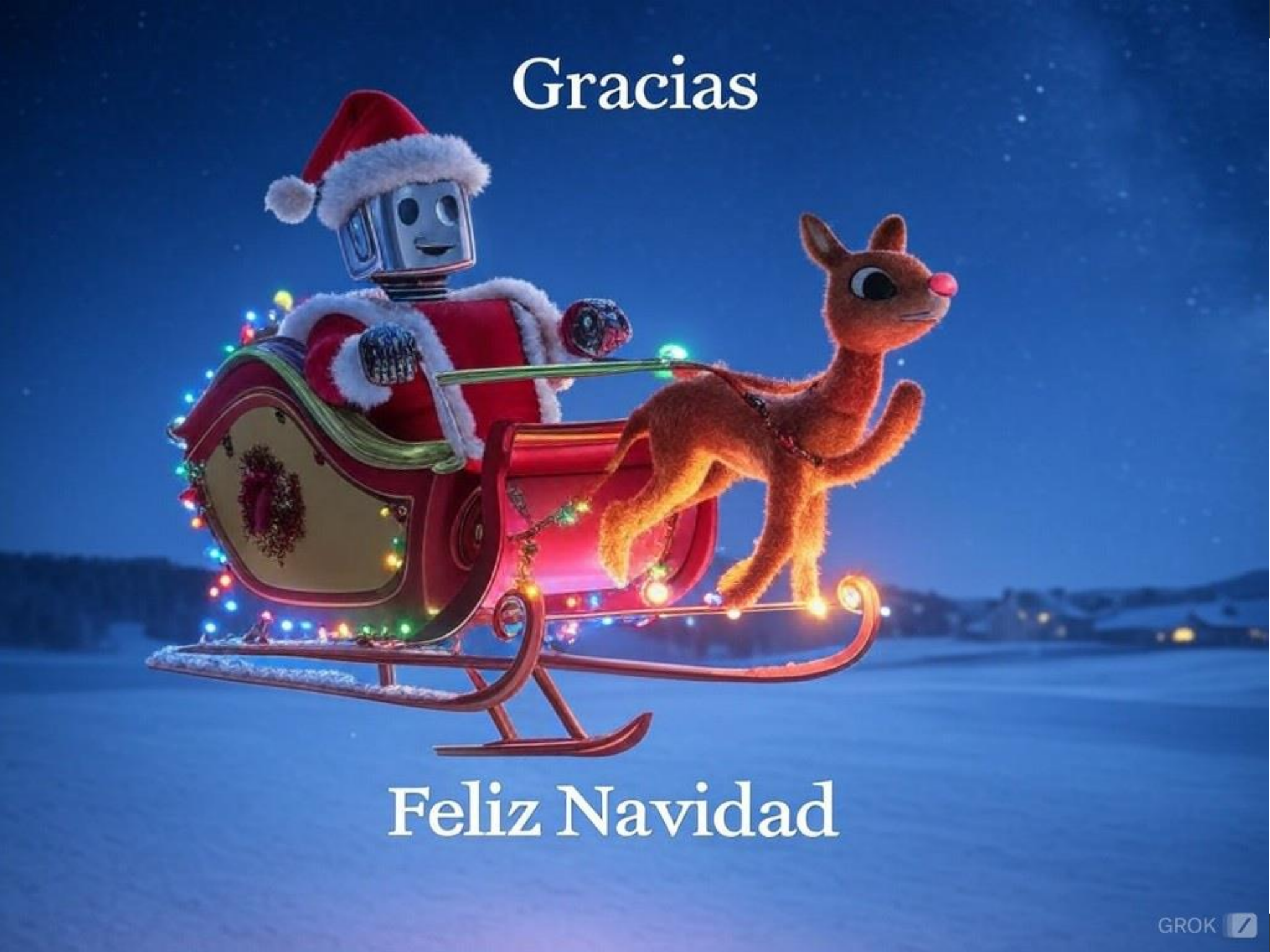
CÓMO PROTEGER LOS DATOS SENSIBLES EN UN ENTORNO IA

Un Enfoque Integral

- MARCO REGULATORIO: RGPD, LOPDyGDD, Ley de Inteligencia Artificial… NORMAS ISO, UNE…

- BUENAS PRÁCTICAS:
- Políticas de uso de la IA en la empresa
- Políticas de seguridad de la información adaptadas a la IA.
- Evaluación de riesgos y mitigación de amenazas.
- Encriptación, anonimización y pseudoanonimización.
- Sistemas de detección de anomalías y respuesta a incidentes.
- Auditorías de seguridad.

- FORMACIÓN Y CONCIENCIACIÓN
- Capacitas a los miembros de la empresa sobre los riesgos de la IA, buenas prácticas y medidas de seguridad.
- Crear una cultura de seguridad de datos en la empresa.


Gracias
Feliz Navidad
GROK

¿PREGUNTAS?

Jacob Peregrina Barahona

@jacobperegrina
www.tecnoiuris.es
tecnoiuris@tecnoiuris.es

TECNOIURIS
GABINETE JURÍDICO

Junta de Andalucía
Consejería de Turismo y Andalucía Exterior

Cámaras Andalucía

enterprise europe network

Red de Información Europea de Andalucía