Inteligencia Artificial y Ciberseguridad

Desafíos y Riesgos en la Era Digital

Elier Cruz | Global Enterprise Security Architect

Septiembre 2024

YOU DESERVE THE BEST SECURITY

Agenda

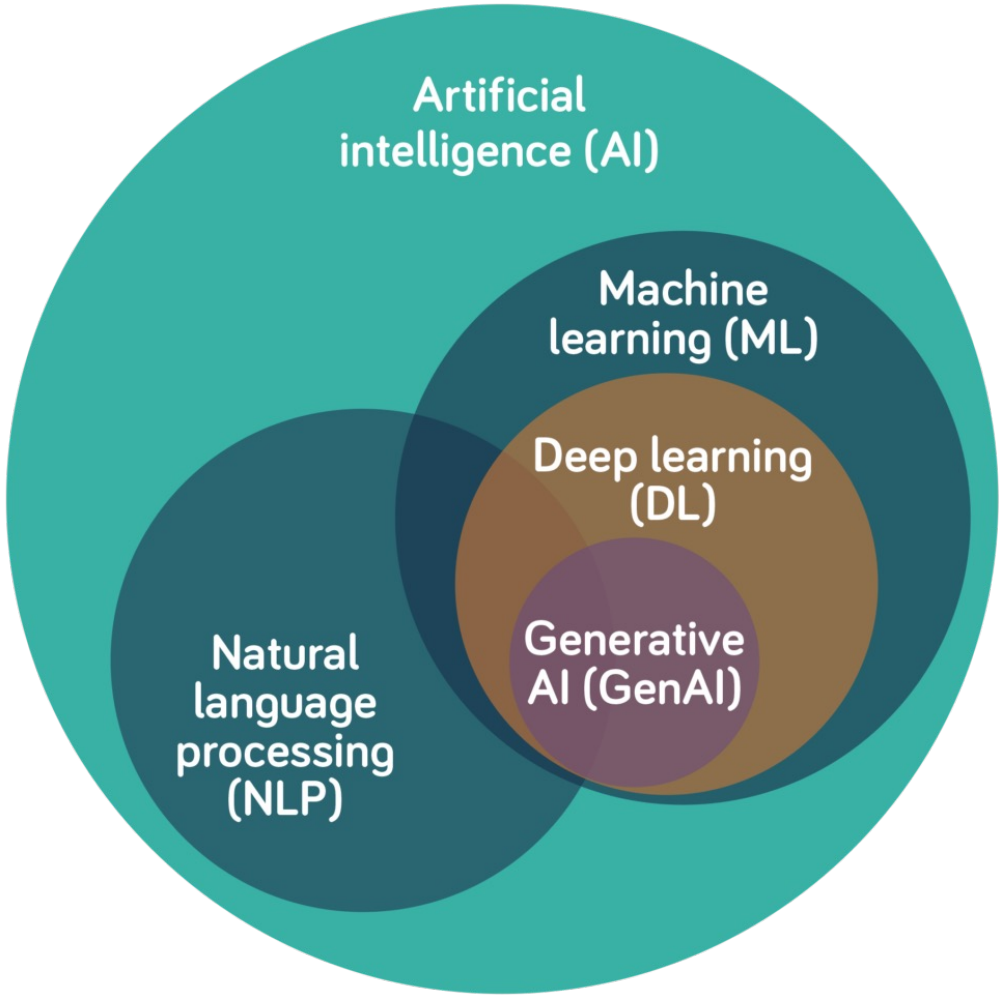
- Historia y evolucion de la IA
- Que problema resuelve la GenAI?
- Usos de la GenAI
- La Industria 5.0 & la GenIA
- Desafios de Ciberseguridad
- Gobernanza, Gestion de Riesgos y Ciberseguridad en la IA

Introducción

- ¿Qué es la GenAI?
- Breve historia y evolución

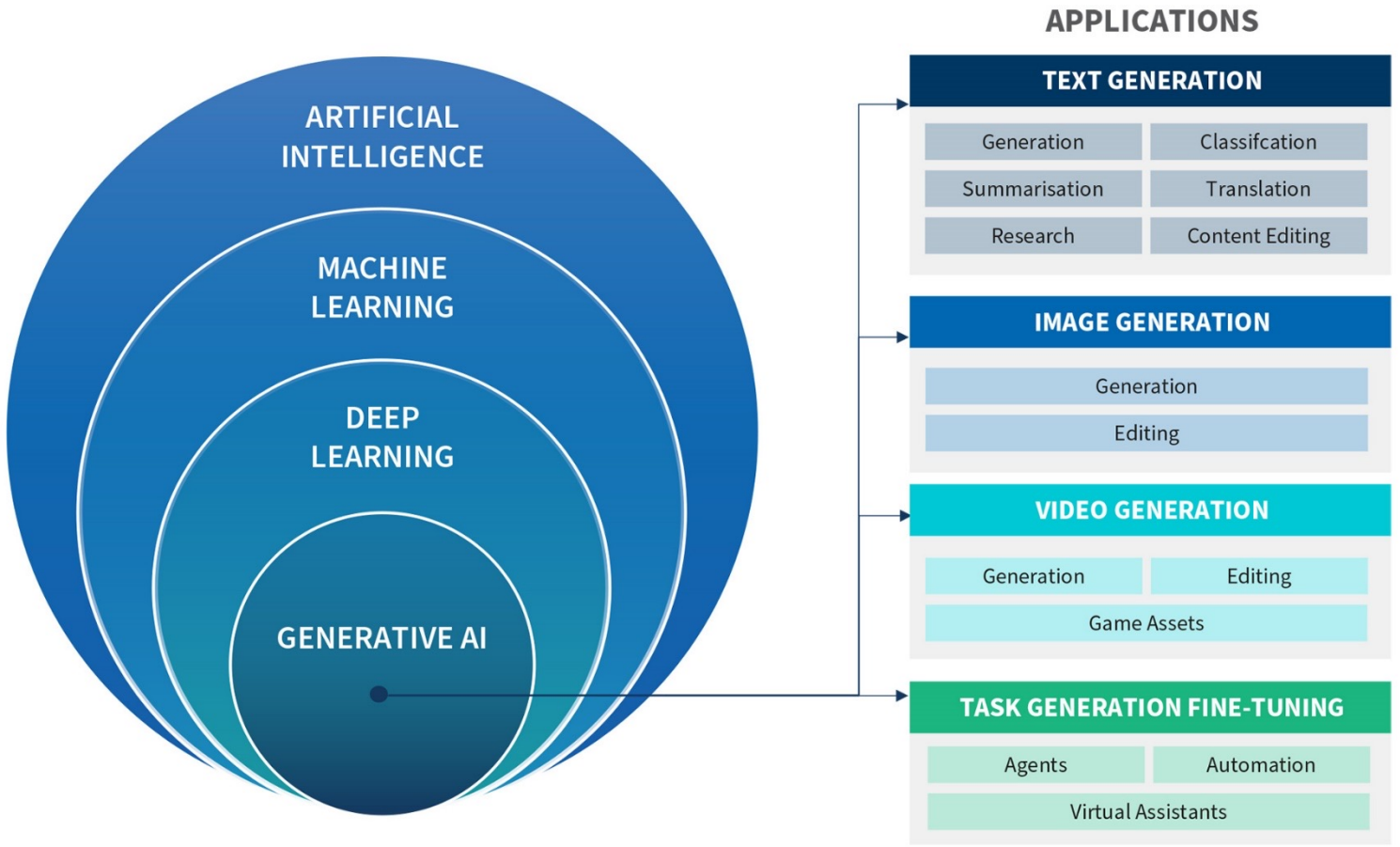
Breve historia y evolución de la IA

- La historia de la Inteligencia Artificial Generativa (GenAI) se remonta a décadas atrás, con pioneros como John von Neumann y Alan Turing, quienes plantearon conceptos fundamentales sobre la creación de máquinas capaces de generar contenido original. En la década de 1950, el primer programa generativo, "Logic Theorist", fue desarrollado por Newell y Simon.
- Sin embargo, el verdadero auge de la GenAI comenzó en la década de 2010 con el desarrollo de las Redes Neuronales Generativas (GANs) por Ian Goodfellow y colaboradores, revolucionando la capacidad de las máquinas para crear arte, música, texto y más.

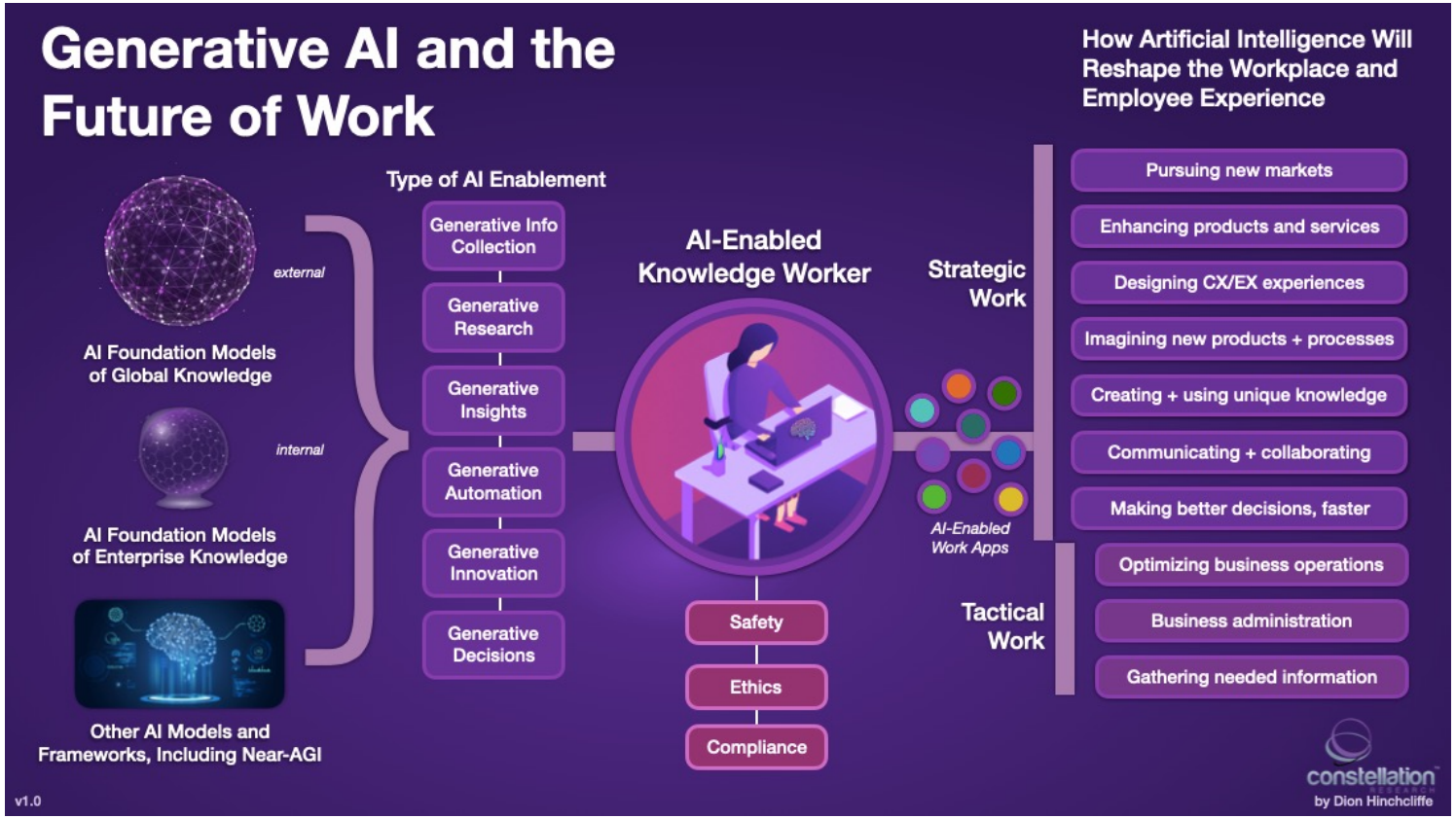


¿Que es la Inteligencia Artificial generativa?

- La Inteligencia Artificial Generativa (GenAI) es una subdisciplina de la inteligencia artificial enfocada en la creación de contenido nuevo y original mediante algoritmos avanzados. A diferencia de otras formas de IA que analizan o clasifican datos, la GenAI se centra en generar datos que no existían previamente. Esto puede incluir imágenes, música, texto, videos y otros tipos de contenido.

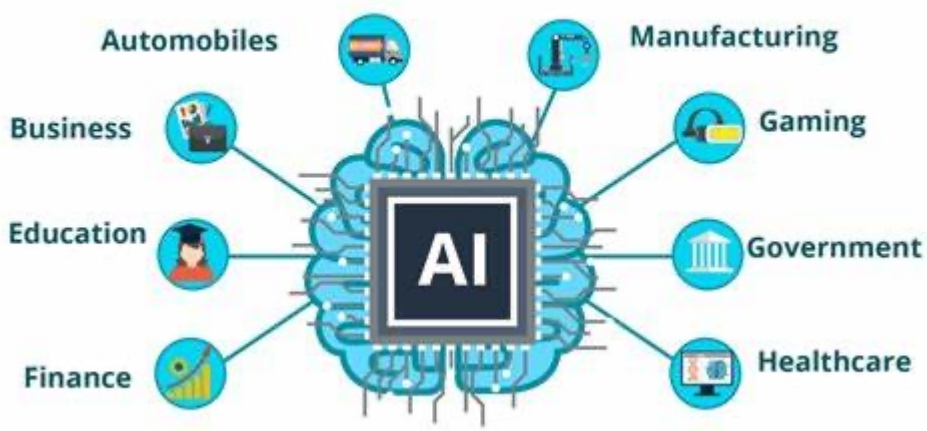


Inteligencia Artificial Generativa y el trabajo en el futuro

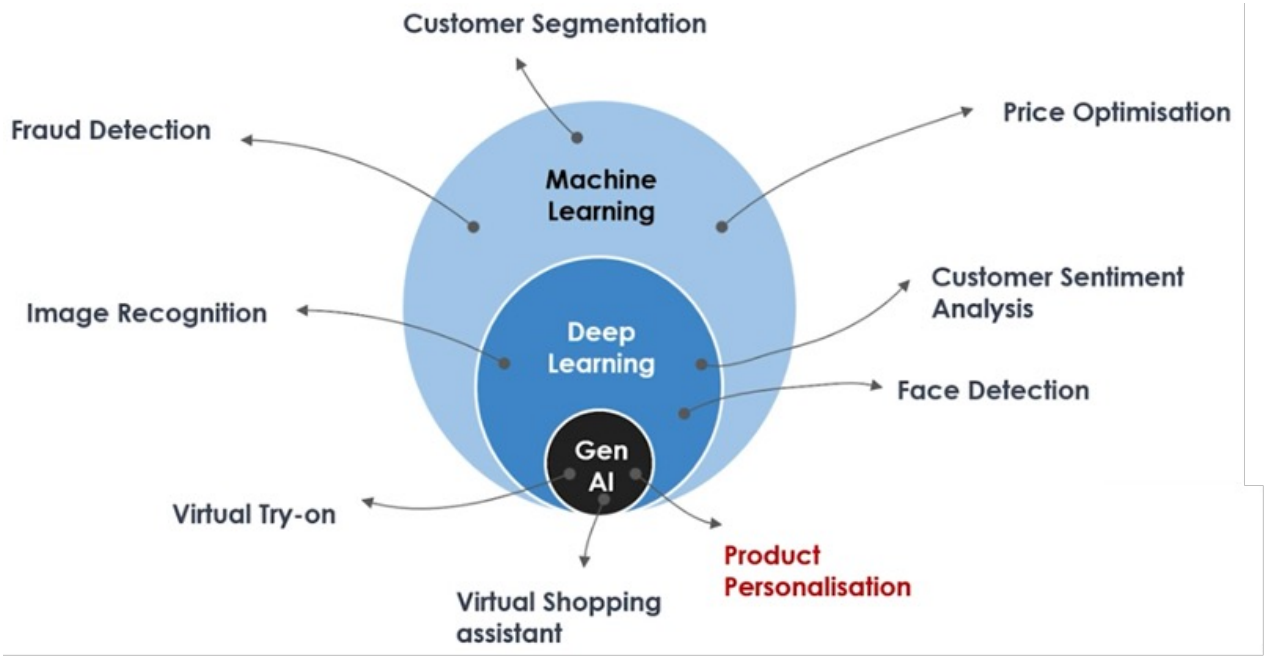


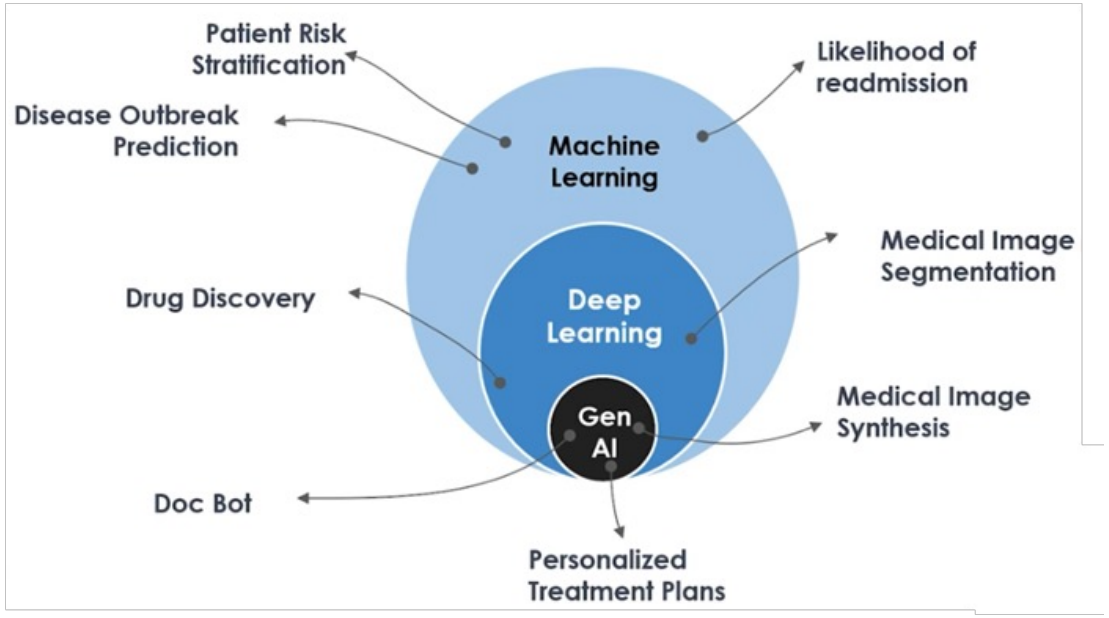
Usos de la IA Generativa

- Arte y creatividad
- Generación de texto y contenido
- Simulación y modelado
- Uso en Ventas & Asesoramiento
- Investigacion
- Creación de Código Fuente para nuevo softw
- Aplicaciones en medicina y biotecnología
- Industria 5.0



Problemas de Negocios resueltos por GenAI

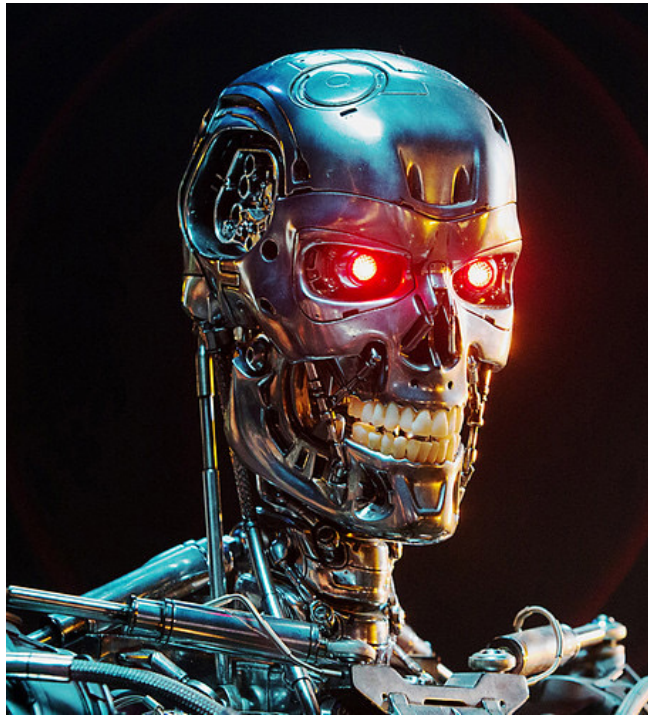


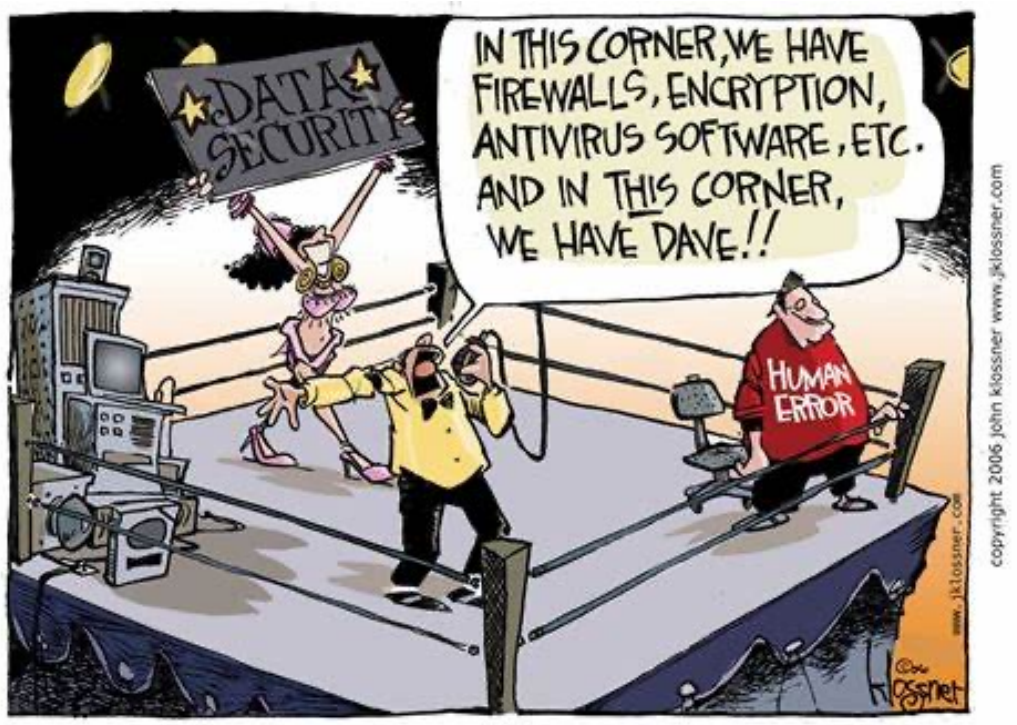


Industria 5.0: Evolución de la Industria 4.0 con GenIA Generativa

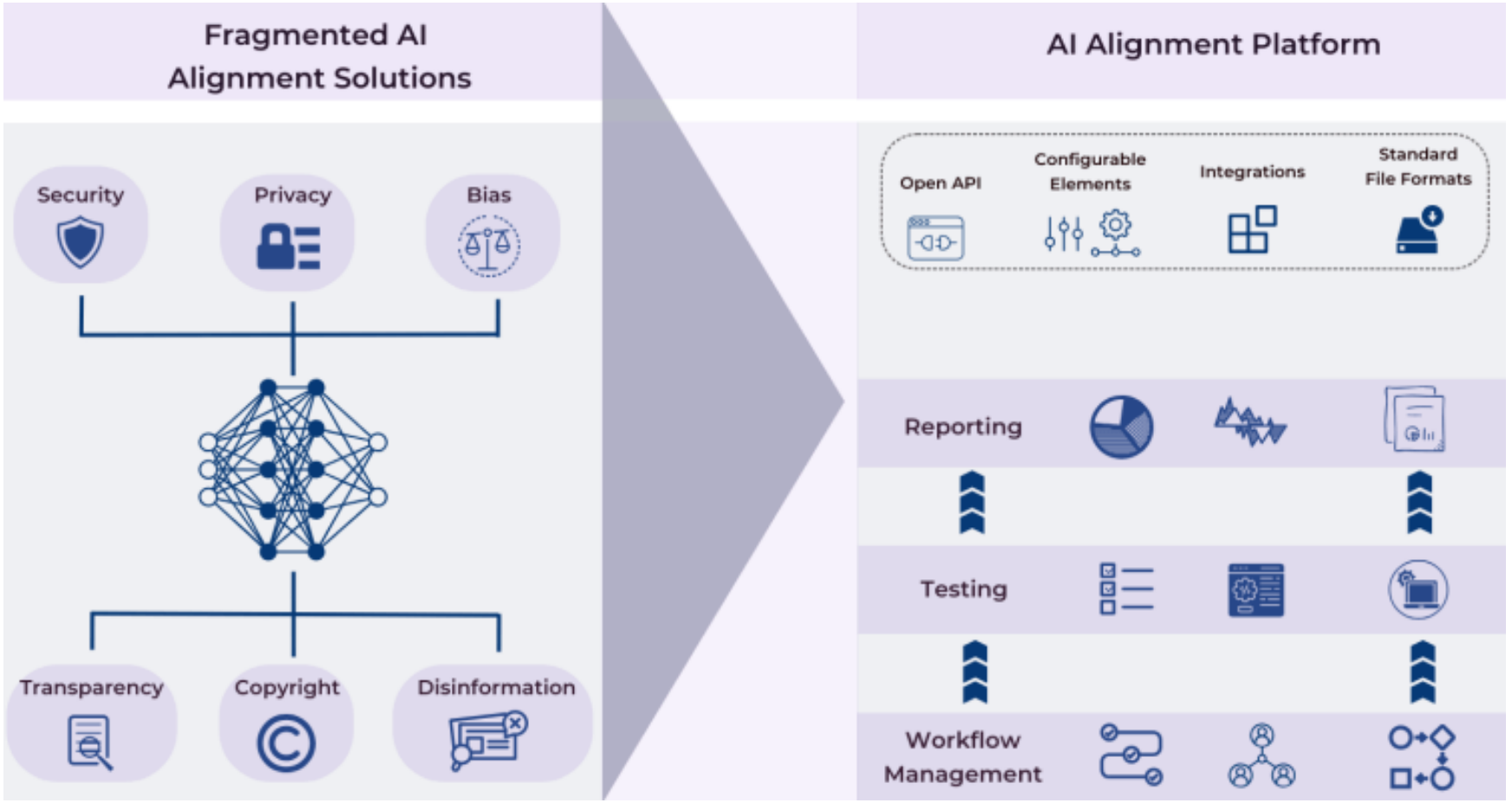
- La Industria 5.0 se centra en la colaboración entre humanos y máquinas inteligentes.
- IA Generativa permite la personalización a gran escala y mayor flexibilidad en la producción.
- Aplicaciones en diseño de productos, personalización masiva y fabricación autónoma.
- Impulso hacia la 'human-centric manufacturing', con robots y sistemas inteligentes que trabajan junto a humanos.
- Retos: interoperabilidad, seguridad de datos, y el equilibrio entre automatización y creatividad humana.

¿Quién es más peligroso?



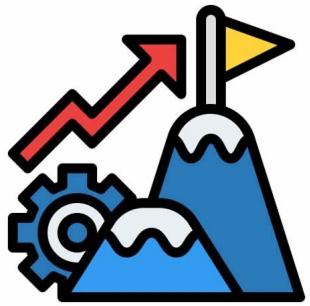


Desafíos de Seguridad con la IA Generativa



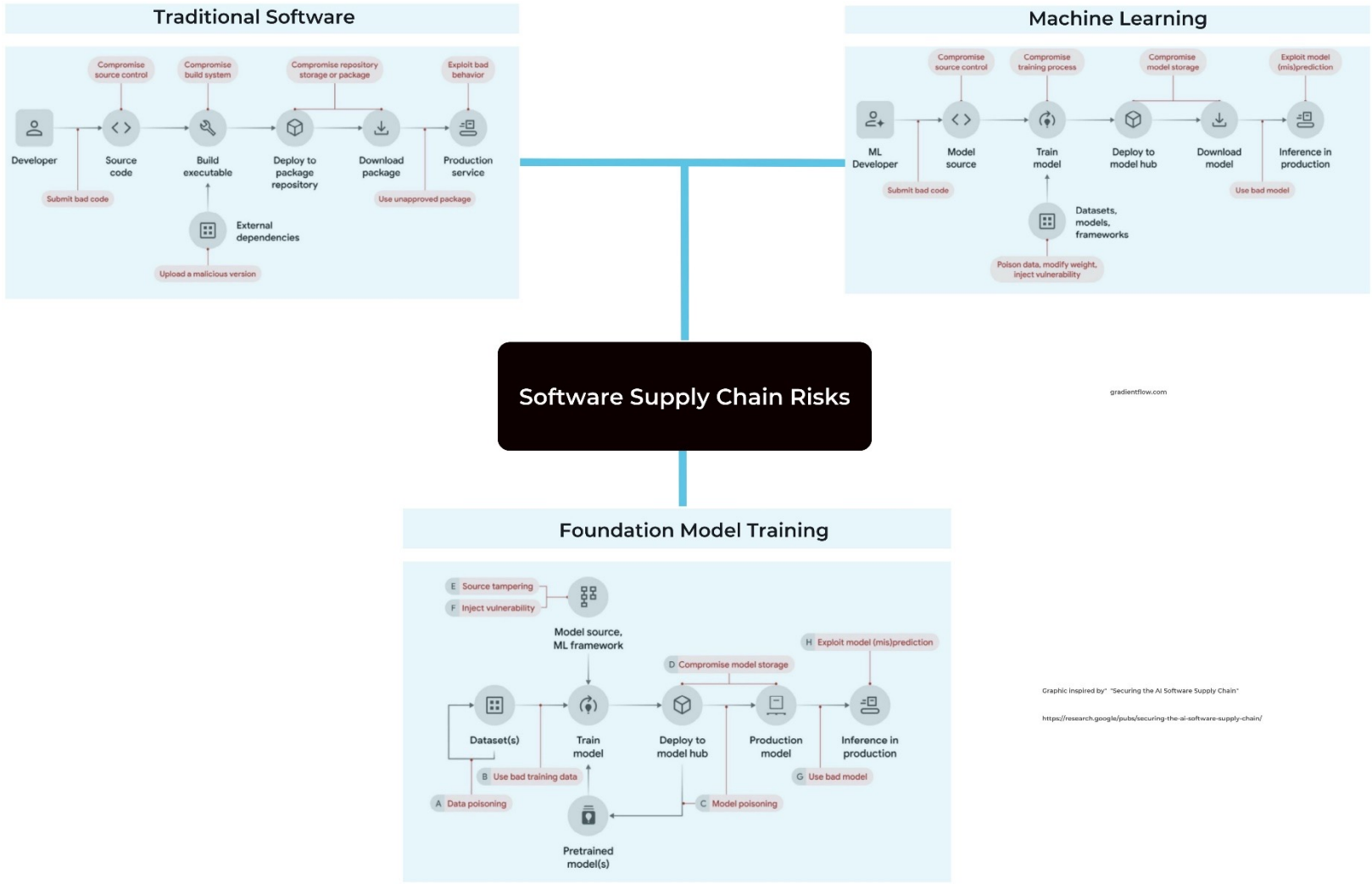
Desafíos de Seguridad con la IA Generativa

- Riesgos de la manipulación de datos
- Riesgos en la manipulación de imágenes y vídeos.
- Desafíos en la detección de contenido falso.
- Posibles impactos en la privacidad

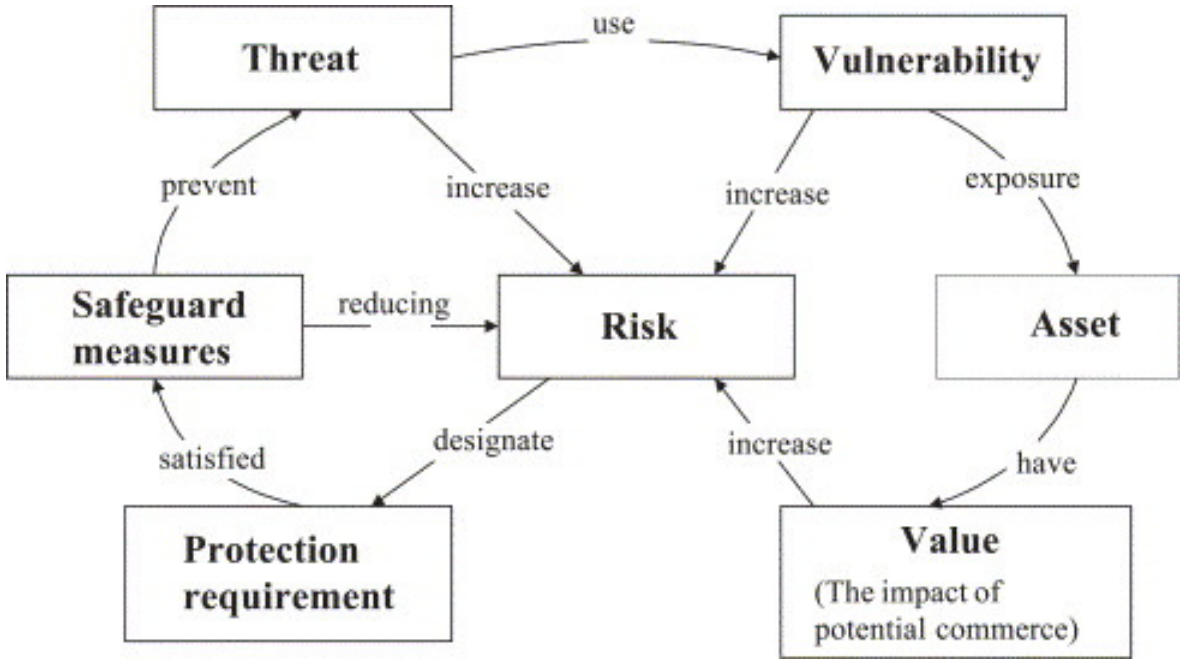


1. **Prevención de Manipulación de Contenido**:
    1. Las tecnologías de GenAI pueden generar contenido extremadamente realista, incluyendo imágenes, vídeos y textos. Sin medidas de seguridad adecuadas, existe el riesgo de que se utilicen para crear contenido falso o engañoso, como deepfakes, que pueden dañar la reputación de individuos y organizaciones.
2. **Protección de Datos Personales**:
    1. Los modelos de GenAI a menudo se entrenan con grandes volúmenes de datos, que pueden incluir información personal y sensible. Es vital asegurar que estos datos estén protegidos contra accesos no autorizados y que los modelos no generen información que pueda comprometer la privacidad de las personas.
3. **Integridad y Confiabilidad de los Sistemas**:
    1. Asegurar la integridad de los sistemas de GenAI es esencial para mantener la confianza de los usuarios. Cualquier manipulación o ataque que comprometa la integridad de los modelos puede resultar en decisiones incorrectas o perjudiciales, especialmente en aplicaciones críticas como la medicina o la seguridad pública.
4. **Mitigación de Riesgos Éticos y Legales**:
    1. La implementación de medidas de seguridad robustas ayuda a mitigar riesgos éticos y legales. Garantiza que las tecnologías de GenAI se utilicen de manera responsable y conforme a regulaciones y normativas, evitando así posibles sanciones y problemas legales.
5. **Fomento de la Innovación Responsable**:
    1. Al establecer un entorno seguro y confiable, se fomenta la innovación responsable en el desarrollo y despliegue de GenAI. Esto permite explorar nuevas aplicaciones y beneficios de la tecnología mientras se minimizan los riesgos asociados.

Riesgos en el Supply Chain



OWASP Top 10 for LLM



OWASP Top 10 for LLM

This is a draft list of important vulnerability types for Artificial Intelligence (AI) applications built on Large Language Models (LLMs).

LLM01: Prompt Injections

Prompt Injection Vulnerabilities in LLMs involve crafty inputs leading to undetected manipulations. The impact ranges from data exposure to unauthorized actions, serving attacker's goals.

LLM02: Insecure Output Handling

These occur when plugins or apps accept LLM output without scrutiny, potentially leading to XSS, CSRF, SSRF, privilege escalation, remote code execution, and can enable agent hijacking attacks.

LLM03: Training Data Poisoning

LLMs learn from diverse text but risk training data poisoning, leading to user misinformation. Overreliance on AI is a concern. Key data sources include Common Crawl, WebText, OpenWebText, and books.

LLM04: Denial of Service

An attacker interacts with an LLM in a way that is particularly resource-consuming, causing quality of service to degrade for them and other users, or for high resource costs to be incurred.

LLM05: Supply Chain

LLM supply chains risk integrity due to vulnerabilities leading to biases, security breaches, or system failures. Issues arise from pre-trained models, crowdsourced data, and plugin extensions.

LLM06: Permission Issues

Lack of authorization tracking between plugins can enable indirect prompt injection or malicious plugin usage, leading to privilege escalation, confidentiality loss, and potential remote code execution.

LLM07: Data Leakage

Data leakage in LLMs can expose sensitive information or proprietary details, leading to privacy and security breaches. Proper data sanitization, and clear terms of use are crucial for prevention.

LLM08: Excessive Agency

When LLMs interface with other systems, unrestricted agency may lead to undesirable operations and actions. Like web-apps, LLMs should not self-police; controls must be embedded in APIs.

LLM09: Overreliance

Overreliance on LLMs can lead to misinformation or inappropriate content due to "hallucinations." Without proper oversight, this can result in legal issues and reputational damage.

LLM10: Insecure Plugins

Plugins connecting LLMs to external resources can be exploited if they accept free-form text inputs, enabling malicious requests that could lead to undesired behaviors or remote code execution.

Riesgos de Seguridad en los LLMs

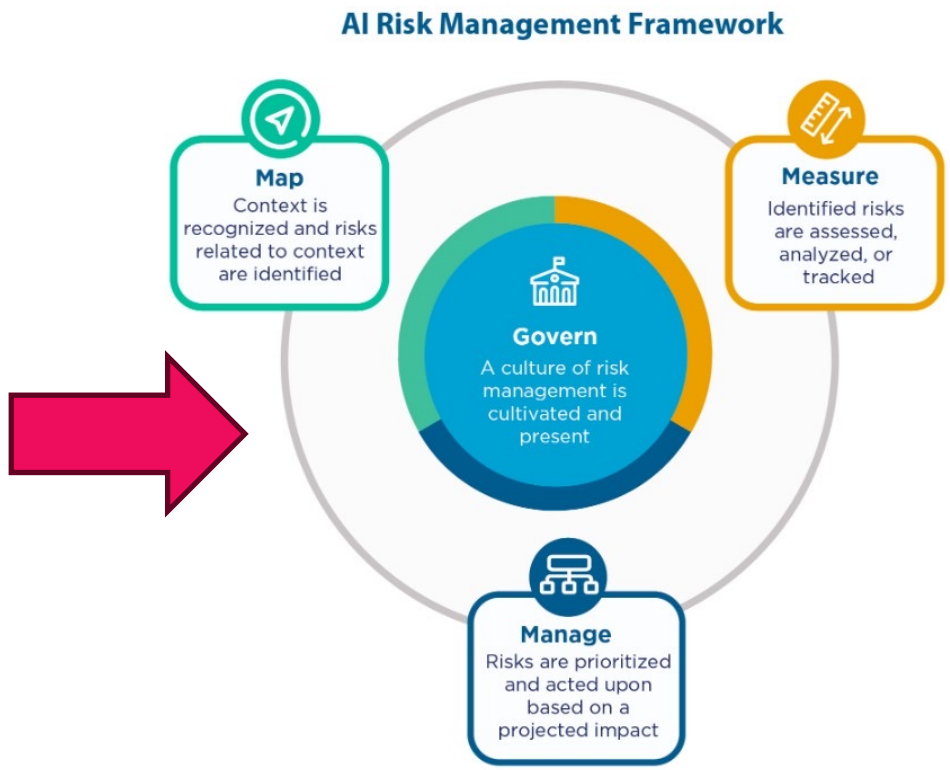
Los proyectos de IA que utilizan Large Language Models (LLM) enfrentan a numerosos riesgos de seguridad si no se protegen adecuadamente.

Los principales problemas son los siguientes

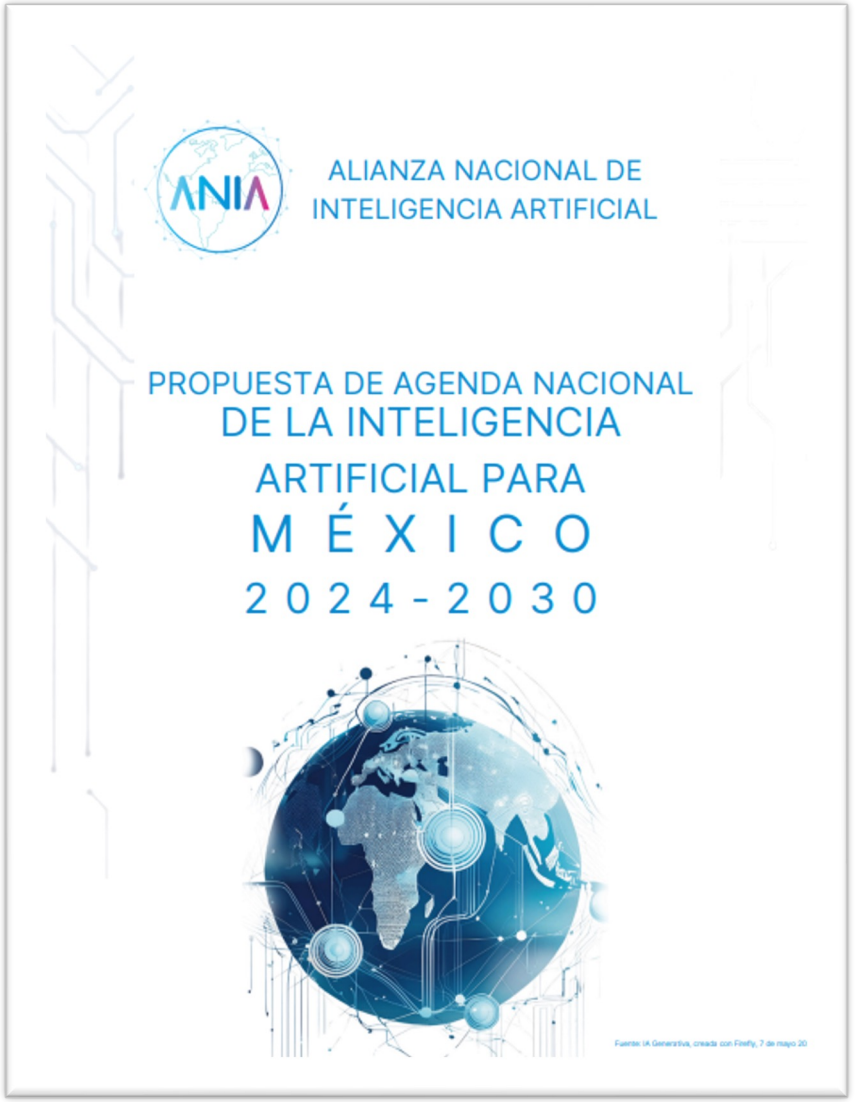
- **Prompt Injection:** Los atacantes pueden manipular el comportamiento del modelo a través de entradas manipuladas, dando lugar a un acceso no autorizado a los datos o a salidas no deseadas.
- **Gestión insegura de los resultados:** Las salidas sin cifrar pueden exponer a los sistemas backend a riesgos como XSS, CSRF o ejecución remota de código.
- **Envenenamiento de datos de entrenamiento:** Agentes malintencionados pueden manipular los datos de entrenamiento, introduciendo vulnerabilidades que comprometan la integridad del modelo.
- **Revelación de información sensible:** Los LLM pueden revelar inadvertidamente datos confidenciales en sus respuestas, provocando violaciones de la privacidad.
- **Denegación de servicio del LLM:** Sobrecargar el modelo con peticiones puede interrumpir su disponibilidad.
- **Vulnerabilidades de la cadena de suministro:** La dependencia de plugins o conjuntos de datos de terceros puede introducir riesgos adicionales

AI RISK Management Framework 1.0 by NIST

1. **Map (Mapear)**:
    1. **Descripción**: Reconocer el contexto y los riesgos relacionados con ese contexto.
    2. **Objetivo**: Identificar y entender el entorno en el que la IA opera y los posibles riesgos asociados.
2. **Measure (Medir)**:
    1. **Descripción**: Los riesgos identificados son evaluados, analizados o monitoreados.
    2. **Objetivo**: Evaluar la magnitud y la probabilidad de los riesgos para poder analizarlos adecuadamente y realizar un seguimiento.
3. **Manage (Gestionar)**:
    1. **Descripción**: Los riesgos se priorizan y se actúa sobre ellos en función del impacto proyectado.
    2. **Objetivo**: Implementar acciones para mitigar, transferir, aceptar o evitar los riesgos priorizados según su impacto potencial.
4. **Govern (Gobernar)**:
    1. **Descripción**: Se cultiva y mantiene una cultura de gestión de riesgos.
    2. **Objetivo**: Establecer y mantener una cultura organizacional que valore y practique la gestión proactiva de riesgos en todas las actividades relacionadas con la IA.



Mexico y la Inteligencia Artificial



El viaje en la Transformación de la IA

Consultoría sobre gobernanza y cumplimiento de la IA

Objetivo:

Establecer estructuras y procesos de gobernanza claros para supervisar las iniciativas de IA.

Consultoría de gestión de riesgos en IA

Objetivo:

Identificar, evaluar y mitigar los riesgos asociados a los sistemas de IA a lo largo de su ciclo de vida.

CDAIO

Consultoría de Responsabilidades en la Organización para IA

Objetivo:

- Definir y asignar funciones y responsabilidades específicas para la gestión de las iniciativas de IA.

Consultoría de Seguridad en Inteligencia Artificial

Objetivo:

Proteger los sistemas de IA de las amenazas a la seguridad y garantizar la integridad, confidencialidad y disponibilidad de los datos y procesos de IA.

Conclusiones

- La Inteligencia Artificial Generativa (GenAI) representa un avance disruptivo en la capacidad de crear contenido nuevo y realista, incluyendo texto, imágenes, y código, a partir de datos existentes. Su uso está revolucionando sectores como la creación de contenido, el diseño y la investigación científica.
- Sin embargo, plantea desafíos de seguridad, como el uso malintencionado en campañas de desinformación o generación de contenido engañoso.
- Su implementación debe equilibrar el potencial innovador con la gestión de riesgos éticos y de ciberseguridad. La regulación y el desarrollo de mecanismos de control son esenciales para su adopción segura y responsable.

CHECK POINT

¡Muchas Gracias!

YOU DESERVE THE BEST SECURITY