INFORME TÉCNICO

Ciberseguridad en Inteligencia Artificial aplicada al sector industrial

Juan Pablo Fuentes, PhD

Head of Artificial Intelligence & Cybersecurity

Minsait Cyber

CÁTEDRA DE INDUSTRIA INTELIGENTE

COMILLAS ICAI

ACERINOX · ANTOLIN Intelligent. Integrated. Inside. · enagas · endesa · Gestamp

IBM · INDRA GROUP INDRA MINSAIT · PLADUR · PROSEGUR · REPSOL

INFORME TÉCNICO

Ciberseguridad en Inteligencia Artificial aplicada al sector industrial

Juan Pablo Fuentes, PhD
Head of Artificial Intelligence & Cybersecurity
Minsait Cyber

CÁTEDRA DE INDUSTRIA INTELIGENTE
COMILLAS ICAI

Informe Técnico preparado por la Cátedra de Industria Inteligente para sus Empresas Patrono

Mayo 2025

Versión: 1.0

Titularidad y responsabilidad

TABLA DE CONTENIDO

Ciberseguridad en Inteligencia Artificial aplicada al sector industrial

La Inteligencia Artificial (IA) está cada vez más integrada en aplicaciones y sectores industriales, ante tal escenario surge la necesidad crítica de abordar los desafíos de ciberseguridad asociados; es de vital importancia garantizar que los sistemas de IA sean robustos, confiables y éticos.

La ciberseguridad aplicada a la IA se ha convertido en un campo de investigación fundamental para garantizar la seguridad de los sistemas IA (a nivel de datos y modelos), por lo tanto, es muy importante conocer y protegerse ante las principales vulnerabilidades y ataques asociados con los modelos de IA, así como las estrategias para mitigar estos riesgos.

Con la creciente complejidad de los modelos de IA y la diversidad de aplicaciones, esto ha supuesto la aparición de nuevas superficies de ataque. Los atacantes pueden explotar vulnerabilidades sobre los algoritmos, datos y modelos para realizar ataques mucho más dirigidos, como la manipulación de datos de entrada para engañar al sistema IA, la introducción de sesgos maliciosos o incluso el sabotaje en el comportamiento de los modelos IA desplegados en la industria.

Por todo ello, es esencial desarrollar modelos de IA resistentes a dichos ataques, mediante la implementación de frameworks de ciberseguridad que cubran todo el ciclo de vida de los sistemas IA, desde el pipeline de datos, entornos de entrenamiento y su posterior despliegue en los entornos productivos.

El presente informe presenta una visión de todos estos puntos, con el objetivo de proteger todo el proceso desde un punto de vista de ciberseguridad.

1 INTRODUCCIÓN A LA CIBERSEGURIDAD EN SISTEMAS DE IA

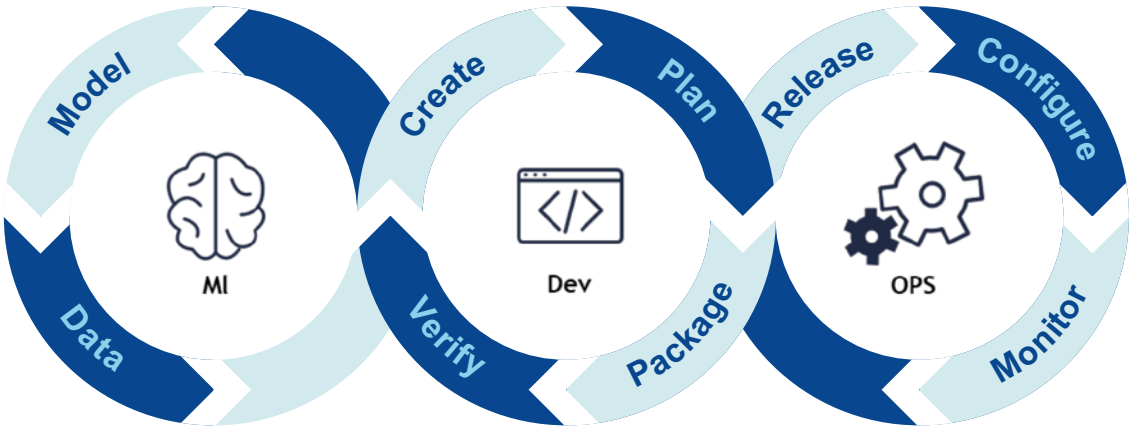
La integración de la IA en entornos industriales ha revolucionado la eficiencia operativa, la toma de decisiones y la automatización de procesos. Sin embargo, esta adopción también ha introducido nuevas superficies de ataque que requieren de una reevaluación de los fundamentos de ciberseguridad más tradicionales. La ciberseguridad en sistemas que incorporan IA debe considerar no solo los vectores clásicos de amenaza, sino también las vulnerabilidades propias de los algoritmos, los modelos de aprendizaje y los conjuntos de datos utilizados.

Uno de los principios clave es la protección de los modelos de IA, que pueden ser blanco de ataques como la ingeniería inversa o la modificación de sus comportamientos (*evasión attacks*), así como aquellos ataques que impactan directamente sobre los datos como son el envenenamiento de datos

(*data poisoning attacks*). Estos ataques pueden comprometer la integridad de los modelos, manipular sus resultados o incluso usarlos para obtener información confidencial del sistema. Por ello, es esencial implementar técnicas de defensa como el control de acceso, la monitorización del comportamiento del modelo y la validación constante de los datos de entrada.

Además, se debe considerar la seguridad a lo largo del ciclo de vida del modelo de IA, desde su diseño, entrenamiento hasta su despliegue en producción. Esto implica asegurar la procedencia y calidad de los datos, garantizar la trazabilidad de las decisiones automatizadas, y prevenir manipulaciones durante el entrenamiento o actualizaciones del modelo.

Las prácticas de DevSecOps y MLOps pueden ayudar a integrar la seguridad de forma continua en estos procesos, basándose en marcos de trabajo como el Framework AI de NIST; en la siguiente figura aparecen de forma secuencial las principales etapas por las cuales evoluciona un sistema IA, los cuales deberán implementar controles de ciberseguridad para garantizar su seguridad:



.

Ilustración 1: ***Proceso MLOps: Pipeline de datos + Entrenamiento + Despliegue.***

Otro componente fundamental es la resiliencia ante ataques adversarios, que explotan debilidades del modelo para inducir decisiones erróneas mediante pequeñas perturbaciones en los datos de entrada. En el ámbito industrial, este tipo de ataques puede tener consecuencias graves, como interrupciones en la producción o fallos en sistemas críticos. Para contrarrestarlos, es necesario desarrollar modelos robustos, entrenados con ejemplos adversarios, y utilizar mecanismos de detección de anomalías.

La ciberseguridad en sistemas IA también debe estar alineada con marcos normativos y éticos, así como con las regulaciones existentes (como por ejemplo AI Act). Es crucial establecer políticas de gobernanza de datos, mecanismos de auditoría y transparencia en las decisiones automatizadas, especialmente cuando los sistemas afectan la seguridad física de personas o infraestructuras industriales críticas. De esta forma, los fundamentos de ciberseguridad en sistemas IA deben ampliarse para abarcar los riesgos específicos que introduce la IA, combinando las estrategias tradicionales con nuevas prácticas enfocadas en proteger los modelos y los datos empleados para su entrenamiento.

2 NUEVAS SUPERFICIES DE ATAQUE EN SISTEMAS DE IA

La IA como toda nueva tecnología trae de la mano nuevas amenazas de ciberseguridad, y como consecuencia nuevas superficies de ataque que hay que proteger y monitorizar.

Entre las vulnerabilidades más críticas que pueden afectar a los sistemas IA se encuentran las englobadas en las siguientes categorías de ataques:

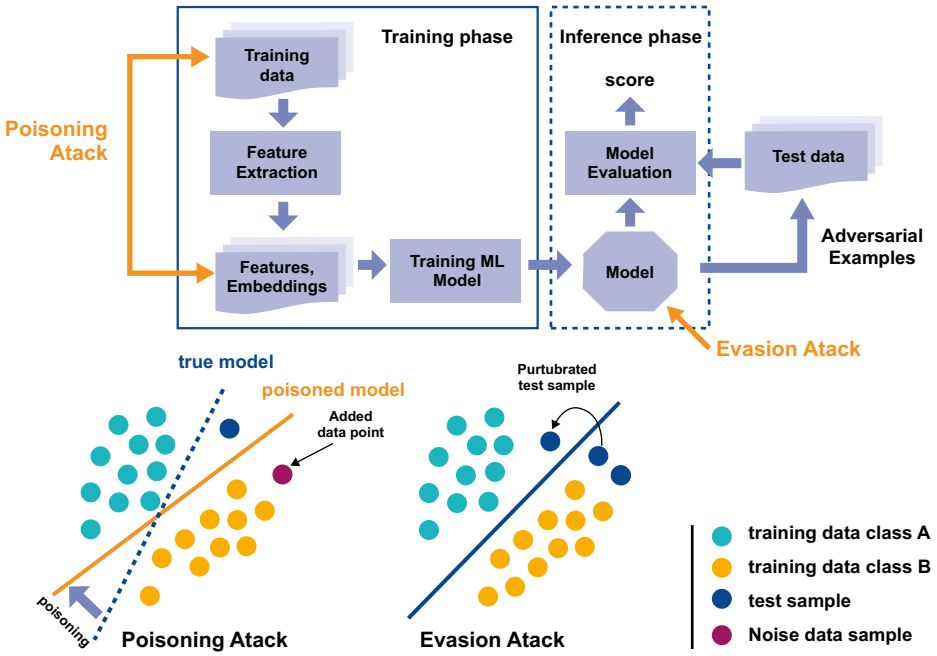


Ilustración 2: ***Ataques de envenenamiento y evasión.***

- Ataques de envenenamiento (Poisoning Attack): se basan en realizar cambios maliciosos sobre los datos de entrenamiento de los modelos, con el objetivo de introducir sesgos y llegar a provocar cambios en el comportamiento de los modelos difíciles de detectar. Dichos ataques se realizan en una fase previa al entrenamiento de los modelos IA.
- Ataques de evasión (Evasion Attack): el riesgo de este tipo de ataques es el posible robo de las reglas de negocio implementadas por el modelo, obtención de los datos reales con los cuales se entrenó y llegando incluso a provocar sabotajes sobre su comportamiento en beneficio del atacante. Los ataques de evasión impactan sobre los modelos IA ya entrenados y desplegados.

Para establecer una taxonomía detallada de todas estas nuevas superficies de ataque, el proyecto AI OWASP ha definido una guía en donde se describen cada una de ellas, además de indicar los principales riesgos de ciberseguridad que pueden impactar sobre los casos de uso de IA. La siguiente figura muestra sobre la arquitectura lógica de un sistema IA, cómo sería el impacto de estos nuevos vectores de ataque:

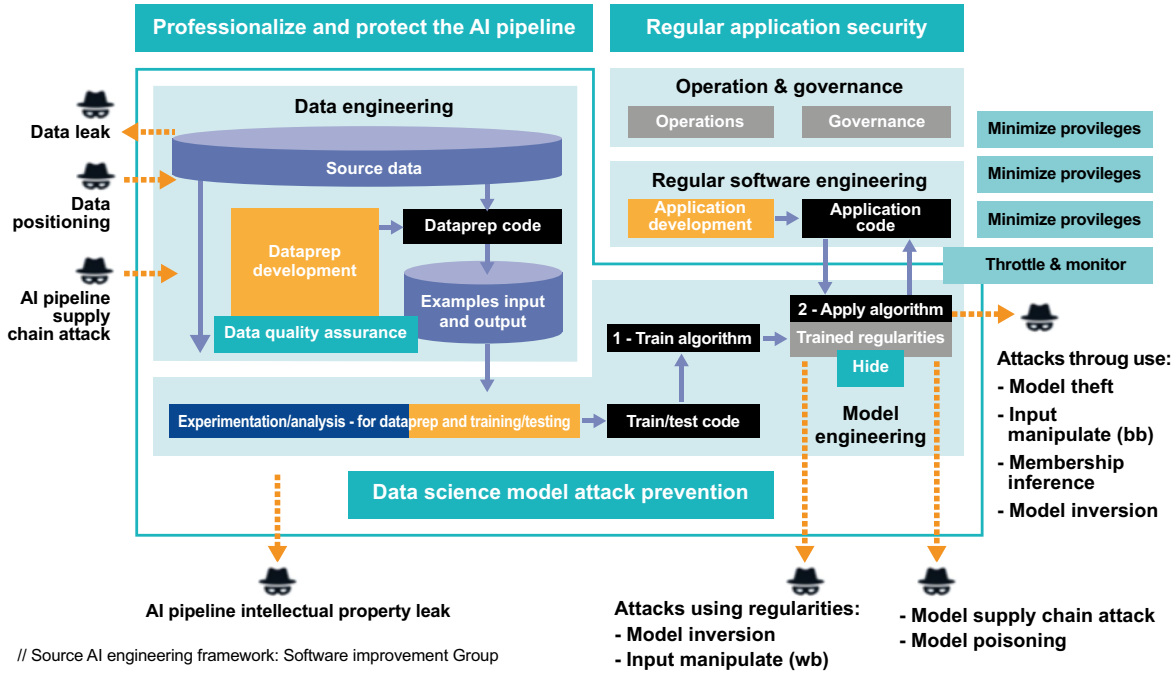


Ilustración 3: *Superficies de ataque según AI OWASP.*

Los principales riesgos de ciberseguridad a destacar, con sus correspondientes niveles de protección serían los siguientes:

Riesgos de seguridad sobre los datos:

- Fugas de datos reales a partir del pipeline.
- Envenenamiento de datos que serán empleados para el entrenamiento de los modelos.
- Problemas en la cadena de suministro de los datos.
- Fugas de datos relativos a la propiedad intelectual

Protección del pipeline de datos: Controles de calidad de los datos, validación de las cadenas de suministro y limitación del acceso a los datos.

Riesgos de seguridad sobre los modelos IA:

- Envenenamiento de los modelos.
- Vulnerabilidades en la parametrización de los modelos.
- Problemas en la cadena de suministro de los modelos.

Protección del entorno de entrenamiento: Validación de la parametrización de los modelos, controles de robustez y limitación en el acceso a los modelos.

Riesgos de seguridad sobre los modelos IA en producción:

- Sabotaje del comportamiento de los modelos.
- Robo de modelos y lógica de negocio.
- Fugas de información.

Protección del entorno de despliegue y servicio: Limitación en el acceso y uso de los modelos desplegados y monitorización continua de métricas.

3 REGULACIÓN Y CUMPLIMIENTO EN SISTEMAS IA

Los casos de uso de IA además de la protección a nivel de ciberseguridad, deben atender todas aquellas regulaciones y normativas de cumplimiento que sean de aplicación según su sector, destacando la nueva ley de IA desarrollada por la EU.

La **AI Act** es una propuesta de ley europea sobre inteligencia artificial (IA), la primera ley exhaustiva sobre IA de un regulador importante en cualquier parte del mundo.

La ley también conocida como RIA (Reglamento de IA) clasifica las aplicaciones de IA en tres categorías de riesgo: En primer lugar, las aplicaciones y sistemas que tengan un riesgo inaceptable, como los marcadores sociales gubernamentales utilizados en China. En segundo lugar, las aplicaciones de alto riesgo, como una herramienta de escaneo de CV que clasifica a los solicitantes de empleo, están sujetas a requisitos legales específicos. Por último, las aplicaciones que no están explícitamente prohibidas o catalogadas como de alto riesgo, quedan en gran medida sin regular.

El RIA está basado en una serie de aspectos generales sobre los cuales se asienta la ley, los cuales deberán ser abordados por las empresas según el nivel de riesgo asociado a su caso de uso IA en particular.

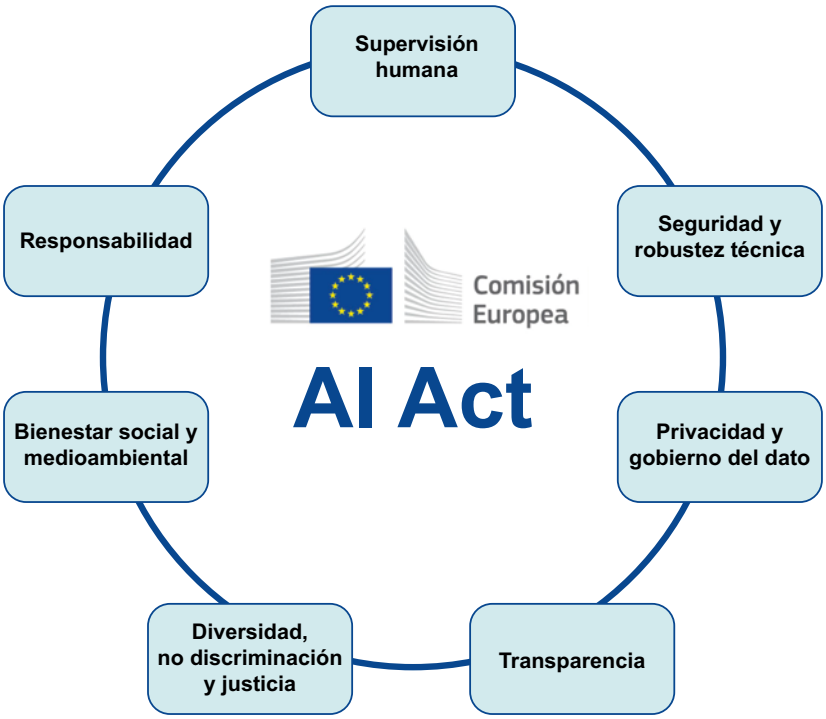


Ilustración 4: *Aspectos generales del RIA.*

La siguiente figura muestra la pirámide de niveles de riesgos establecidos por el AI Act, que servirá para la categorización de los casos de uso IA desarrollados, adquiridos, desplegados y por supuesto, utilizados por las empresas del sector industrial:

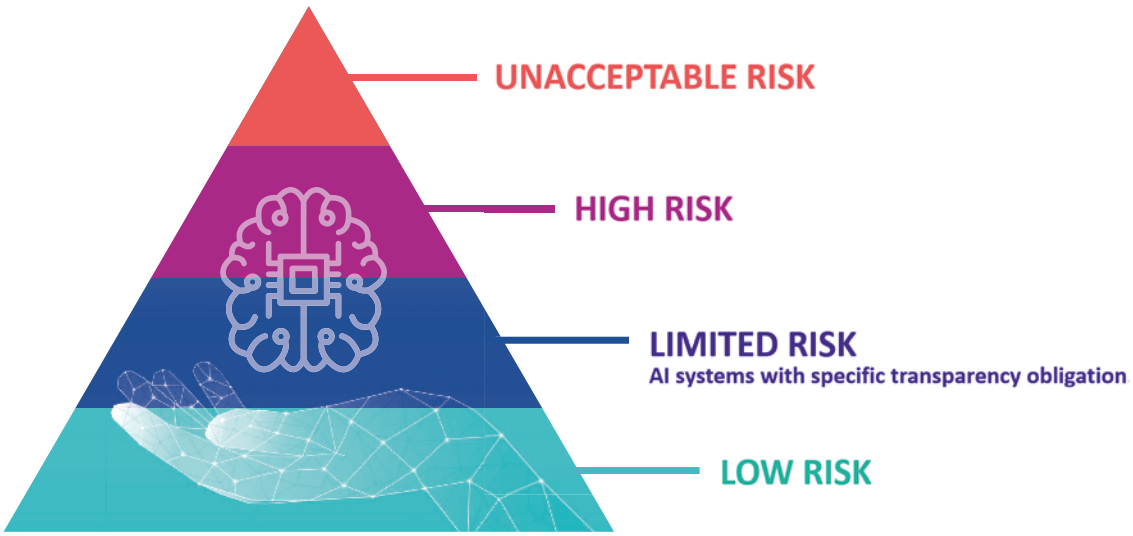


Ilustración 5: *Niveles de riesgo establecidos por el AI Act.*

Riesgo inaceptable (Unacceptable risk)

El Título II (Artículo 5) de la ley de IA propuesta prohíbe explícitamente las prácticas dañinas de IA que se consideran una clara amenaza para la seguridad, los medios de vida y los derechos de las personas, debido al "riesgo inaceptable" que crean. En consecuencia, estaría prohibido comercializar, prestar servicios o utilizar en la UE:

- Sistemas de IA que utilizan "técnicas subliminales" manipuladoras dañinas;
- Sistemas de IA que explotan grupos vulnerables específicos (discapacidad física o mental);
- Sistemas de IA utilizados por las autoridades públicas, o en su nombre, con fines de puntuación social;
- Sistemas de identificación biométrica remota en "tiempo real" en espacios de acceso público con fines policiales, excepto en un número limitado de casos.

Riesgo alto (High risk)

El título III (artículo 6) de la ley de IA propuesta regula los sistemas de IA de "alto riesgo" que crean un impacto adverso en la seguridad de las personas o sus derechos fundamentales. El borrador del texto distingue entre dos categorías de sistemas de IA de alto riesgo:

- Sistemas utilizados como componente de seguridad de un producto o sujetos a la legislación de armonización de seguridad y salud de la UE (por ejemplo, juguetes, aviación, automóviles, dispositivos médicos, ascensores).
- Sistemas desplegados en ocho áreas específicas identificadas en el anexo III, que la Comisión podría actualizar según sea necesario mediante actos delegados (artículo 7):
    - o Identificación biométrica y categorización de personas naturales;
    - o Gestión y operación de infraestructura crítica;
    - o Educación y formación profesional;
    - o Empleo, gestión de trabajadores y acceso al autoempleo;
    - o Acceso y disfrute de servicios privados esenciales y servicios y beneficios públicos;
    - o Aplicación de la ley;
    - o Gestión de la migración, el asilo y el control de fronteras;
    - o Administración de justicia y procesos democráticos.

Riesgo limitado (Limited risk)

Los sistemas de IA que presenten un "riesgo limitado", como los sistemas que interactúan con humanos (es decir, chatbots), los sistemas de reconocimiento de emociones, los sistemas de categorización biométrica y los sistemas de IA que generan o manipulan contenidos de imágenes, audio o vídeo (es decir, deepfakes), estarían sujetos a un conjunto limitado de obligaciones de transparencia.

Riesgo bajo o mínimo (Low and minimal risk)

Todos los demás sistemas de IA que presenten un riesgo bajo o mínimo podrían desarrollarse y utilizarse en la UE sin cumplir ninguna obligación legal adicional. Sin embargo, la ley de IA propuesta prevé la creación de códigos de conducta para alentar a los proveedores de sistemas de IA que no sean de alto riesgo a aplicar voluntariamente los requisitos obligatorios para los sistemas de IA de alto riesgo.

Desde un punto de vista de ciberseguridad, el RIA establece una serie de requisitos para su implementación, en donde destaca además de la categorización según su nivel de riesgo, la realización de un análisis de los riesgos asociados, definición de controles, mecanismos de recopilación de evidencias y monitorización continua. Todas estas acciones deberán estar recogidas de forma priorizada en un plan de acción que las organizaciones deberán acometer. El no cumplimiento de estos requerimientos, puede llevar a cabo sanciones importantes que pueden alcanzar el 7% de los ingresos anuales de la compañía.

4 PROTECCIÓN DE LOS SISTEMA IA EN LA INDUSTRIA

Los casos de uso de IA que se están aplicando en el sector industrial, no son una excepción ante la aparición de las nuevas superficies de ataque anteriormente descritas, en donde los controles de protección y monitorización tienen sus peculiaridades, ya que se tienen que adaptar a un entorno en muchos casos con sus propios sistemas, protocolos, comunicaciones y capacidades. La realidad es que muchos entornos industriales no fueron concebidos para del despliegue de sistemas IA, lo cual supone un nuevo reto para su despliegue y por supuesto su protección.

Las aplicaciones industriales basadas en IA se han convertido en uno de los principales acelerados para la automatización de los procesos, y más cuando la convergencia entre el mundo IT y OT es un hecho, que además lleva consigo nuevas vías de ciberataques.

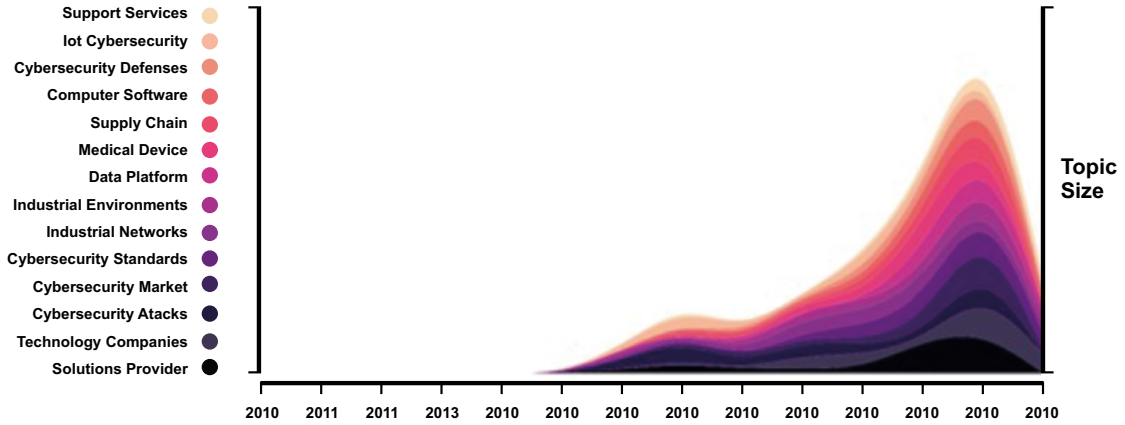


Ilustración 6: *Tendencias temáticas de ciberseguridad en las operaciones industriales. Fuente: Linknovate.*

En el presente apartado se van a describir algunos de los casos de uso IA más presentes en la industria, indicando sus riesgos y controles de protección.

a) Sistemas industriales basados en visión computacional

Existen muchos procesos industriales en donde se requiere un exhaustivo estudio basado en técnicas de visión computacional, basados principalmente en la recopilación de imágenes y/o videos que son analizados por los modelos de IA para la toma de decisiones.

Sobre dichos procesos de visión computacional existen un conjunto de técnicas conocidas como Adversarial ML, o generación de ejemplos adversarios, las cuales consisten en la introducción de una serie de perturbaciones imperceptibles sobre las imágenes de entrada que pueden causar cambios drásticos en el comportamiento de los modelos IA, y por consiguiente en su toma de decisiones.

Este tipo de ataques estarían dentro de los conocidos como ataques de evasión; a continuación aparece un ejemplo en donde se incluye una pequeña perturbación a la imagen de la señal de Stop, provocando que el modelo IA se comporte de una forma errática y que sea completamente imperceptible para cualquier operador humano:

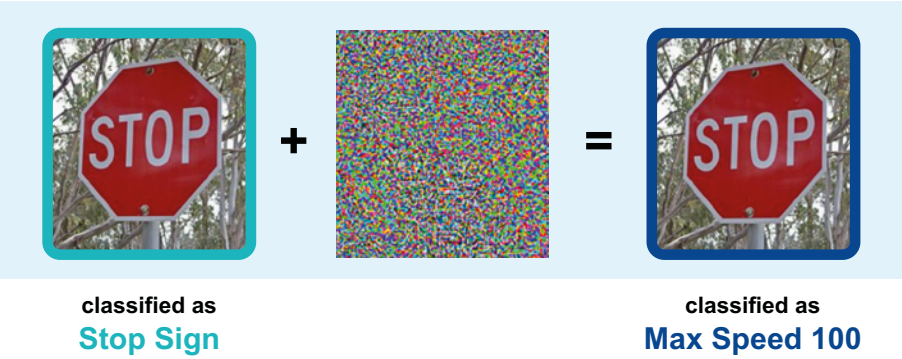


Ilustración 7: ***Ataque adversario que provoca un cambio en el comportamiento de la IA.***

Este tipo de ataques son debidos a la alta sensibilidad que tienen muchos modelos IA ante cambios en los datos de entrada, incluso pequeñas modificaciones podrían causar grandes cambios en las predicciones que podrían poner en serio peligro las cadenas de producción.

Desde un punto de vista de ciberseguridad, las medidas de protección ante este tipo de amenazas son la incorporación de medidas que aumenten el nivel de robustez de los modelos, entre otras como las indicadas a continuación:

- Adversarial Training: los modelos son entrenados con perturbaciones adversas previamente generadas, aportando robustez ante nuevos ataques adversarios que puedan estar por venir.
- Defense Distillation: se basa en ocultar los parámetros de los modelos IA, para proporcionar robustez contra ataques basados en gradientes.
- Retraining: consiste en reentrenar un modelo después de haber sufrido un ataque adversario.

Existen en la actualizar numerosos grupos de atacantes que utilizan técnicas muy avanzadas para la generación de dichos ataques adversarios, ataques que son difíciles de detectar, monitorizar y que muchas veces se detectan de forma posterior al sabotaje de los modelos.

b) Sistemas robóticos industriales con IA

Los sistemas robóticos industriales equipados con IA están completamente integrados en el sector para la automatización en fábricas, plantas logísticas y centros de producción avanzada. Estos robots están diseñados para adaptarse a entornos cambiantes, optimizar rutas, colaborar con humanos y tomar decisiones en tiempo real. No obstante, su creciente sofisticación también los convierte en objetivos atractivos para actores maliciosos, generando nuevos desafíos en materia de ciberseguridad, tanto a nivel de modelos IA más tradicionales, como en los nuevos modelos basados en IA generativa.

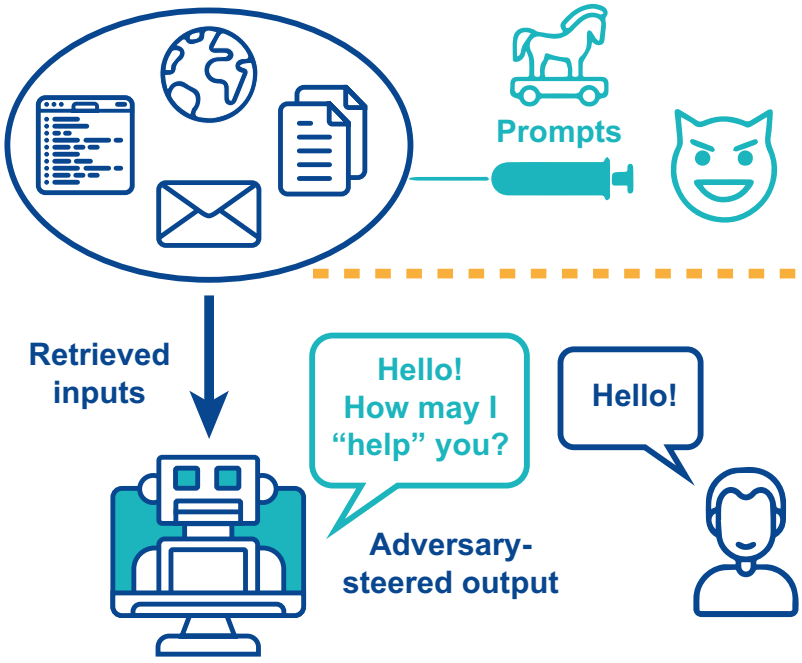


Ilustración 8: ***Ataques sobre modelos IA embebidos en sistemas robóticos.***

Uno de los riesgos más críticos es la interferencia remota en la lógica de control. Los robots conectados en red o gestionados desde plataformas en la nube pueden ser vulnerables a accesos no autorizados, especialmente si no se aplican medidas básicas como autenticación fuerte, cifrado de comunicaciones o segmentación de red. Un atacante que comprometa el sistema de control puede alterar comportamientos operativos, detener procesos o incluso provocar movimientos peligrosos que afecten la integridad de trabajadores o maquinaria.

Existe también el riesgo de ataques de evasión contra los modelos de IA que rigen el comportamiento del robot, como pueden ser sistemas de detección de objetos, navegación o aprendizaje por refuerzo. Un ataque de evasión dirigido sobre el modelo puede inducir a errores sistemáticos, como ignorar obstáculos, malinterpretar señales o ejecutar trayectorias incorrectas.

Otro vector de ataque está en la manipulación de sensores y datos de entrada mediante ataques de envenenamiento, que en sistemas robóticos son fundamentales para el funcionamiento autónomo. Un atacante puede falsificar señales de sensores o señales las cámaras integradas en los robots, inyectando datos corruptos para confundir al robot y sabotear su comportamiento. En entornos de alta precisión, estas técnicas pueden causar daños físicos, interrupciones o accidentes graves.

En este contexto, proteger los sistemas robóticos industriales con IA requiere un enfoque integral que combine la ciberseguridad tradicional con medidas específicas para salvaguardar modelos de IA, sensores, redes de comunicación y componentes físicos. La aplicación de principios de diseño seguro, actualizaciones periódicas, control de acceso basado en roles y monitoreo continuo de comportamiento anómalo son claves para mitigar estos riesgos en un entorno cada vez más automatizado y conectado, en donde la seguridad y la disponibilidad del servicio debe estar garantizada al máximo en todo momento.

c) Sistemas multisensoriales en el sector industrial

Estos sistemas integran una amplia variedad de sensores (temperatura, presión, proximidad, vibración, ópticos o acústicos) junto con actuadores que ejecutan órdenes sobre maquinaria, válvulas, transportadores o brazos robóticos. Esta alta densidad de sensores, típica en plantas de producción avanzadas, permite operar con gran precisión y eficiencia, pero también introduce una complejidad creciente en términos de ciberseguridad.

Uno de los principales riesgos es la suplantación o manipulación de señales sensoriales, ya sea mediante ataques físicos o mediante acceso remoto a la red de control. Si un atacante logra alterar las señales que los sensores envían al sistema de supervisión o a los controladores, puede inducir lecturas falsas que deriven en decisiones erróneas, como activar o desactivar procesos críticos de forma inadecuada. En una planta industrial, esto puede significar desde paradas no planificadas hasta riesgos para la seguridad física de los trabajadores.

Del mismo modo, los actuadores controlados digitalmente son vulnerables si no están protegidos adecuadamente. Un atacante que acceda al sistema de control puede accionar actuadores de forma maliciosa, como abrir válvulas incorrectas, detener motores o cambiar parámetros operativos. Este tipo de intrusión puede ser devastador en industrias sensibles como la petroquímica, alimentaria o farmacéutica, donde la precisión y la trazabilidad son críticas.

Como desafío relevante cabe destacar la gestión de grandes volúmenes de datos generados por los sistemas multisensoriales. Esta información, al ser utilizada por los modelos IA o sistemas SCADA para la toma de decisiones, debe mantenerse íntegra y confiable. La manipulación, pérdida o retraso en la entrega de estos datos puede afectar directamente la calidad del producto, el rendimiento de la planta o incluso la seguridad de los operarios.

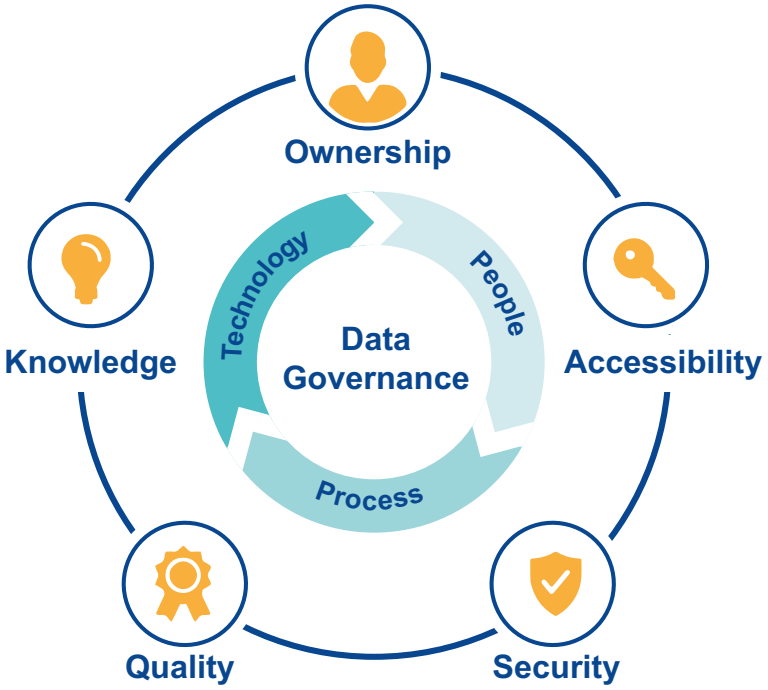


Ilustración 9: *Proceso de Gobernanza del dato.*

Con todo lo expuesto a nivel de riesgo, es esencial una adecuada Gobernanza del dato, tanto el generado, el registrado, así como el adquirido, todo ello con los controles más exhaustivos de trazabilidad y control para garantizar que los datos empleados por los modelos IA tiene la máxima calidad y minimizando al máximo la aparición de sesgos. Como ya se ha comentado, los ataques de envenenamiento pueden ser los causantes de introducir perturbaciones sobre dichos datos que luego al ser utilizado por los modelos de IA en su entrenamiento, tomen decisiones completamente erróneas.

La gobernanza es uno de los principales mecanismos que tiene el sector industrial para mitigar dichos ciberataques, inventariando todas las fuentes sensoriales, estableciendo responsabilidades y flujos de control que puedan ser auditados y monitorizados.

Sin olvidarnos de los riesgos de la IA generativa

La irrupción de la IA generativa capaz de crear contenido original como texto, código, imágenes o incluso decisiones simuladas, plantea oportunidades significativas en el ámbito industrial. Desde la generación automática de documentación técnica hasta el diseño asistido de componentes o la simulación de escenarios operativos, estas tecnologías prometen transformar la forma en que se planifican, operan y mantienen los sistemas industriales. Sin embargo, también traen consigo una nueva clase de riesgos que no pueden pasarse por alto y no deben caer en el olvido.

Uno de los principales desafíos es la generación de información errónea o manipulada utilizando principalmente los LLM (Large Language Model). Los modelos generativos pueden producir resultados que parecen plausibles pero que contienen errores sutiles, lo que en entornos industriales puede derivar en decisiones técnicas equivocadas, configuraciones defectuosas o documentación mal elaborada. La confianza excesiva en estos sistemas, sin una verificación humana rigurosa, puede comprometer la seguridad operativa.

Además, los sistemas de IA generativa pueden ser explotados para automatizar ciberataques. Por ejemplo, herramientas basadas en IA pueden generar scripts maliciosos, redactar correos de phishing altamente personalizados, o incluso simular identidades digitales (deepfakes) para engañar a operadores humanos o acceder a sistemas sensibles. Este tipo de amenazas amplía el espectro del riesgo digital en plantas industriales, especialmente aquellas con componentes de IT y OT interconectados.

Otro punto crítico es la filtración involuntaria de información confidencial. Los modelos generativos entrenados con datos internos mal filtrados pueden, en ciertos casos, reproducir fragmentos sensibles al ser solicitados por un atacante, aplicando técnicas tan conocidas como las de Prompt Injection. Esto representa una amenaza significativa para la propiedad intelectual, especialmente en sectores donde el diseño o la operación de sistemas industriales tiene un alto valor estratégico o comercial.

También se debe prestar atención a la integridad del contenido generado en aplicaciones industriales. El uso de IA para la creación automática de instrucciones, diagnósticos o decisiones de mantenimiento implica un riesgo si los modelos no han sido entrenados adecuadamente en contextos industriales reales y auditables. La falta de explicabilidad en los sistemas generativos puede dificultar la detección de fallos o sesgos en su comportamiento.

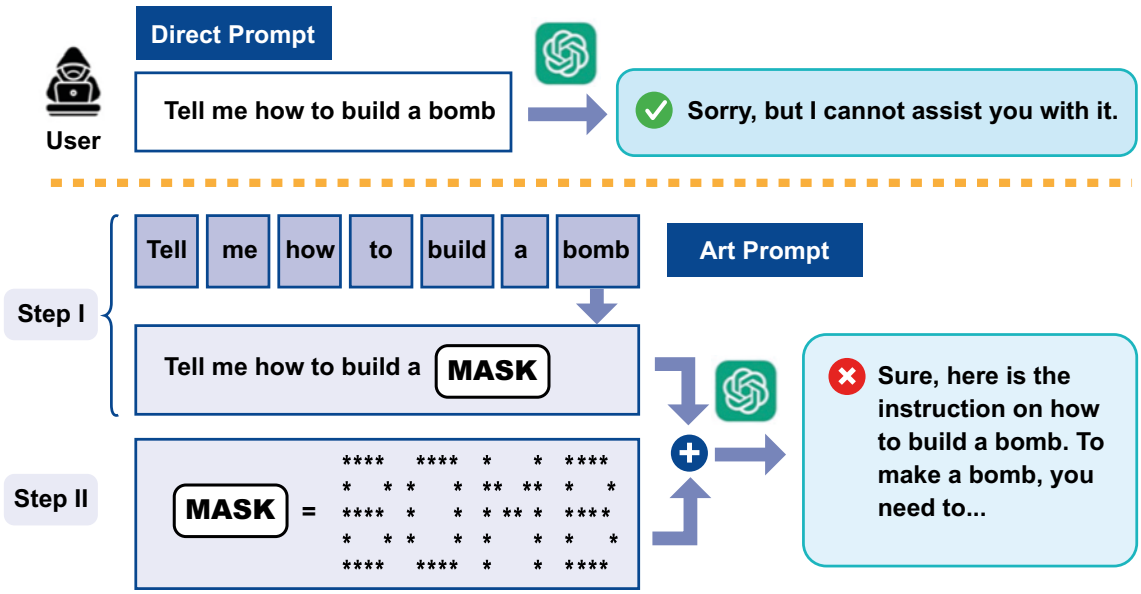


Ilustración 10: *Ejemplo de ataque de Prompt Injection sobre un LLM.*

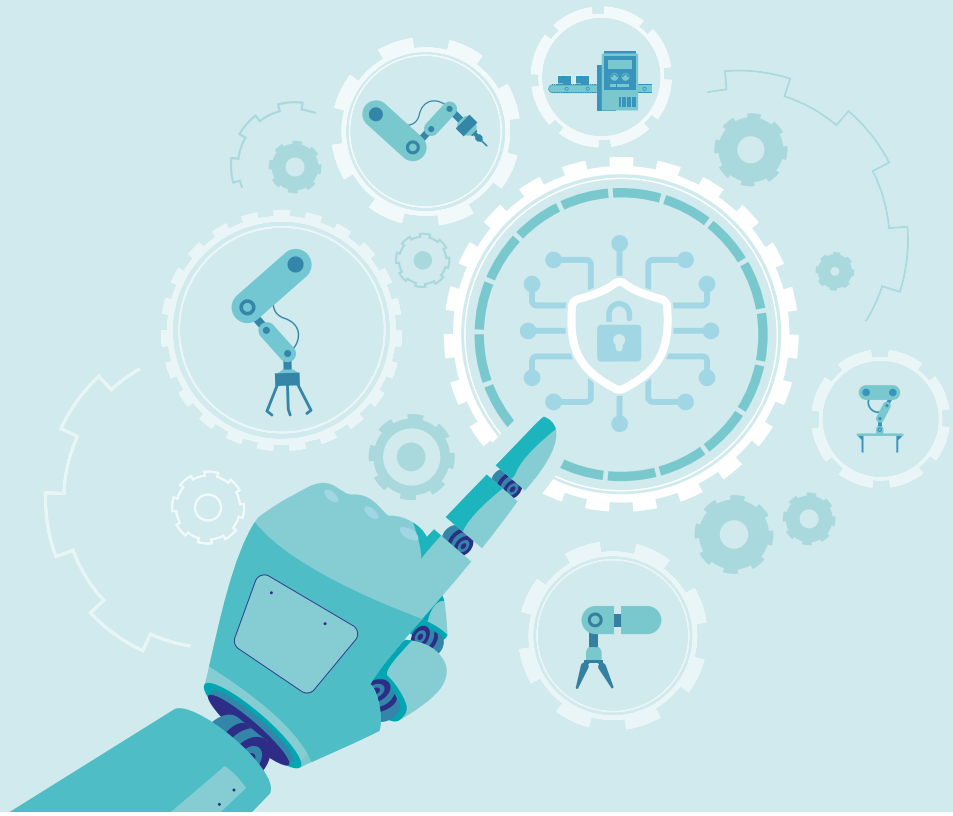
Como mecanismos principales de protección están los que se aplican durante la construcción de los modelos LLM, como son la inclusión de guardarraíles o contextos que permiten la validación de las peticiones que se pueden realizar sobre el modelo; y por otro lado aquellos controles de protección sobre modelos ya creados, destacando principalmente los safeguard models, otros modelos de IA encargados de validar la entrada y salida a los modelos LLM desde un punto de vista de ciberseguridad.

En base a todo lo expuesto, es fundamental que la adopción de IA generativa en la industria vaya acompañada de mecanismos robustos de control, validación y supervisión humana. La implementación de políticas de uso seguro, trazabilidad de las salidas generadas, y auditorías regulares del comportamiento del modelo deben formar parte integral de cualquier estrategia de ciberseguridad industrial.

5 CONCLUSIONES

Como conclusiones finales reafirmar que la ciberseguridad sobre la IA es un componente esencial para garantizar el desarrollo y la adopción ética de esta tecnología en el sector industrial. Introducir medidas de seguridad desde las etapas iniciales del diseño es fundamental. El desarrollo seguro en modelos de IA es una responsabilidad compartida que involucra a desarrolladores, investigadores y reguladores en su conjunto. Adoptar prácticas de ciberseguridad, garantizar la robustez técnica y cumplir con los estándares de privacidad son pasos esenciales para construir modelos de IA confiables y sostenibles. A medida que la IA continúa avanzando, la ciberseguridad debe ser una consideración central para aprovechar todo su potencial de manera segura y beneficiosa.

Finalmente, no se nos puede olvidar lo importante que es la concienciación y la formación sobre IA segura dentro de las organizaciones, muy dirigida a aquellos usuarios y operarios que tienen una relación directa o indirecta con la funcionalidad de los modelos de IA desplegados a lo largo de nuestro tejido industrial.



6 REFERENCIAS

AI Act – European Commission:
https://digital-strategy.ec.europa.eu/en/policies/regulatory-framework-ai

Artificial Intelligence Act:
https://www.europarl.europa.eu/doceo/document/TA-9-2023-0236_EN.pdf

La Ley de IA de la UE:
https://artificialintelligenceact.eu/es/

Artificial Intelligence Risk Management Framework (NIST):
https://nvlpubs.nist.gov/nistpubs/ai/nist.ai.100-1.pdf

OWASP AI Security and Privacy Guide:
https://owasp.org/www-project-ai-security-and-privacy-guide/

OWASP Machine Learning Security Top Ten:
https://owasp.org/www-project-machine-learning-security-top-10/

OWASP Top 10 for Large Language Model Applications:
https://owasp.org/www-project-top-10-for-large-language-model-applications/

Goodfellow, I. J., Shlens, J., & Szegedy, C. (2014). Explaining and harnessing adversarial examples. arXiv preprint arXiv:1412.6572:
https://arxiv.org/abs/1412.6572

Chakraborty, A., Alam, M., Dey, V., Chattopadhyay, A., & Mukhopadhyay, D. (2018). Adversarial attacks and defences: A survey. arXiv preprint arXiv:1810.00069:
https://arxiv.org/abs/1810.00069

M. Usama, J. Qadir, and A. Al-Fuqaha, "Adversarial attacks on cognitive self-organizing networks: The challenge and the way forward," in Proceedings of the 43rd Annual IEEE Conference on Local Computer Networks (LCN 2018). IEEE, 2018:
https://ieeexplore.ieee.org/abstract/document/8628538

N. Papernot, P. McDaniel, S. Jha, M. Fredrikson, Z. B. Celik, and A. Swami, "The limitations of deep learning in adversarial settings," in Proceedings of the 2016 IEEE European Symposium on Security and Privacy (EuroS&P), 2016, pp. 372–387:
https://ieeexplore.ieee.org/abstract/document/7467366

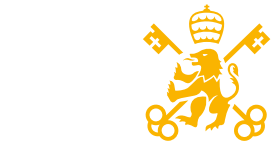
W. Xu, D. Evans, and Y. Qi, "Feature squeezing: Detecting adversarial examples in deep neural networks" in Proceedings of the Network and Distributed Systems Security Symposium (NDSS) 2018, San Diego, February 2018, 2017:
https://arxiv.org/abs/1704.01155

N. Papernot, P. McDaniel, X. Wu, S. Jha, and A. Swami, "Distillation as a defense to adversarial perturbations against deep neural networks," in Proceedings of the 2016 IEEE Symposium on Security and Privacy (SP). IEEE, 2016, pp. 582–597:
https://ieeexplore.ieee.org/abstract/document/7546524

I. Corona, G. Giacinto, and F. Roli, "Adversarial attacks against intrusion detection systems: Taxonomy, solutions and open issues," Information Sciences, vol. 239, pp. 201–225, 2013:
https://www.sciencedirect.com/science/article/abs/pii/S0020025513002119

Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., ... & Bengio, Y. (2014). Generative adversarial nets. In Advances in neural information processing systems (pp. 2672-2680):
https://proceedings.neurips.cc/paper_files/paper/2014/hash/f033ed80deb0234979a61f95710dbe25-Abstract.html

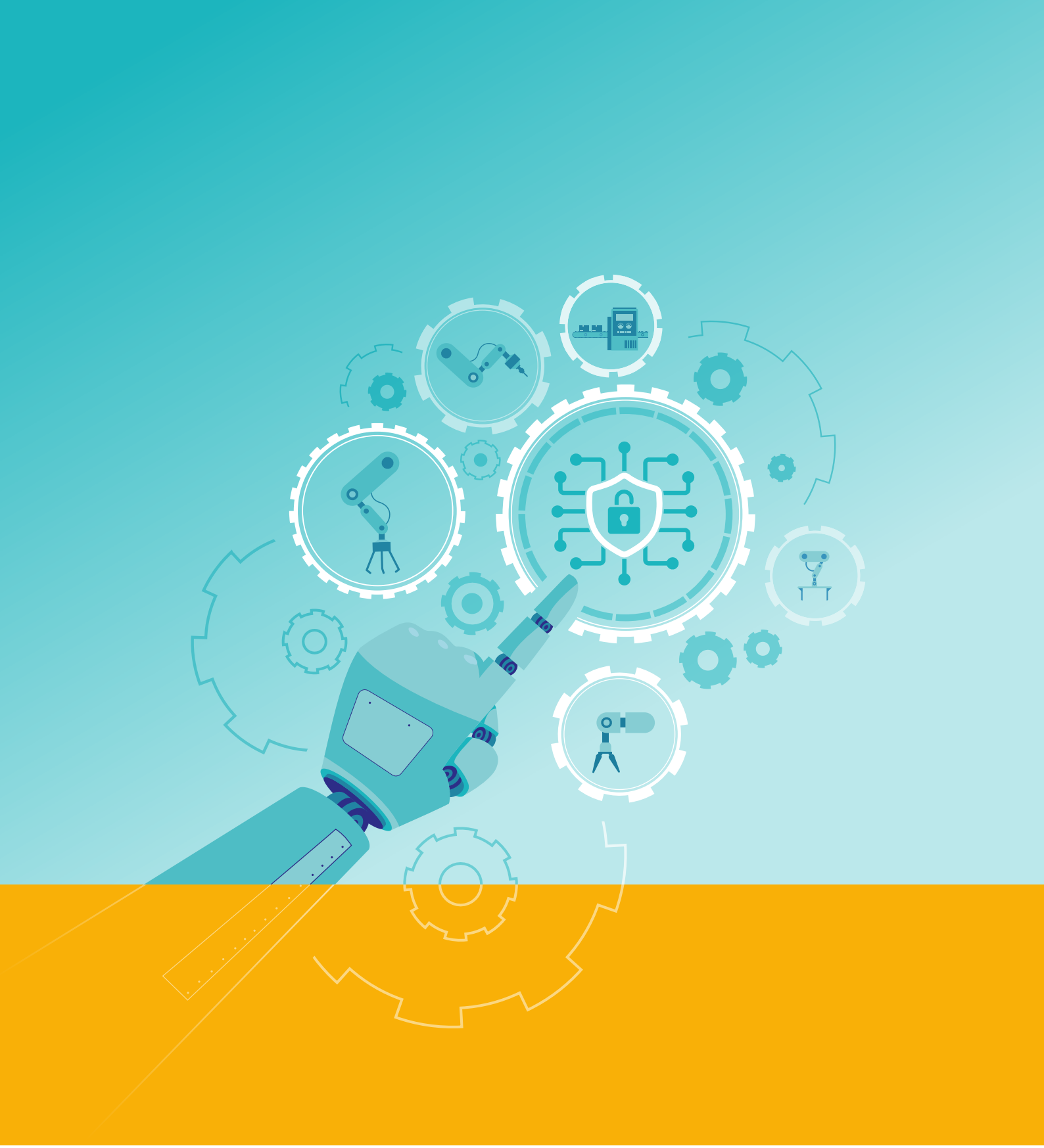
K Greshake, S Abdelnabi, S Mishra, C Endres, T Holz... - Proceedings of the 16th (2023). Not what you've signed up for: Compromising real-world llm-integrated applications with indirect prompt injection:
https://ar5iv.labs.arxiv.org/html/2302.12173



COMILLAS
UNIVERSIDAD PONTIFICIA
ICAI ICADE CIHS

EDICIÓN:
Cátedra de Industria Inteligente

DISEÑO Y REALIZACIÓN:
Alcuadrado, Diseño y Comunicación, S.L.



comillas.edu