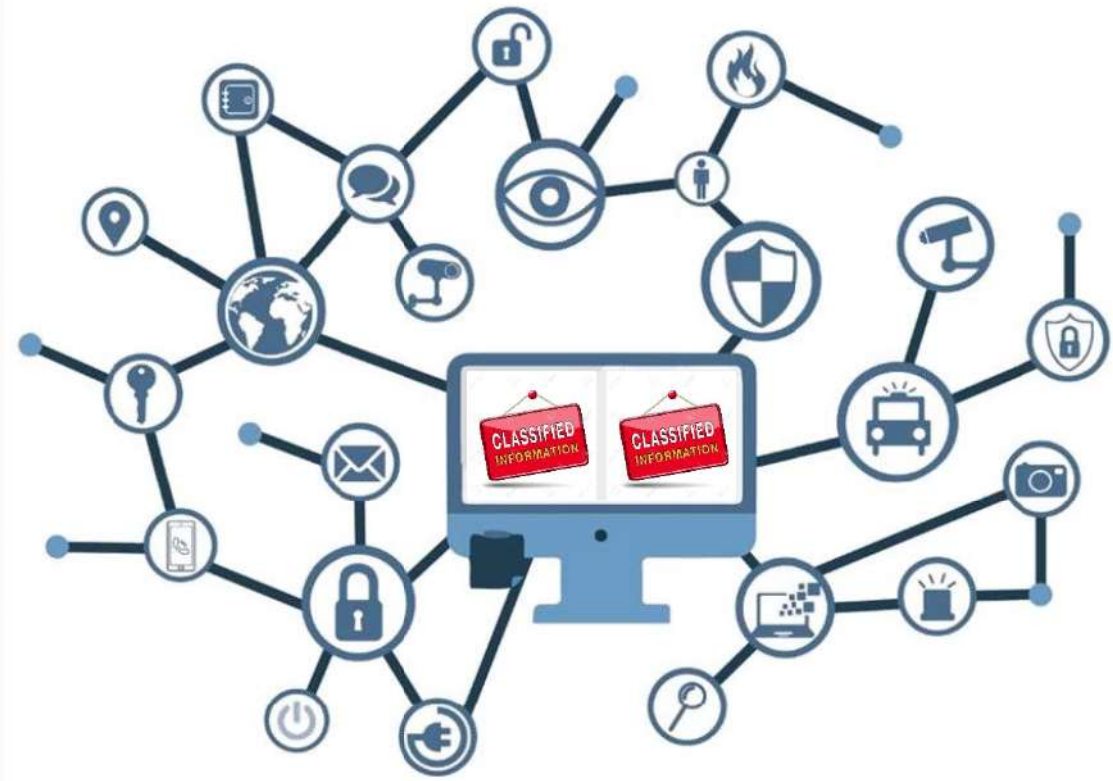
Inteligencia Artificial en Entornos Sensibles: Riesgos y Vulnerabilidades en el Manejo de Información Clasificada



Enero 2026

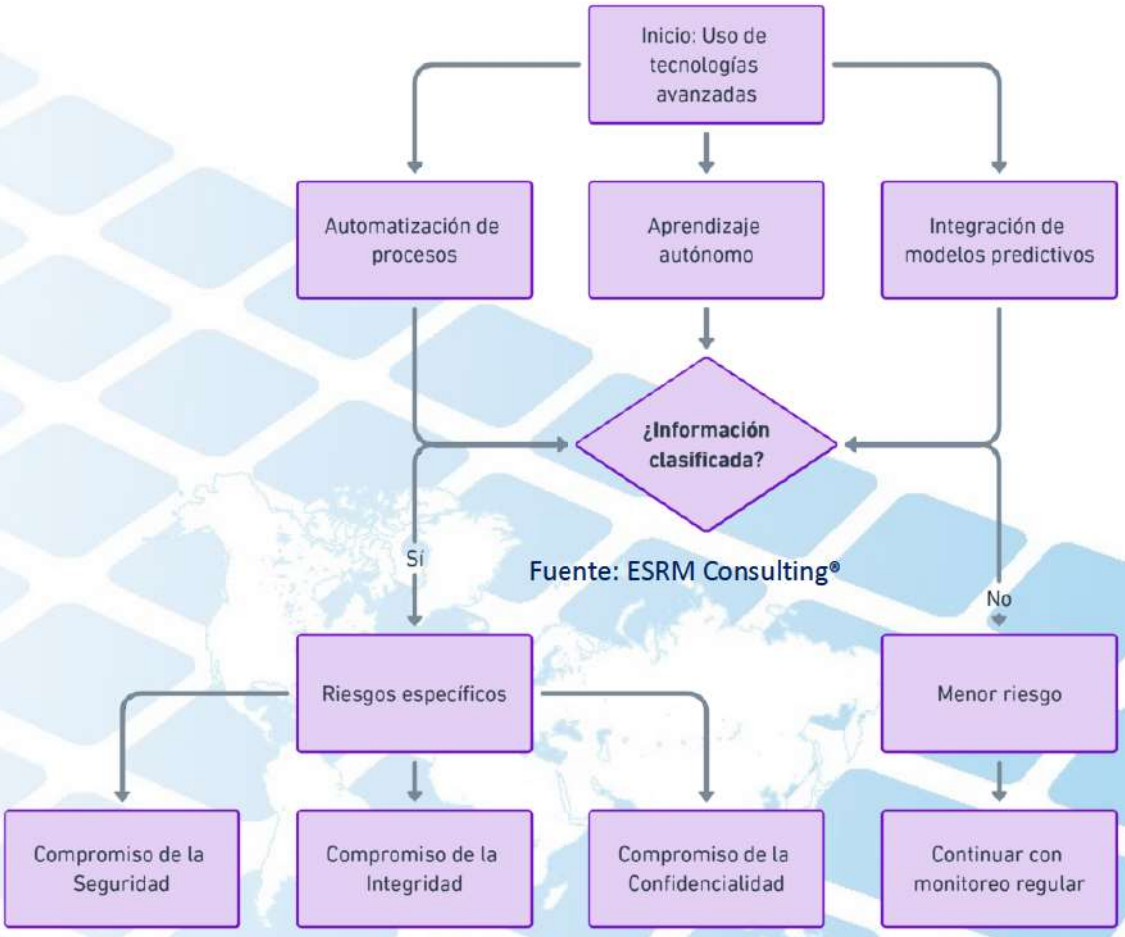
LIMITACIÓN DE RESPONSABILIDAD

El presente documento se proporciona de acuerdo con los términos en él recogidos, rechazando expresamente cualquier tipo de garantía implícita que se pueda encontrar relacionada. En ningún caso, ESRM Consulting® puede ser considerada responsable del daño directo, indirecto, fortuito o extraordinario derivado de la utilización de la información que se indica incluso cuando se advierta de tal posibilidad.

INDICE

1 Introducción

La inteligencia artificial (IA) se ha convertido en una herramienta de uso transversal, capaz de transformar procesos, automatizar tareas y optimizar la toma de decisiones en sectores tan diversos como la medicina, la industria, la defensa o los servicios públicos. Sin embargo, su aplicación en entornos donde se gestiona Información Confidencial (C) o Clasificada (IC) plantea una serie de interrogantes y **preocupaciones legítimas que no deben ser ignoradas.** La automatización, el aprendizaje autónomo y la integración de modelos predictivos en trabajos sensibles conllevan riesgos específicos que pueden comprometer la seguridad, la integridad y la confidencialidad de la información tratada.



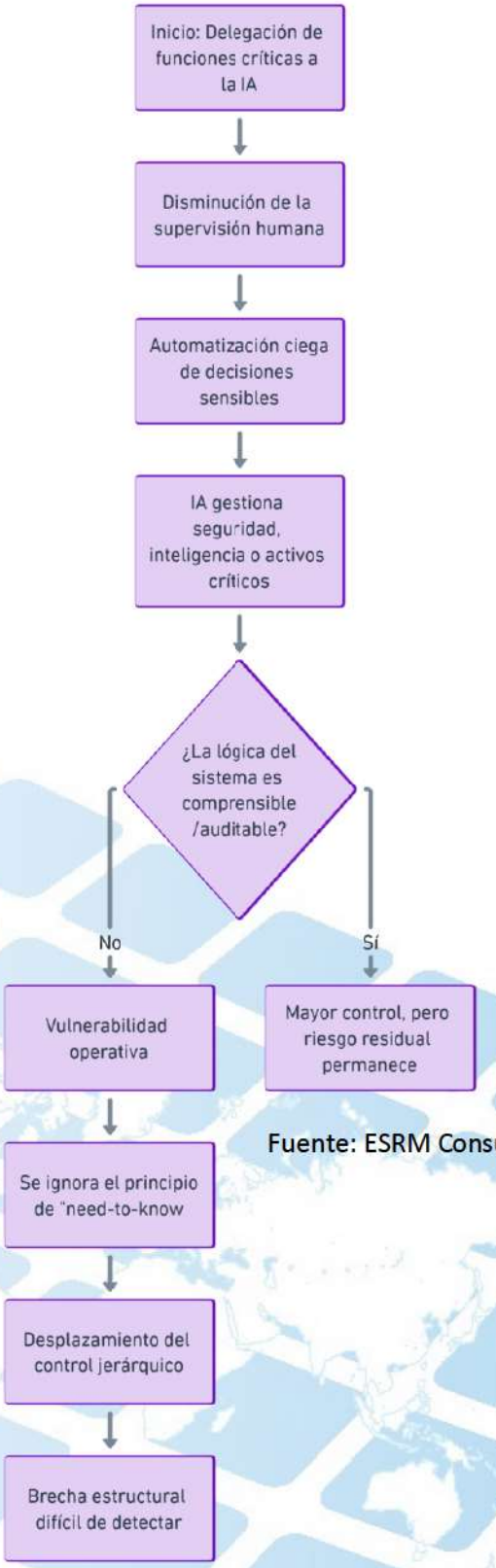
Fuente: ESRM Consulting®

2. Riesgo de exposición no autorizada de información

Uno de los principales peligros es el riesgo de fuga de datos, ya sea de forma accidental o maliciosa. Los sistemas de IA, especialmente aquellos entrenados en la nube o que interactúan con bases de datos externas, pueden generar salidas que revelen fragmentos de información sensible sin que el usuario sea plenamente consciente. **En contextos donde se manejan secretos empresariales, datos clasificados por el Estado o información crítica para la seguridad nacional, este tipo de filtración podría tener consecuencias devastadoras.** Además, muchos modelos de lenguaje o análisis predictivo no están diseñados para diferenciar entre información pública y clasificada, lo que aumenta la probabilidad de que datos protegidos acaben siendo reutilizados o integrados en contextos inapropiados.

3. Dependencia tecnológica y pérdida de control humano

La delegación de funciones críticas en sistemas de IA puede provocar una disminución progresiva de la supervisión humana. Esta tendencia a la "automatización ciega" implica que decisiones sensibles —como las relacionadas con la seguridad de instalaciones, el tratamiento de información de inteligencia o la gestión de activos críticos— pueden quedar en manos de algoritmos cuya lógica operativa no siempre es comprensible ni auditable. En entornos donde el principio de necesidad de conocimiento y el control jerárquico son pilares fundamentales, el desplazamiento del criterio humano por el cálculo estadístico introduce una vulnerabilidad estructural difícil de detectar hasta que ya se ha producido una brecha.



Fuente: ESRM Consulting®

3. Falta de trazabilidad y explicabilidad del modelo[1]

Los algoritmos de IA, especialmente los de tipo "caja negra" como los basados en redes neuronales profundas, carecen en muchas ocasiones de trazabilidad. Esto significa que, ante una decisión errónea, sesgada o incluso peligrosa, puede resultar extremadamente complejo reconstruir la cadena lógica que llevó al sistema a actuar de una determinada manera. En contextos donde se exige una rendición de cuentas estricta —como ocurre con la Información Clasificada (IC) de carácter estatal o en empresas sujetas a normativa de seguridad industrial—, la falta de explicabilidad constituye una vulnerabilidad de primer orden que puede derivar en responsabilidades penales o contractuales.

4. Riesgo de manipulación y ataques dirigidos

Los sistemas de IA son susceptibles de ser manipulados mediante técnicas específicas, como la introducción de datos contaminados (data poisoning) o ataques adversariales diseñados para alterar su comportamiento. **En escenarios donde se utilicen sistemas inteligentes para el análisis de amenazas, la evaluación de riesgos o la clasificación de documentos sensibles, la posibilidad de que un actor hostil influya en el modelo o en sus decisiones representa una amenaza real y creciente.** Además, los actores con intenciones maliciosas pueden aprovecharse del funcionamiento estadístico de los sistemas de IA para generar ataques que pasen desapercibidos para los sistemas tradicionales de defensa, accediendo así a capas protegidas de información sin levantar sospechas inmediatas.

5. Incompatibilidad con marcos normativos de seguridad

- El uso de IA en trabajos que implican el tratamiento de Información Clasificada (IC) puede entrar en conflicto con las normativas nacionales e internacionales que regulan el acceso, almacenamiento y transmisión de dicha información.
- En el ámbito europeo, la Directiva 2013/488/UE sobre protección de Información Clasificada (IC), así como las políticas de los Estados

[1] La **explicabilidad del modelo** en inteligencia artificial significa la **capacidad de entender y explicar cómo y por qué un sistema de IA llega a una decisión o resultado**

Miembros, establecen requisitos específicos sobre entornos de confianza, control de acceso y segregación física y lógica de la información.

- En España, la ANPIC (Autoridad Nacional de Protección de la Información Clasificada) establece que no está permitido conectar un ordenador o sistema de información (STIC) al exterior, salvo que esté acreditado bajo sus propias normas.
- Además, resulta obligatorio aplicar las guías y normativa del Centro Criptológico Nacional (CCN), que definen los estándares técnicos y organizativos para garantizar la seguridad de los sistemas y la protección de la Información Clasificada (IC).

La utilización de sistemas de IA que interactúan con servidores externos, proveedores no controlados o plataformas de código abierto puede contravenir estas normativas, con consecuencias legales y reputacionales severas para las organizaciones implicadas.

6. Falsa sensación de seguridad o fiabilidad

Otro riesgo menos evidente, pero igual de relevante, es el de confiar en la IA como una herramienta infalible o completamente segura. Esta falsa sensación de control o precisión puede llevar a que decisiones de alto impacto se basen en datos incorrectos, interpretaciones erróneas o supuestos no verificados. En el ámbito de la seguridad o la defensa, por ejemplo, una clasificación errónea o una correlación mal interpretada puede tener consecuencias estratégicas, operativas e incluso humanas.

Conclusión

La inteligencia artificial ofrece un potencial indiscutible, también en contextos de alta sensibilidad, pero su integración en entornos que manejan información confidencial o clasificada debe hacerse con extrema cautela. No se trata solo de valorar los beneficios operativos o los avances tecnológicos, sino de entender que el uso de IA modifica de manera estructural los modelos de seguridad, control y gobernanza.

Antes de implantar sistemas de IA en entornos clasificados, es imprescindible: Realizar una evaluación exhaustiva de riesgos.

- Definir estrictas políticas de gobernanza tecnológica.

- Establecer mecanismos de supervisión humana continuada.
- Garantizar la compatibilidad con los marcos normativos, en particular con las normas de la ANPIC y las guías del CCN.

La protección de la información sensible, al igual que la confianza en los procesos críticos, no puede ser delegada por completo a una máquina, por inteligente que sea.