CAPACITACIÓN USACH
UNIVERSIDAD DE SANTIAGO DE CHILE

# ISO 42001:
# Sistema de Gestión de Inteligencia Artificial

## Conceptos Claves, Riesgos y Controles

ALIGNMENT
SOLUCIONES ESTRATÉGICAS

DIPLOMADO
CIBERSEGURIDAD

# Agenda

- Un poco de contexto
- IA Risk
- ISO 42001 – Sistema de Gestión
- ISO 42001 - Controles

# Profesor

Carlos Lobos de Medina

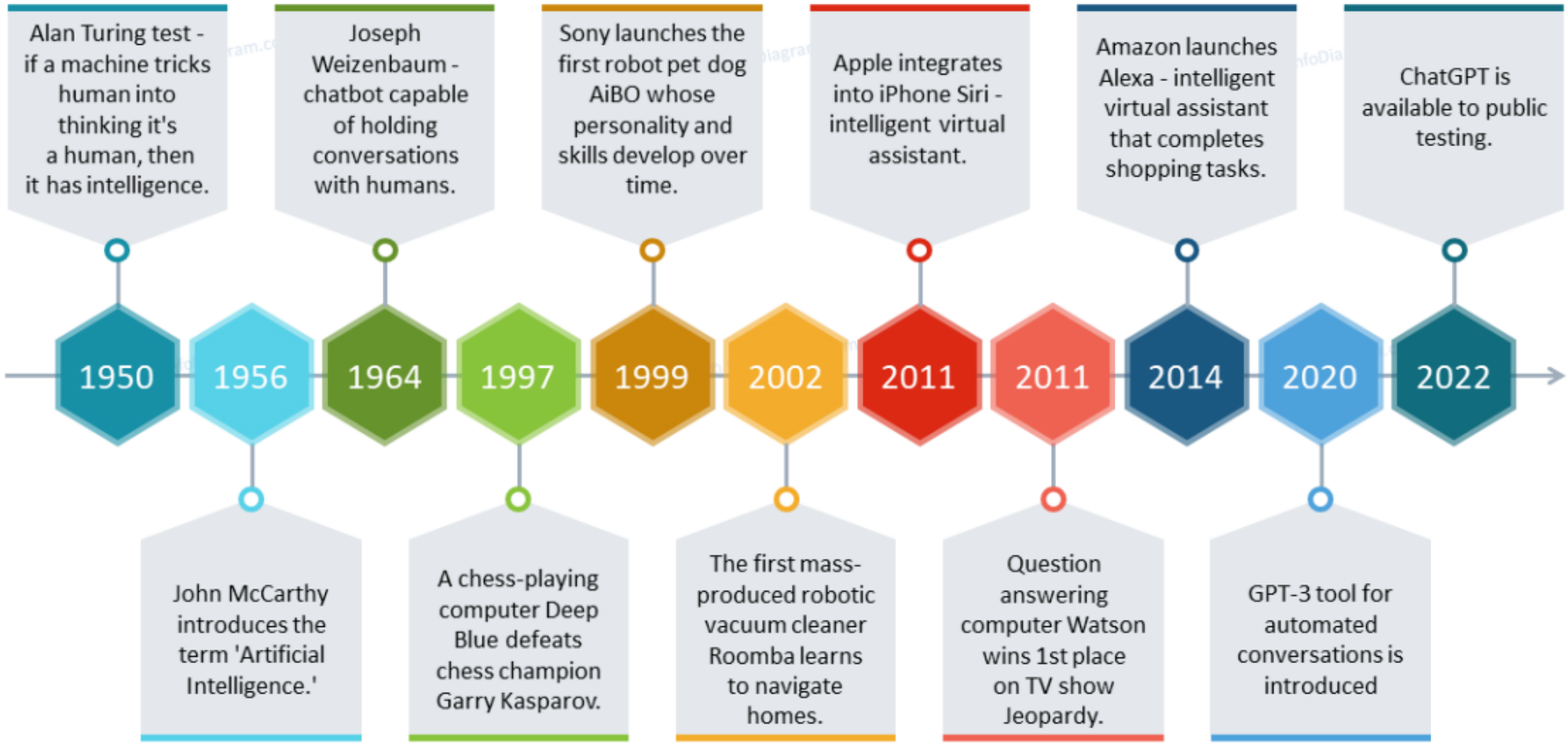
Ingeniero Civil en Informática y Magister en Informática © de la Universidad de Santiago de Chile, Diplomado en Auditoría de Sistemas, Postitulado en Seguridad Computacional de la Universidad de Chile.

Especialista en gobernanza, gestión y control de tecnologías de información empleando modelos como COBIT, ITIL, ISO 27001 e ISO 22301. Posee certificaciones internacionales CISA, CISM, COBIT, ISO Lead Auditor 27001, ISO Lead Auditor BS 25599 e ITIL V3, entre otras.}

https://www.linkedin.com/in/clobos/

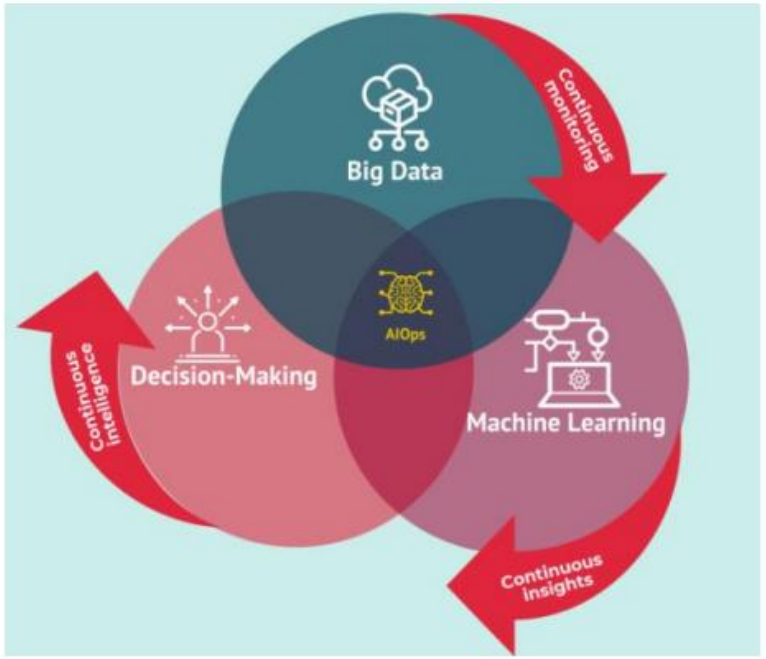
# Un poco de Contexto

# Contexto y evolución histórica

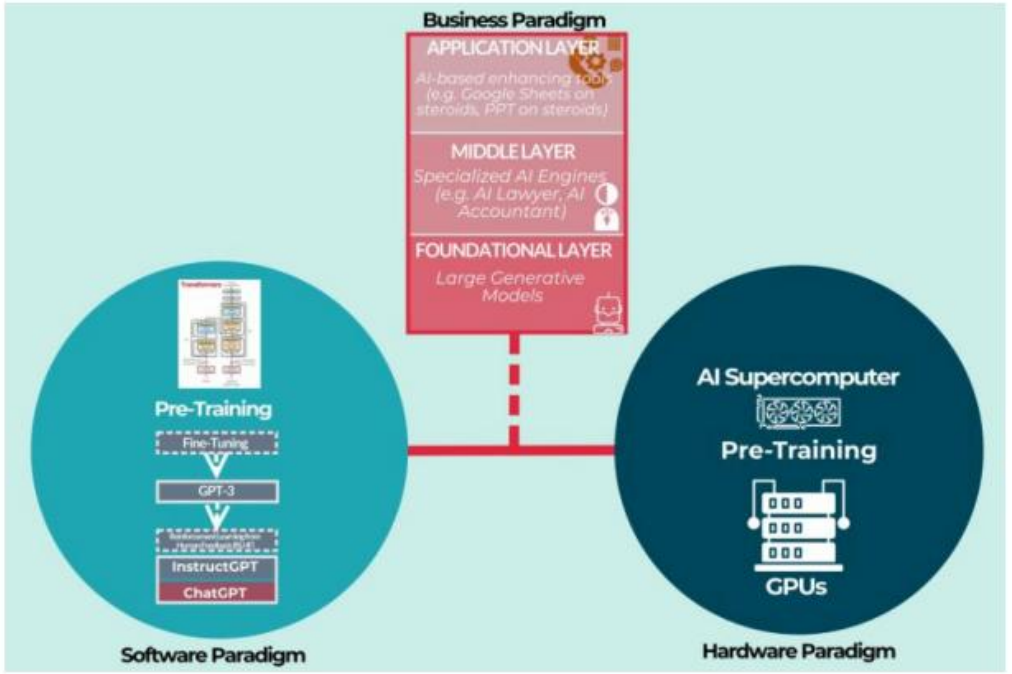


# Provisión de Servicios de IA

Visión General



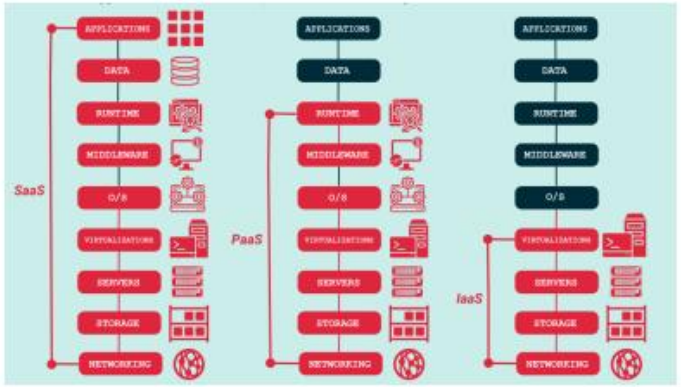

Visión más especifica



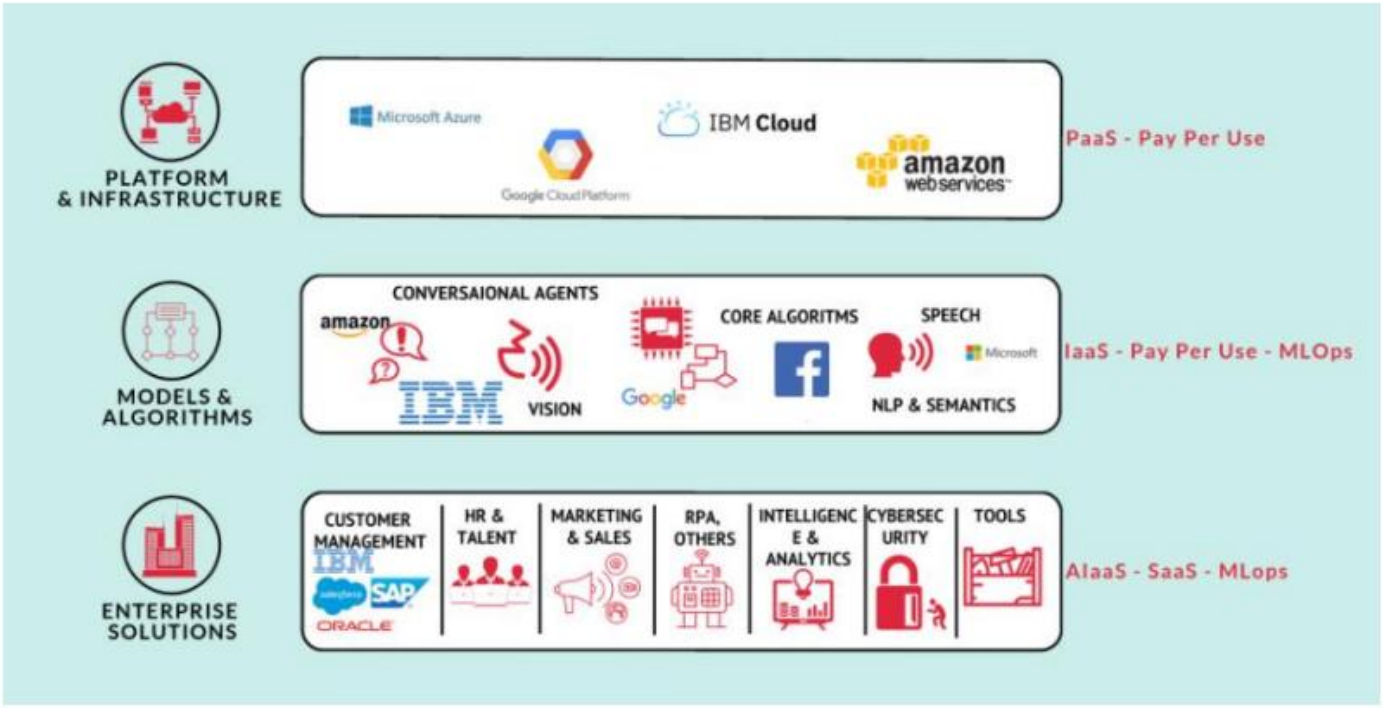

Fuente:ForWeekMba

# Servicios y usos de la IA en la Ciberseguridad

Provisión de Servicios Cloud




Provisión de Servicios de IA




Fuente:ForWeekMba

# Provisión de Servicios de IA

| | | |
|---|---|---|
| Data Protección | Security Operations | Network Security |
| EndPoint | Cloud Security | GRC |
| Application Security | IAM | Threat Intelligence |

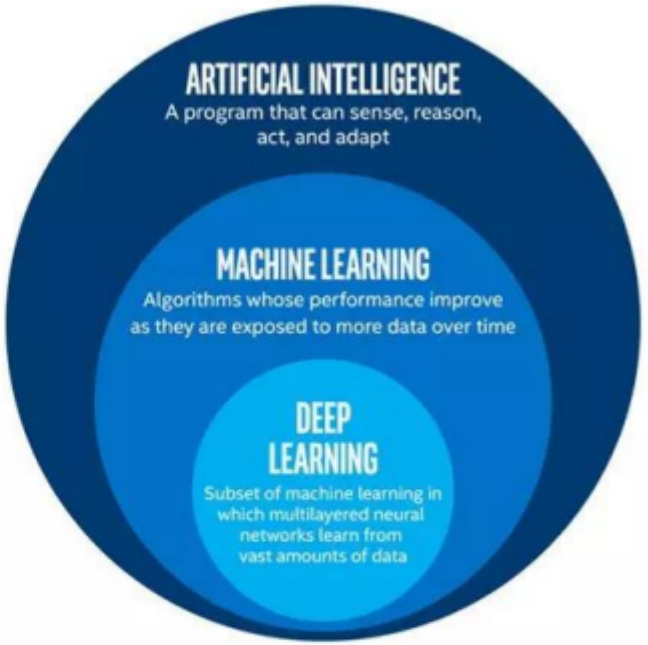

Más de un 90% de las soluciones comerciales reconoce el uso de IA

¡¡¡Prácticamente está en todo lo ocupamos!!!

Fuente: Navigating the Security Landscape - Optiv

# AI methods in cybersecurity functions

| Security function\AI | DT | SVM | NB | K-means | HMM | GAS | ANN | CNN | RNN | Encoders | SNN |
|---|---|---|---|---|---|---|---|---|---|---|---|
| Intrusion detection | X | X | X | X | X | X | X | X | X | | X |
| Malware detection | X | X | X | X | | | | X | X | | |
| Vulnerability assessment | X | | | | | | | | | | |
| Spam filtering | | | X | | | | | | | | |
| Anomaly detection | | | | | X | | | | | X | |
| Malware classification | | | | | | X | X | | | | X |
| Phishing detection | | | | | | | X | | | | |
| Traffic analysis | | | | | | | | X | X | | |
| Data compression | | | | | | | | | | X | |
| Feature extraction | | | | | | | | | | X | |




- Muchos de los usos de la IA en ciberseguridad emplean técnicas clásicas de Data Mining, Machine Learning, Deep Learning, la gran diferencia es que no hay que programarlas específicamente en cada escenario.

Fuente: Artificial Intelligence and Cybersecurity Research - ENISA

# Ejemplos en la Prevención/Detección de Ataque

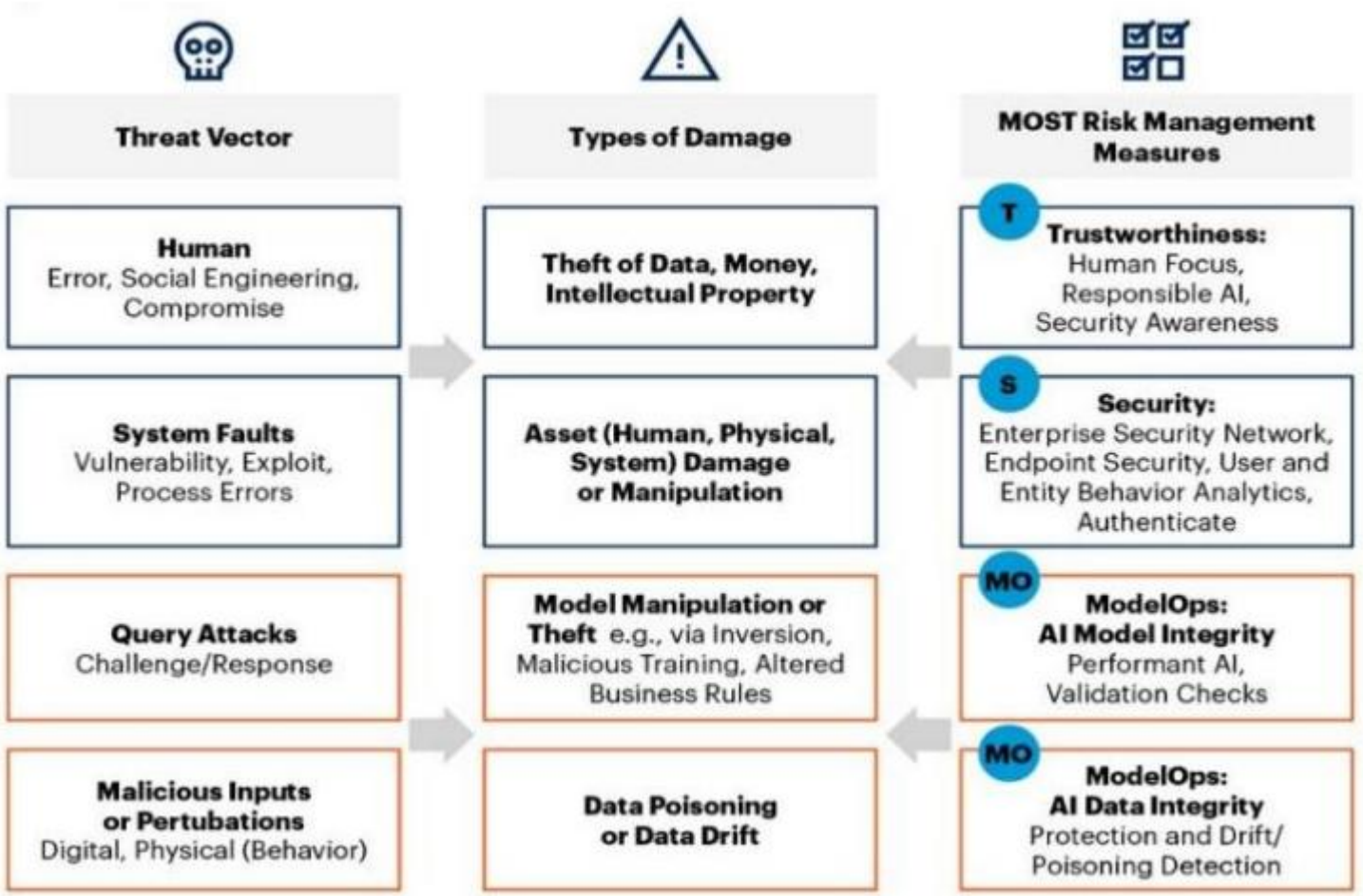
Prevención

| Task | Example of AI methods, techniques, approaches |
|---|---|
| *Fuzzers* | DL |
| *Pen-testing* | Reinforcement learning |
| *Vulnerability assessment* | NLP, traditional ML |

Detección

| Task | Examples of AI techniques |
|---|---|
| *Spam detection* | SVM, DT |
| *Intrusion detection* | Supervised and unsupervised approaches, bio-inspired algorithms |
| *Malware detection* | Standard ML classifiers, DL |
| *Attack detection* | SVM, DT |

# Riesgos Generales de la IA: Gartner



# Riesgos Generales de la IA

| MONITORIZACIÓN CONTINUA DE RIESGOS | SUPERVISIÓN DE LA ARQUITECTURA DE DATOS Y TI | REVISIÓN DE LOS MODELOS DE IA | SUPERVISIÓN DE LA IMPLEMENTACIÓN | MONITORIZACIÓN TRAS LA PUESTA EN PRODUCCIÓN |
|---|---|---|---|---|
| Identificación de riesgos generales por el despliegue de sistemas de IA, y *risk-assessment* específico en función de la complejidad de los algoritmos y objetivos perseguidos con cada modelo de IA | • Actividades definidas para garantizar el acceso, tratamiento, privacidad, protección y destrucción de los datos, especialmente en aquellos en modelos predictivos sobre comportamiento humano.<br>• Proporcionar seguridad razonable sobre los sistemas de TI y prevención de ciber-ataques. | • Asegurar un entendimiento adecuado de la operatividad de los algoritmos y *output* esperado.<br>• Definición de métricas e indicadores de desempeño para la monitorización continua.<br>• Revisión del comportamiento de los modelos en fases tempranas del despliegue de los motores. | • Desarrollar las pruebas de implementación suficientes para garantizar que el despliegue de los motores atiende los objetivos esperados.<br>• Aprobaciones pertinentes de los Comités de Proyecto establecidos previo al *go-live*. | • Revisión periódica de las métricas e indicadores de desempeño.<br>• Mecanismos de control para la identificación de comportamientos anómalos de desempeño, incluyendo las acciones correctivas necesarias a los algoritmos. |

Fuente: Fabrica del Pensamiento

# Aplicaciones en Ciberseguridad

# Detección temprana de ataques análisis (UBA/UEBA)

**Descripción detallada**

La IA permite crear perfiles de comportamiento normal de usuarios, endpoints, aplicaciones y servicios. Comparando el comportamiento actual con el histórico y el esperado, la IA detecta anomalías complejas y patrones invisibles para reglas tradicionales:

- acceso a datos atípicos
- movimientos laterales lentos (low and slow)
- cambios de privilegios irregulares
- uso extraño de credenciales
- exfiltración gradual
- sesiones simultáneas desde ubicaciones imposibles

Es clave en ataques avanzados y silenciosos (APT, ransomware preparado, insiders).

**Criticidad: ALTA**

Porque la ventana de detección temprana determina si un ataque queda contenido o llega a cifrar, robar o destruir información.

**Elementos de control**

- Política de monitoreo avanzado
- Modelos IA de comportamiento basados en telemetría completa
- Integración con SIEM/SOC
- Alertas de riesgo adaptativo (RBA)
- Umbrales dinámicos automatizados
- Revisión manual periódica de falsos positivos y drift del modelo
- Segmentación de red basada en comportamiento

**Ejemplos específicos de aplicación**

- Detectar que un analista de backoffice accede a 3.000 fichas de beneficiarios en menos de 10 minutos (comportamiento inusual y con patrón de exfiltración).
- Identificar que una cuenta de servicio comienza a ejecutar comandos de reconocimiento propios de ransomware (“whoami”, “net view”, “wmic”).
- Detectar accesos simultáneos desde Santiago y Turquía → secuestro de sesión.

**Riesgos**

- Exfiltración silenciosa de datos sensibles.
- Movimientos laterales prolongados sin detección.
- Escalamiento de privilegios no autorizado.
- Compromiso de cuentas privilegiadas sin evidencia visible.

**Controles**

- UEBA integrado a SIEM.
- Telemetría completa (endpoint, red, identidad).
- Perfiles dinámicos de usuario y máquina.
- Alertas basadas en contexto (riesgo, hora, ubicación).
- Validación humana de anomalías críticas.
- Segmentación de red y microsegmentación para limitar movimientos.

“La clave es anticiparse: detectar comportamientos anómalos evita exfiltraciones silenciosas y se controla con perfiles dinámicos, telemetría completa y segmentación inteligente.”

# Análisis automatizado de malware y clasificación inteligente

### Descripción detallada

La IA analiza ejecutables, scripts, macros, llamadas al sistema y patrones de comportamiento para identificar malware desconocido. La clasificación automática permite detectar variantes polimórficas, malware sin firmas y campañas específicas contra entidades públicas.
El análisis dinámico en sandbox con IA aprende comportamientos y reduce la dependencia de firmas.

### Criticidad: ALTA

Porque muchas instituciones públicas dependen de antivirus tradicionales que no detectan variantes nuevas.

### Elementos de control

- Integración de EDR con IA
- Sandboxing avanzado
- Contención automática por políticas
- Uso de features comportamentales (no sólo firmas)
- Modelos reforzados con datasets limpios y auditados
- Pruebas adversariales regulares

### Ejemplos específicos de aplicación

- Detectar que un documento Word ejecuta en segundo plano "powershell -nop -w hidden" para conectarse a un C2.
- Identificar variantes nuevas de Emotet que cambian hash en cada ejecución.
- Análisis dinámico de adjuntos en correos de municipalidades con sandboxing IA.

### Riesgos

- Infecciones zero-day no detectadas por firmas tradicionales.
- Entrada de malware polimórfico en estaciones de trabajo.
- Compromiso de servidores debido a explotación de macros o scripts.

### Controles

- EDR con IA y detección por comportamiento.
- Sandboxing automatizado para archivos desconocidos.
- Bloqueo automático de extensiones de alto riesgo.
- Listado de procesos permitidos (application control).
- Actualización continua de modelos de clasificación.

"Cuando el malware evoluciona más rápido que las firmas, el control está en análisis IA por comportamiento y sandboxing que reduzca el riesgo de infecciones zero-day."

# Predicción de amenazas mediante modelos de inteligencia anticipatoria

### Descripción detallada

La IA analiza tendencias globales, TTP de grupos criminales, flujos de explotación, vectores más usados por ransomware, patrones de campañas de phishing y datos OSINT.

Crea modelos predictivos que permiten anticipar:

- sectores más atacados en el mes
- CVE que se volverán críticos
- campañas dirigidas
- georreferencias de ataques

### Criticidad: MEDIA-ALTA

Reduce exposición anticipando parches, reforzando controles y priorizando activos críticos.

### Elementos de control

- Integración de OSINT y Threat Intelligence automatizado
- Modelos de predicción basados en MLOps
- Métricas de precisión y recall para evaluar el modelo
- Comité de riesgo cibernético para acción anticipatoria
- Ajustes mensuales de reglas SOC según predicciones

### Ejemplos específicos de aplicación

- El sistema alerta que en los próximos días habrá explotación masiva del CVE-2023-27997 (Fortinet SSL-VPN).
- Detección de incremento de campañas dirigidas a instituciones educativas previo al inicio de clases.
- Identificación de credenciales filtradas de funcionarios en foros de dark web.

### Riesgos

- Ataques dirigidos sin preparación previa.
- Parches críticos sin aplicar en tiempo.
- Exposición a vulnerabilidades de alto impacto (RCE).

### Controles

- Integración de plataformas OSINT con IA.
- Dashboard de riesgo anticipado por activo.
- Procesos de parchado priorizado basado en predicciones.
- Comité de riesgos cibernéticos para decisiones rápidas.

“Prever ataques antes de que ocurran reduce exposición; el riesgo se mitiga con modelos predictivos, parchado priorizado y decisiones aceleradas del comité de riesgo.”

# Automatización inteligente de SOC y respuesta a incidentes (SOAR + IA)

### Descripción detallada

La IA conecta herramientas del SOC y automatiza acciones repetitivas:

- aislar hosts
- bloquear IOCs
- cortar sesiones sospechosas
- revocar tokens de autenticación
- generar tickets priorizados
- recopilar evidencia automatizada

Reduce drásticamente los tiempos de contención y mejora la calidad del análisis.

### Criticidad: ALTA

En ransomware, minutos significan diferencias millonarias en impacto.

### Elementos de control

- Playbooks automáticos supervisados
- Aprobación humana para acciones críticas
- Métricas de MTTD/MTTR
- Versionamiento de reglas de automatización
- Registro auditado de todas las acciones automáticas

### Ejemplos específicos de aplicación

- Aislamiento automático de un endpoint cuando detecta cifrado masivo de archivos.
- Revocar tokens de sesión O365 al detectar actividad imposible de geolocalización.
- Enviar automáticamente al firewall la IP maliciosa identificada por varias fuentes OSINT.

### Riesgos

- Respuestas lentas ante incidentes críticos.
- Saturación del SOC por alertas repetitivas.
- Alto riesgo de ransomware sin contención inicial.

### Controles

- Playbooks automáticos preaprobados.
- Registro 100% auditable de acciones del SOAR.
- Doble validación para acciones destructivas.
- Integración con sistemas de ticketing.
- Métricas MTTR y MTTD optimizadas por IA.

“La velocidad salva: automatizar la respuesta reduce el impacto del ransomware, siempre bajo control de playbooks auditados y validación humana en acciones críticas.”

# Protección contra phishing con análisis semántico y contextual

### Descripción detallada

La IA analiza lenguaje, intención, estructura, URL, reputación del remitente y hasta el estilo de escritura.

Detecta phishing sin depender de listas negras o firmas.

Puede bloquear, alertar o educar al usuario en tiempo real.

Incluye análisis de deepfake de voz e imágenes generadas con IA.

### Criticidad: ALTA (sector público ya sufre campañas masivas)

El 80% de los incidentes graves comienzan por ingeniería social.

### Elementos de control

- Filtros IA basados en NLP
- Clasificación en tiempo real
- Sandboxing de adjuntos
- Simulaciones de phishing automatizadas
- Control de reputación del dominio y DKIM/DMARC/SPF
- Registro de campañas por usuario y estadísticas de riesgo

### Ejemplos específicos de aplicación

- Detectar que un correo que simula ser del BancoEstado utiliza construcción lingüística típica de campañas nigerianas (NLP).
- Detectar deepfakes de voz pidiendo una transferencia "urgente".
- Identificar que URLs acortadas redirigen a clones de sitios institucionales.

### Riesgos

- Robo de credenciales.
- Transferencias fraudulentas.
- Infección interna mediante adjuntos maliciosos.
- Compromiso de cuentas de correo institucional.

### Controles

- Filtros IA antiphishing (NLP, análisis semántico).
- DMARC, DKIM, SPF reforzados.
- Sandboxing para adjuntos sospechosos.
- MFA obligatorio en accesos remotos.
- Simulaciones periódicas de phishing con IA.

"El phishing ya no se detecta por firmas, sino por intención; el riesgo baja combinando análisis semántico, sandboxing y políticas estrictas de autenticación y correo seguro."

# Identificación inteligente de vulnerabilidades y soporte DevSecOps

### Descripción detallada

Modelos IA integrados en SAST/DAST/IAST revisan código, infraestructura como código (IaC), APIs, contenedores y pipelines CI/CD.

Priorizan vulnerabilidades según:

- exposición real
- criticidad del activo
- rutas de explotación
- impacto potencial

Reduce deuda técnica y mejora la calidad del software institucional.

### Criticidad: MEDIA-ALTA

Porque muchas brechas se deben a código vulnerable y falta de priorización.

### Elementos de control

- Integración CI/CD con IA
- Escaneo continuo
- Priorización basada en riesgo real
- Métricas de cobertura
- Capacitación en seguridad para equipos de desarrollo
- Reglas de bloqueo automático para builds críticas

### Ejemplos específicos de aplicación

- El pipeline CI/CD detiene un despliegue porque detectó una inyección SQL en un API ciudadano.
- Identificar hardcoded passwords en repositorios internos.
- Detectar configuraciones peligrosas en Kubernetes (pods con privilegios elevados).

### Riesgos

- Explotación remota de servicios críticos.
- Fugas de datos desde APIs mal configuradas.
- Publicación de código vulnerable en producción.

### Controles

- Escaneo continuo en cada commit.
- Análisis de infraestructura como código (IaC).
- Priorización automática según exposición real.
- Políticas de "build fail" para vulnerabilidades críticas.
- Capacitación obligatoria a desarrolladores.

"El código seguro se asegura en el pipeline: identificar y bloquear vulnerabilidades antes del despliegue reduce riesgos y se controla con SAST/DAST/IAST integrados."

# Autenticación basada en riesgo (Adaptive MFA + IA)

## Descripción detallada

La IA evalúa en tiempo real el riesgo de una autenticación:

- ubicación inesperada
- cambios de dispositivo
- patrones biométricos irregulares
- accesos fuera de horario
- velocidad de desplazamiento imposible

Si el riesgo es alto, obliga MFA reforzado o bloquea.

**Criticidad: ALTA (identidades son el nuevo perímetro)**

## Elementos de control

- IA para scoring de riesgo
- MFA adaptativo
- Reglas para contextos de alto riesgo
- Políticas de passwordless
- Auditoría continua de accesos privilegiados
- Integración con PAM

## Ejemplos específicos de aplicación

- Solicitar MFA adicional cuando un usuario intenta acceder desde un dispositivo nuevo.
- Bloquear un acceso porque la ubicación cambia de Valparaíso a China en 2 minutos.
- Reforzar autenticación cuando el usuario accede a bases con datos sensibles.

## Riesgos

- Secuestro de sesiones.
- Suplantación de identidad.
- Robos de credenciales mediante phishing.

## Controles

- Motor de scoring de riesgo dinámico (IA).
- MFA adaptativo por contexto.
- Controles de acceso basados en atributo (ABAC).
- Auditoría continua de accesos privilegiados.
- Integración con PAM.

"La identidad es el nuevo perímetro: el riesgo se controla evaluando contexto y comportamiento, aplicando MFA adaptativo y supervisando accesos privilegiados."

# Monitoreo inteligente de proveedores y cadena de suministro digital

### Descripción detallada

La IA analiza conexiones de terceros, APIs, integraciones, cambios en patrones de uso y actividad anómala en servicios críticos.
Detecta vulnerabilidades o comportamientos de riesgo incluso si provienen de software externo (SolarWinds, MOVEit).

### Criticidad: ALTA

La mitad de los incidentes graves hoy proviene de proveedores.

### Elementos de control

- Evaluaciones automáticas de riesgo de terceros
- IA para analizar telemetría externa
- Inventario dinámico de integraciones
- Score de riesgo continuo
- Alertas ante cambios en contratos o endpoints
- Pruebas de seguridad periódicas a proveedores

### Ejemplos específicos de aplicación

- Detectar que un proveedor de RRHH accede a la API fuera del horario pactado.
- Identificar que un proveedor SaaS envía datos a servidores en EE.UU. sin autorización formal.
- Detectar uso anómalo de credenciales de terceros en integraciones API.

### Riesgos

- Ataques indirectos vía proveedores (tipo SolarWinds).
- Exfiltración de datos personales por terceros.
- Dependencia tecnológica sin control.

### Controles

- Score de riesgo continuo con IA.
- Monitoreo de API e integraciones.
- Auditorías periódicas a proveedores críticos.
- Clasificación de proveedores por impacto.
- Restricción de permisos por mínimo privilegio.

“Los proveedores son parte del perímetro: reducir riesgos exige monitoreo continuo, control de APIs, evaluación de terceros y mínimos privilegios en cada integración.”

# Análisis inteligente de logs y reducción de falsos positivos (Noise Reduction)

### Descripción detallada

La IA clasifica, correlaciona, agrupa eventos y reduce ruido.

Detecta incidentes escondidos entre millones de logs.

Identifica patrones repetitivos y elimina falsos positivos del SOC.

### Criticidad: MEDIA-ALTA

Reduce la fatiga del personal, aumenta precisión y evita alertas ignoradas.

### Elementos de control

- Modelos IA en SIEM
- Clustering inteligente
- Correlación basada en patrones
- Priorización por impacto
- Revisión periódica de "alert drift"
- Mejora continua del SOC

### Ejemplos específicos de aplicación

- Agrupar 10.000 alertas de autenticación fallida en una sola alerta priorizada.
- Correlacionar logs para detectar exfiltración lenta vía protocolo DNS.
- Identificar ataque de password spraying antes de que comprometa cuentas.

### Riesgos

- Fatiga del SOC.
- Ignorar alertas críticas entre ruido.
- Alertas falsas que esconden incidentes reales.

### Controles

- Modelos de clustering de eventos.
- Correlación inteligente por caso de uso.
- Reducción automática de ruido (noise reduction).
- Priorizar alertas con base en impacto.
- Revisión mensual de alert drift.

"El SOC gana cuando entiende el ruido: IA reduce falsos positivos y los riesgos se controlan con correlación avanzada, priorización por impacto y revisión del alert drift."

# Uso de IA generativa para análisis, documentación y soporte técnico

### Descripción detallada

La IA generativa permite:

- redactar informes ejecutivos
- generar análisis técnicos
- crear scripts para mitigación
- resumir eventos complejos
- extraer indicadores de compromiso
- documentar playbooks
- crear manuales y políticas

Es un copiloto para analistas, reduciendo tiempos y aumentando calidad.

### Criticidad: MEDIA

Optimiza recursos y libera tiempo humano para lo crítico.

### Elementos de control

- Restricciones de privacidad (no subir datos sensibles sin anonimizar)
- Versionado de prompts y resultados
- Auditorías internas de uso de IA
- Controles para evitar fugas de información
- Validación humana obligatoria
- Políticas de uso ético de IA

### Ejemplos específicos de aplicación

- Generar automáticamente un informe técnico de un incidente a partir de logs del SIEM.
- Crear scripts para cerrar puertos abiertos o revocar credenciales en masa.
- Documentar un playbook completo en minutos.

### Riesgos

- Filtración accidental de datos sensibles en prompts.
- Dependencia excesiva de sistemas generativos externos.
- Generación de recomendaciones incorrectas si no hay supervisión.

### Controles

- Política de uso responsable de IA.
- Anonimización automática antes de enviar datos al modelo.
- Validación humana obligatoria.
- Versionado de prompts.
- Auditorías trimestrales del uso de IA generativa.

"La IA generativa potencia al analista, pero su riesgo está en la fuga de datos; se controla anonimizando información, auditando el uso y validando siempre cada salida."

# ISO 42001: Sistema de Gestión de IA

# Sistema de Gestión de IA: ISO 42001

- **Un AI-MS (Artificial Intelligence Management System) es un sistema formal basado en riesgos**, diseñado para **gobernar, controlar y asegurar** el ciclo de vida de los sistemas de IA.

- **Se estructura bajo la High Level Structure (HLS)**: política, liderazgo, roles, planificación, soporte, operación, evaluación del desempeño y mejora continua.

- **Introduce *riesgo algorítmico* como categoría obligatoria**, incorporando evaluación de sesgos, drift del modelo, ataques adversarios, explicabilidad, calidad del dataset, estabilidad del pipeline y dependencias tecnológicas.

- **Define gobernanza técnica obligatoria:**
  – AI Officer / Comité de IA
  – Dueños de modelos
  – Dueños de datasets
  – Responsables del pipeline MLOps

**Exige control sobre el ciclo de vida completo de IA:**
– diseño y arquitectura del modelo
– adquisición, procesamiento y gobernanza de datos
– entrenamiento y tuning
– validación y verificación técnica
– despliegue seguro
– monitoreo continuo y retraining
– retiro y destrucción de artefactos

# ¿Qué ofrece realmente la ISO/IEC 42001? (Beneficio técnico y de control

- **Marco técnico para IA confiable:** requisitos para explicabilidad, robustez, detección de drift, validación algorítmica y protección contra ataques adversarios (model inversion, model extraction, poisoning).

- **Gestión avanzada de datos:** clasificación, trazabilidad del dataset, auditoría del origen, licencias, equilibrio, minimización y protección de PII, integrando privacidad por diseño.

- **Controles de seguridad específicos para IA:**
  - protección del pipeline MLOps
  - integridad del modelo
  - aislamiento de entornos de entrenamiento
  - gobernanza de características (feature governance)
  - controles anti-manipulación y anti-contaminación de datos

- **Transparencia y responsabilidad:** obliga a registrar decisiones algorítmicas, documentar modelos, explicar limitaciones y asegurar mecanismos de revisión humana.

- **Control de proveedores de IA:** evaluación de plataformas externas (ChatGPT, Gemini, Bedrock), contratos con obligaciones técnicas, trazabilidad de versiones y riesgos de uso de modelos cerrados.

# Posee una estructura de un Sistema de Gestión Estándar

## PLANIFICAR

- 4 CONTEXTO DE LA ORGANIZACIÓN
  - COMPRENSIÓN DE LA ORGANIZACIÓN Y SU CONTEXTO
  - COMPRENSIÓN DE LAS NECESIDADES Y EXPECTATIVAS DE LAS PARTES INTERESADAS
  - DETERMINACIÓN DEL ALCANCE DEL SISTEMA DE GESTIÓN DE INTELIGENCIA ARTIFICIAL
  - SISTEMA DE GESTIÓN DE INTELIGENCIA ARTIFICIAL
- 5 LIDERAZGO
  - LIDERAZGO Y COMPROMISO
  - POLÍTICA DE PRIVACIDAD
  - ROLES, RESPONSABILIDADES Y AUTORIDADES
- 6 PLANIFICACIÓN
  - ACCIONES PARA ABORDAR LOS RIESGOS Y OPORTUNIDADES
  - OBJETIVOS DE PRIVACIDAD Y PLANIFICACIÓN PARA LOGRARLOS
  - PLANIFICACIÓN DE LOS CAMBIOS

## HACER

- 7 APOYO
  - RECURSOS
  - COMPETENCIA
  - CONCIENCIZACIÓN
  - COMUNICACIÓN
  - INFORMACIÓN DOCUMENTADA
- 8 OPERACIÓN
  - PLANIFICACIÓN Y CONTROL OPERACIONAL
  - EVALUACIÓN DE RIESGSO DE PRIVACIDAD
  - TRATAMIENTO DE RIESGOS DE PRIVACIDAD

## VERIFICAR

- 9 EVALUACIÓN DE DESEMPEÑO
  - SEGUIMIENTO, MEDICIÓN ANÁLISIS Y EVALUACIÓN
  - AUDITORÍA INTERNA
  - REVISIÓN POR LA DIRECCIÓN

## ACTUAR

- 10 MEJORA
  - MEJORA CONTINUA
  - NO CONFORMIDAD Y ACCIONES CORRECTIVAS

# Elementos de control claves

## 1. Gobernanza, roles y responsabilidades de IA

Controles orientados a asegurar que la IA tenga dueño, gestión y supervisión formal.

- Comité de IA o estructura equivalente.
- AI Officer / Responsable del AI-MS.
- Asignación de dueños de modelos y dueños de datasets.
- Procedimientos formales para aprobación de casos de uso de IA.
- Mecanismos de rendición de cuentas (accountability).
- Alineamiento del AI-MS con políticas organizacionales.

**Objetivo:** asegurar control institucional, no solo técnico.

## 2. Gestión del riesgo algorítmico

Controles para identificar, analizar y tratar riesgos únicos de IA.

- Identificación de riesgos específicos del sistema (bias, drift, ataques adversarios).
- Evaluación de impacto algorítmico (AIA).
- Priorización de riesgos según impacto ético, social, legal y operacional.
- Mecanismos de mitigación y monitoreo continuo.
- Evidencias de revisión periódica de riesgos.

**Objetivo:** evitar modelos dañinos, no confiables o manipulables.

## 3. Gobierno de datos y datasets

Controles para garantizar calidad, legalidad y trazabilidad del dataset.

- Identificación de origen del dataset.
- Evaluación de calidad, integridad y completitud de datos.
- Control de sesgo y balance de datos.
- Validación de licencias y derechos de uso.
- Gobernanza del procesamiento, limpieza y transformación.
- Minimización y protección de PII (alineado a ISO 27701).

**Objetivo:** asegurar datos éticos, seguros, justificados y auditables.

## 4. Ciclo de vida del modelo (ML Lifecycle Controls)

Controles aplicables desde el diseño hasta el retiro del modelo.

- Requisitos técnicos para el diseño del algoritmo.
- Documentación de arquitectura del modelo.
- Procedimientos para entrenamiento reproducible.
- Validación técnica y funcional antes del despliegue.
- Pruebas de robustez y seguridad.
- Monitoreo continuo del rendimiento.
- Gestión del model drift (desviación del patrón).
- Procedimientos de retraining.
- Planes de retiro seguro del modelo.

**Objetivo:** asegurar modelos confiables, estables y controlados en todo su ciclo.

# Elementos de control claves

## 5. Controles de seguridad específicos para IA

Nuevos controles exclusivos para proteger modelos y pipelines.

- Protección contra ataques adversarios (evasion, poisoning).
- Salvaguardas para evitar model inversion y model extraction.
- Seguridad en pipelines MLOps (entrenamiento, testing, deployment).
- Control de integridad de modelos y features.
- Aislamiento de entornos de entrenamiento y validación.
- Control de acceso granular (ABAC para IA).

**Objetivo:** impedir manipulación, robo o degradación de los modelos.

## 6. Transparencia, explicabilidad y ética

Controles para garantizar decisiones responsables y auditables.

- Información explícita al usuario cuando interactúa con IA.
- Documentación del propósito del modelo.
- Explicabilidad técnica o funcional según contexto.
- Controles para evitar discriminación o impactos indebidos.
- Evidencias de supervisión humana.
- Límites de uso definidos y aprobados.

**Objetivo:** IA responsable, entendible y legítima.

## 7. Gestión de proveedores y servicios de IA externos

Controles para IA comprada, integrada o tercerizada.

- Evaluación técnica y de riesgos del proveedor.
- Obligaciones contractuales sobre transparencia del modelo.
- Trazabilidad de versiones del modelo y del dataset.
- Restricciones claras sobre uso de datos proporcionados.
- Indicadores de desempeño y robustez del proveedor.
- Mecanismos de monitoreo y auditoría.

**Objetivo:** controlar IA de terceros como parte del perímetro.

## 8. Documentación, evidencia y auditoría

Controles para demostrar cumplimiento del AI-MS.

- Registro del diseño, entrenamiento y validación.
- Evidencias de decisión algorítmica.
- Logs del pipeline MLOps.
- Reportes de supervisión humana.
- Auditorías internas del AI-MS.
- Medición de desempeño y mejora continua.

**Objetivo:** demostrar cumplimiento técnico y organizacional para certificación.

# 4.1. Contexto Externo

**1. Analiza el marco regulatorio aplicable a IA, datos y ciberseguridad.**

Incluye leyes de protección de datos, requisitos sectoriales, normas éticas, Ley Marco de Ciberseguridad, guías regulatorias y estándares internacionales como el AI Act europeo.

**2. Identifica expectativas y preocupaciones de partes interesadas externas.**

Ciudadanos, clientes, reguladores, auditores, organismos fiscalizadores y proveedores; determina qué esperan de la IA en términos de transparencia, explicabilidad y seguridad.

**3. Evalúa tendencias tecnológicas globales en IA y su impacto.**

Incluye modelos generativos, LLM, automatización avanzada, riesgo reputacional, cambios rápidos en proveedores, y nuevas capacidades de machine learning.

**4. Considera amenazas externas específicas para IA.**

Ataques adversarios, model extraction, model inversion, poisoning de datasets, manipulación de modelos y uso abusivo de IA por actores maliciosos.

**5. Analiza la dinámica del mercado y la presión competitiva.**

Competidores, actores del sector, capacidad de adopción tecnológica, velocidad de innovación y expectativas de rendimiento basadas en IA.

**6. Estudia las dependencias tecnológicas externas.**

Servicios cloud, proveedores de APIs, modelos de IA de terceros (ChatGPT, Gemini, Bedrock, Claude), licencias de datasets, y riesgos asociados a la cadena de suministro digital.

**7. Evalúa el impacto social, cultural y ético del uso de IA.**

Riesgos de discriminación, impacto en comunidades, transparencia ante ciudadanos, sesgo sistémico y consecuencias sociales de automatizar decisiones sensibles.

**8. Considera el ecosistema de riesgos geopolíticos y de ciberseguridad.**

Conflictos globales, tensiones tecnológicas, riesgo país, ciberamenazas transnacionales, disponibilidad de infraestructura y vulnerabilidades de la región.

"Comprender el contexto externo es esencial porque define las reglas, riesgos y expectativas que rodean a la IA; sin esta visión, el sistema de gestión queda ciego ante regulaciones, amenazas y demandas de transparencia que impactan directamente su confiabilidad."

# 4.1 Contexto Interno

**1. Evalúa la madurez organizacional en datos, IA y analítica.**

Identifica si existen capacidades reales en gestión de datos, MLOps, aprendizaje automático, infraestructura y cultura digital para soportar un AI-MS robusto.

**2. Identifica los sistemas de IA existentes y sus dependencias internas.**

Modelos en producción, prototipos, automatizaciones, sistemas de recomendación, chatbots y cualquier algoritmo que influya en decisiones institucionales.

**3. Define roles, responsabilidades y capacidades del equipo.**

Competencias en ciencia de datos, ingeniería, seguridad, ética algorítmica, auditoría de modelos y gobernanza; detecta brechas críticas.

**4. Revisa políticas, normas y procesos ya implementados.**

SGSI (ISO 27001), privacidad (ISO 27701), gestión de proveedores, riesgos, continuidad, desarrollo seguro, y cómo se integrarán al AI-MS.

**5. Analiza los flujos internos de datos y su calidad.**

Trazabilidad, integridad, disponibilidad, sesgos, origen, licencias, reglas de procesamiento y su alineación con privacidad por diseño.

**6. Identifica capacidades técnicas internas reales.**

Infraestructura (cloud, on-prem, GPUs), pipelines de datos, herramientas de MLOps, repositorios de modelos y capacidades de monitoreo.

**7. Determina el apetito de riesgo interno respecto a IA.**

Tolerancia al error algorítmico, impacto permitido en decisiones sensibles, nivel de transparencia requerido y expectativas del comité ejecutivo.

**8. Evalúa restricciones internas que condicionan el AI-MS.**

Presupuesto, personal, cultura, resistencia al cambio, prioridades estratégicas, contratos vigentes, limitaciones en proveedores o sistemas legados.

“Comprender el contexto interno es crítico porque revela las capacidades, limitaciones y riesgos reales de la organización; sin esta claridad, el AI-MS se diseña sobre supuestos y no sobre la estructura operativa, tecnológica y cultural que lo debe sostener.”

# 4.2 Partes Interesadas

## 1. Clientes (empresas públicas y privadas)

**Intereses:**

- Servicios TI estables, seguros y confiables.
- Cumplimiento de SLA, continuidad y soporte oportuno.
- Protección de datos y privacidad.
- Transparencia en el uso de IA y decisiones automatizadas.

**Expectativas:**

- Alta disponibilidad, respuestas rápidas, trazabilidad.
- Cumplimiento normativo (ISO 27001, 27701, Ley de Ciberseguridad, AI-MS).
- Comunicación clara frente a incidentes o cambios.

**Riesgos asociados:**

- Incidentes que afecten continuidad o seguridad del cliente.
- Pérdida de confianza por fallas o brechas.
- Mal uso de datos o algoritmos opacos.

**Necesidades de control:**

- SLA robustos, SOC maduro, controles de seguridad y privacidad.
- Transparencia en IA, reportes claros y gestión ágil de incidentes.
- Gestión contractual sólida.

## 2. Proveedores tecnológicos / Cloud / IA

**Intereses:**

- Relación comercial estable, volumen de uso y continuidad.

**Expectativas:**

- Integraciones claras, APIs seguras, pagos oportunos.
- Gestión adecuada de configuraciones y consumo.

**Riesgos asociados:**

- Fallas del proveedor → impacto en disponibilidad del servicio TI.
- Cambios en precios o modelos de licencia.
- Riesgos algorítmicos en IA de terceros (drift, sesgo, vulnerabilidades).

**Necesidades de control:**

- Evaluación de proveedores, monitoreo continuo, acuerdos de soporte.
- Análisis de riesgos en servicios cloud e IA generativa.

## 3. Reguladores y entidades fiscalizadoras

**Intereses:**

- Cumplimiento normativo, seguridad de la información, privacidad, IA responsable.

**Expectativas:**

- Controles basados en estándares (ISO 27001, 27701, 42001).
- Evidencias auditables, registros de decisiones, protocolos de incidentes.

**Riesgos asociados:**

- Sanciones, incumplimientos, daño reputacional.
- Falta de trazabilidad o gobernanza.

**Necesidades de control:**

- AI-MS, SGSI, PIMS y registros de auditoría robustos.
- Gestión transparente y documentada.

# 4.2 Partes Interesadas

## 4. Usuarios finales / Consumidores de servicios

**Intereses:**

- Experiencia fluida, soporte rápido, seguridad y privacidad garantizada.

**Expectativas:**

- Sistemas estables, interfaces claras, decisiones automatizadas explicables.

**Riesgos asociados:**

- Pérdida de confianza por fallas o brechas.
- Sesgos algorítmicos que afecten decisiones de servicio.

**Necesidades de control:**

- UX segura, MFA accesible, explicabilidad algorítmica.
- Controles de privacidad y protección de datos personales.

## 5. Equipo interno / Personal TI

**Intereses:**

- Capacitación, herramientas adecuadas y claridad de roles.

**Expectativas:**

- Infraestructura estable, procesos definidos, gobernanza clara.

**Riesgos asociados:**

- Errores operativos por falta de capacitación.
- Dependencia excesiva de personas clave.
- Brechas internas por configuraciones erróneas.

**Necesidades de control:**

- Capacitación continua, MLOps documentado, roles y responsabilidades claras.
- Automatización segura y monitoreo constante.

## 6. Dirección ejecutiva / Gerencia

**Intereses:**

- Rentabilidad, continuidad, reputación y diferenciación competitiva.

**Expectativas:**

- Información oportuna y decisiones basadas en datos.
- Cumplimiento normativo y riesgos controlados.

**Riesgos asociados:**

- Impacto reputacional.
- Pérdidas económicas por incidentes o mala gestión de IA.

**Necesidades de control:**

- Reportes ejecutivos, métricas de riesgo, indicadores de desempeño del AI-MS.
- Alineamiento estratégico del uso de IA.

## 7. Socios estratégicos / Integradores

**Intereses:**

- Negocios estables, proyectos conjuntos, integraciones seguras.

**Expectativas:**

- Estándares alineados y arquitectura interoperable.

**Riesgos asociados:**

- Integraciones inseguras.
- Responsabilidades difusas en incidentes compartidos.

**Necesidades de control:**

- Contratos claros, RACI, controles de API y evaluaciones conjuntas.

## 8. Comunidad, sociedad civil y opinión pública

**Intereses:**

- IA ética, decisiones justas, protección de derechos ciudadanos.

**Expectativas:**

- Transparencia, explicabilidad y responsabilidad social.

**Riesgos asociados:**

- Daño reputacional por fallas o sesgos.
- Desconfianza en soluciones automatizadas.

**Necesidades de control:**

- Política de ética en IA, reportes de impacto, mecanismos de reclamo.

# 5.2 Política del SGIA

### 1. Compromiso con el uso responsable, seguro y ético de la IA

La política debe declarar que la organización solo desarrollará, desplegará y utilizará IA bajo principios de responsabilidad, transparencia, equidad, explicabilidad y respeto a los derechos de las personas.

### 2. Enfoque basado en riesgos algorítmicos

Debe establecer que todos los sistemas de IA serán evaluados considerando riesgos como: sesgo, discriminación, drift, ataques adversarios, filtración de datos, impacto social, seguridad y privacidad.

### 3. Alcance y aplicación del AI-MS

La política debe definir claramente:

- qué modelos, algoritmos y casos de uso quedan bajo el AI-MS
- qué unidades, procesos y servicios están cubiertos
- qué tecnologías y proveedores forman parte del sistema

### 4. Roles y responsabilidades internas

La política debe presentar la estructura de gobernanza:

- AI Officer o Responsable de IA
- Comité de IA
- Dueños de modelos
- Dueños de datasets
- Responsables de MLOps, datos y seguridad

  Cada rol con sus deberes, autoridad y rendición de cuentas.

### 5. Lineamientos para el ciclo de vida del modelo

Debe establecer que todos los modelos de IA deben seguir un ciclo controlado:
diseño → dataset → entrenamiento → validación → despliegue → monitoreo → retraining → retiro.
Cada fase con criterios documentados y aprobados.

### 6. Transparencia y explicabilidad

La política debe exigir que los sistemas de IA:

- sean explicables en el nivel adecuado
- informen al usuario cuando interactúa con IA
- registren decisiones automatizadas
- documenten sus limitaciones y supuestos

### 7. Controles de seguridad y privacidad aplicables a IA

Debe incluir lineamientos mínimos sobre:

- protección de datos
- control de acceso
- seguridad del pipeline MLOps
- mitigación de ataques adversarios
- gobernanza del dataset
- minimización y trazabilidad de datos
- cumplimiento con ISO 27001 y 27701

### 8. Control de proveedores y servicios de IA externos

La política debe declarar cómo se gestionan proveedores de IA:

- evaluación de riesgos
- exigencias contractuales
- protección de datos
- transparencia de versiones
- controles sobre el uso de modelos cerrados o APIs externas

# 5.3 Roles y funciones claves

## 1. AI Officer / Responsable del Sistema de Gestión de IA

**Funciones principales:**

- Liderar la implementación, operación y mejora del SGIA.
- Coordinar procesos de evaluación de riesgos algorítmicos.
- Supervisar cumplimiento normativo, ético y técnico del uso de IA.
- Centralizar reportes de desempeño, incidentes y desviaciones de modelos.
- Coordinar auditorías internas y externas del SGIA.

**Enfoque técnico:**

Garantiza la gobernanza transversal y la trazabilidad del uso de IA.

## 2. Comité de IA (AI Governance Committee)

**Funciones principales:**

- Aprobar casos de uso de IA y evaluar su impacto.
- Revisar riesgos, sesgos, decisiones automatizadas y criterios éticos.
- Supervisar el desempeño global de los modelos de la organización.
- Autorizar el retiro, actualización o restricción de modelos problemáticos.
- Tomar decisiones estratégicas y de alto impacto.

**Enfoque técnico:**

Define límites, criterios de transparencia y políticas de uso responsable.

## 3. Dueño del Modelo (Model Owner)

**Funciones principales:**

- Definir propósito, alcance y requisitos funcionales del modelo.
- Asegurar que el modelo cumpla con criterios técnicos y regulatorios.
- Aprobar cambios en arquitectura, datasets o configuraciones.
- Supervisar el rendimiento, drift, performance y errores.

**Enfoque técnico:**

Responsable del modelo como "activo vivo" durante todo su ciclo de vida.

## 4. Dueño del Dataset (Data Owner)

**Funciones principales:**

- Supervisar calidad, integridad, trazabilidad y licencias del dataset.
- Asegurar minimización, balance y protección de PII.
- Autorizar el uso de datos para entrenamiento o validación.
- Detectar sesgos o fallas de origen en los datos utilizados.

**Enfoque técnico:**

Custodia los cimientos del modelo y controla el impacto de la materia prima.

## 5. Equipo MLOps / Ingenieros de IA

**Funciones principales:**

- Construir pipelines de entrenamiento, validación, deployment y monitoreo.
- Asegurar reproducibilidad de versiones del modelo y datasets.
- Implementar controles de seguridad y robustez técnica.
- Realizar retraining, tuning y actualizaciones controladas.

**Enfoque técnico:**

Garantiza que el ciclo de vida del modelo sea seguro, estable y automatizado.

## 6. Equipo de Seguridad de la Información / Ciberseguridad

**Funciones principales:**

- Realizar pruebas de adversarial ML (evasion, poisoning, extraction).
- Asegurar integridad del modelo y del pipeline MLOps.
- Implementar controles de acceso y protección contra fugas.
- Monitorear riesgos de proveedores externos.

**Enfoque técnico:**

Protege al modelo contra ataques específicos de IA.

## 7. Equipo de Privacidad / Oficial de Datos (DPO)

**Funciones principales:**

- Evaluar impacto en privacidad (PIA + elementos de AIA).
- Supervisar licencias, minimización, uso legítimo y transferencias de datos.
- Validar que el modelo cumpla con ISO 27701, leyes de datos y principios éticos.
- Revisar decisiones automatizadas que afecten derechos de las personas.

**Enfoque técnico:**

Alinea la IA con privacidad por diseño y responsabilidad legal.

## 8. Usuario Experto del Modelo / Stakeholder Funcional

**Funciones principales:**

- Operar el modelo o interactuar con la interfaz donde se aplica IA.
- Reportar errores funcionales, decisiones inesperadas o fallas de servicio.
- Aportar criterio experto del dominio para validar el uso del modelo.

**Enfoque técnico:**

Punto de control operacional para detectar fallas o desviaciones reales.

## 9. Auditor Interno del SGIA

**Funciones principales:**

- Evaluar conformidad con ISO 42001.
- Auditar evidencia de decisiones, diseño, entrenamiento y monitoreo.
- Revisar transparencia, explicabilidad, seguridad y control de riesgos.
- Emitir hallazgos y recomendaciones de mejora continua.

**Enfoque técnico:**

Asegura la madurez y la trazabilidad del manejo de IA en toda la organización.

# Riesgos y controles ISO 42001

# Gestión de Riesgos de IA

## 🔥 1. Sesgo Algorítmico (Bias)

### Descripción técnica

Ocurre cuando el modelo aprende patrones históricos desbalanceados o discriminatorios. El origen puede estar en el dataset, la ingeniería de features o en la arquitectura del modelo.

### Ejemplo

Un modelo de asignación de beneficios prioriza solicitudes urbanas por sobre rurales debido a datos históricos con sobrerrepresentación urbana.

**Criticidad: ALTA – impacto legal, ético y reputacional.**

**Cláusulas ISO 42001: 4.2, 6.2, 8.4, 8.5, 8.3**

### Controles

- Evaluación de sesgos del dataset (representatividad, balance).
- Métricas de equidad (DP, EO, DI).
- Evaluación de impacto algorítmico (AIA).
- Comité de ética/IA para aprobación de modelos.
- Documentación del propósito y limitaciones.

## 🔥 2. Deriva del Modelo (Model Drift / Data Drift)

### Descripción técnica

El rendimiento baja con el tiempo porque cambió la distribución de datos o el comportamiento real.

### Ejemplo

El modelo de detección de phishing baja del 92% al 71% de precisión porque las campañas de phishing adoptaron nuevas técnicas.

**Criticidad: ALTA – afecta decisiones automatizadas.**

**Cláusulas: 8.5, 8.6**

### Controles

- Monitoreo continuo del rendimiento.
- Alarmas por variaciones de precisión/recall.
- Trazabilidad de datasets y modelos.
- Retraining planificado y adaptativo.

# Gestión de Riesgos de IA

## 🔥 3. Envenenamiento de Datos (Data Poisoning)

### Descripción técnica

Actores maliciosos insertan datos falsificados en el dataset de entrenamiento para manipular el modelo.

### Ejemplo

Un actor altera logs internos para que el modelo de detección de anomalías ignore ciertos patrones.

**Criticidad: MUY ALTA – compromiso total del modelo.**

**Cláusulas: 8.4, 8.7**

### Controles

- Validación de integridad (hashing, firmas).
- Detección de outliers anómalos en dataset.
- Aislamiento de entornos de entrenamiento.
- Dataset governance estricta.

## 🔥 4. Filtración de Datos (Model Inversion / Leakage)

### Descripción técnica

El modelo permite inferir datos personales del set de entrenamiento.

### Ejemplo

Consultas repetitivas a un modelo permiten reconstruir partes del perfil de usuarios.

**Criticidad: ALTA – riesgo de privacidad severo.**

**Cláusulas: 8.4, 8.7**

### Controles

- Minimización y anonimización.
- Límites de consulta (rate limiting).
- Auditoría de accesos.
- Privacidad diferencial.

# Gestión de Riesgos de IA

## 🔥 5. Robo del Modelo (Model Extraction)

**Descripción técnica**

Un atacante replica un modelo consultándolo miles de veces para reproducir su comportamiento.

**Ejemplo**

Un competidor reconstruye un modelo de scoring de riesgo de TI a través de queries masivas.

**Criticidad: MEDIA-ALTA – propiedad intelectual y seguridad.**

**Cláusulas: 8.7.3, 7.5**

**Controles**

- Rate limiting.
- Autenticación por token.
- Monitoreo de patrones de consulta.
- Aislamiento de endpoints críticos.

## 🔥 6. Ataques Adversarios (Adversarial Attacks)

**Descripción técnica**

Pequeñas perturbaciones en entradas engañan al modelo.

**Ejemplo**

Imágenes alteradas que evitan reconocimiento o detección automática.

**Criticidad: ALTA – manipulación directa.**

**Cláusulas: 8.7.3**

**Controles**

- Pruebas adversariales periódicas.
- Robustez del modelo (FGSM, PGD testing).
- Endurecimiento de features.
- Aislamiento en inferencia.

# Gestión de Riesgos de IA

## 🔥 7. Decisiones Automatizadas No Explicables

### Descripción técnica

El modelo no puede explicar criterios usados para decisiones que afectan personas.

### Ejemplo

Un sistema aprueba o rechaza solicitudes sin explicación.

**Criticidad: ALTA**

**Cláusulas: 8.3, 5.3**

### Controles

- SHAP, LIME, explicabilidad técnica.
- Registro de decisiones automáticas.
- Human-in-the-loop para decisiones críticas.

## 🔥 8. Uso Indebido del Modelo

### Descripción técnica

Usar el modelo para un propósito distinto del aprobado.

### Ejemplo

Reutilizar un modelo de clasificación documental para perfilamiento de personas.

**Criticidad: ALTA**

**Cláusulas: 5.2, 7.3, 8.1**

### Controles

- Declaración de propósito.
- Límites de uso.
- Auditorías internas del SGIA.

# Gestión de Riesgos de IA

## 🔥 9. Fallas en el Pipeline MLOps

### Descripción técnica

Problemas de seguridad o integridad en el flujo de entrenamiento, testeo y despliegue.

### Ejemplo

Un ingeniero sube una versión del modelo sin validación.

**Criticidad: MUY ALTA**

**Cláusulas: 8.7, 8.5**

### Controles

- Integridad del pipeline.
- Control de versiones.
- Separación de ambientes.
- Validación obligatoria.

## 🔥 10. Calidad Deficiente del Dataset

### Descripción técnica

Datos incompletos, corruptos o inconsistentes generan modelos inexactos.

### Ejemplo

Datos con campos vacíos o erróneos afectan predicciones.

**Criticidad: ALTA**

**Cláusulas: 8.4**

### Controles

- Controles de calidad.
- Validación automática del dataset.
- Gobernanza de datos (data stewardship).

# Gestión de Riesgos de IA

## 🔥 11. Riesgos Éticos y de Impacto Social

### Descripción técnica

El modelo perjudica grupos o vulnera principios éticos.

### Ejemplo

Un bot institucional da respuestas discriminatorias.

### Criticidad: ALTA

### Cláusulas: 5.3, 6.2, 8.3

### Controles

- Evaluación de impacto ético.
- Comité de IA.
- Límites a decisiones automatizadas.

## 🔥 12. Error Operacional Humano en IA

### Descripción técnica

Errores en la preparación del dataset, actualización o interpretación del modelo.

### Ejemplo

Un analista carga mal un dataset de retraining.

### Criticidad: MEDIA-ALTA

### Cláusulas: 7.2, 8.4, 8.6

### Controles

- Capacitación.
- SOP documentados.
- Doble validación humana.

# Gestión de Riesgos de IA

## 🔥 13. Riesgos en Proveedores de IA

### Descripción técnica

Dependencia de modelos o APIs externas sin suficiente control.

### Ejemplo

Caída de un proveedor deja inoperante un sistema de screening documental.

**Criticidad: MEDIA-ALTA**

**Cláusulas: 7.4, 7.5, 8.7.5**

### Controles

- Evaluación de riesgo del proveedor.
- Acuerdos contractuales robustos.
- Monitoreo continuo.

## 🔥 14. Propagación de Errores por Automatización

### Descripción técnica

Pequeños errores se amplifican por sistemas automatizados.

### Ejemplo

Un modelo mal calibrado envía miles de solicitudes incorrectas.

**Criticidad: ALTA**

**Cláusulas: 8.5, 8.6**

### Controles

- Validación previo al despliegue.
- Pruebas exhaustivas.
- Mecanismos de rollback.

# Gestión de Riesgos de IA

## 🔥 15. Riesgos de Privacidad por Entrenamiento o Inferencia

**Descripción técnica**

El modelo utiliza PII sin base legal, o la expone en inferencia.

**Ejemplo**

IA generativa que reconstruye datos sensibles.

**Criticidad: ALTA**

**Cláusulas: 8.4, relación con PIMS/ISO 27701**

**Controles**

- Minimización.
- Seudonimización/anónimos.
- Control de acceso.

## 🔥 16. Riesgos de Seguridad en Infraestructura IA

**Descripción técnica**

Fallas en GPU, cloud, orquestadores o pipelines permiten intrusiones.

**Ejemplo**

Acceso no autorizado al storage donde se guardan modelos.

**Criticidad: ALTA**

**Cláusulas: 7.5, 8.7**

**Controles**

- Zero Trust MLOps.
- Doble autenticación.
- Segmentación.
- Auditoría de logs.

# Fuentes de Referencia


- Modelo de provisión de servicios - FourweekMba
- https://fourweekmba.com/es/aiaas/

- Mapa de Herramientas de Ciberseguridad – Optiv
- https://www.optiv.com/sites/default/files/images/Cybersecurity-Technology-Map-Web-min.png

- Riesgos en IA – Gartner
- https://blogs.gartner.com/avivah-litan/2021/01/21/top-5-priorities-for-managing-ai-risk-within-gartners-most-framework/

- Auditoría Interna de la Inteligencia Artificial aplicada a procesos de negocio
- https://auditoresinternos.es/centro-de-conocimiento/fabrica-de-pensamiento/

- Artificial Intelligence and Cybersecurity Research – ENISA
- https://www.enisa.europa.eu/publications/artificial-intelligence-and-cybersecurity-research?v2=1

CAPACITACIÓN USACH
UNIVERSIDAD DE SANTIAGO DE CHILE

# ISO 42001:
# Sistema de Gestión de Inteligencia Artificial

## Conceptos Claves, Riesgos y Controles

ALIGNMENT
SOLUCIONES ESTRATÉGICAS

DIPLOMADO
CIBERSEGURIDAD