Ampliación Riesgos y Límites

Riesgo	Descripción	Gravedad	Prob.	Prevención	Contingencia
Sesgo algorítmico	El modelo asigna mayor riesgo a determinados grupos por correlaciones espurias	Alta	Media	Auditorías de equidad trimestrales; control del pipeline de datos; anonimización y balanceo de datasets; métricas (demographic parity, equalized odds)	Congelar despliegue de la versión afectada; reentrenamiento con dataset corregido; informe público y medidas correctoras.
Falsos positivos	Alertas que señalan riesgo cuando no lo hay; generan intervenciones innecesarias	Alta	Media	Calibración continua de umbrales; requisito de revisión humana antes de cualquier acción; pruebas en entornos piloto.	Anulación del registro erróneo; sesión de revisión por comité ético; ajuste de umbral y notificación a afectados.
Falsos negativos	No detectar un conflicto real en tiempo útil.	Alta	Media	Modelos redundantes y validación cruzada; pruebas de estrés; sensores alternativos (presencia, puertas).	Supervisión humana intensificada; revisión inmediata del caso; parche del modelo y advertencia al resto de centros.
Fuga de datos	Acceso no autorizado a vídeo, audio o registros.	Alta	Baja	Cifrado en reposo/transit (AES-256/TLS); control de acceso por roles; logging y rotación de claves; servidores en la UE.	Notificación al DPO y AEPD (art.33 RGPD) en plazo legal; bloqueo de accesos; investigación forense; comunicación interna y externa.
Manipulación	Internos o empleados intentan engañar/evitar detección	Media	Media	Multimodalidad (audio+visión+contexto); detección de anomalías; hardening físico de	Reentrenamiento; parche de detección de evasión;

	(camuflaje, interferencia).			cámaras.	medidas disciplinarias y refuerzo del protocolo físico.
Dependencia de proveedor / lock-in	Software/servicio propietario impide migración o encarece mantenimiento	Media	Media	Preferir componentes libres, APIs abiertas, exportación de datos y modelos; cláusulas contractuales sobre interoperabilidad.	Reversión a alternativas open; migración planificada; compra de código/servicios críticos si necesario.
Fallo crítico de sensores/software	Caída masiva de cámaras/sensores o bug crítico que genera ruido masivo	Alta	Media	Redundancia hardware; monitorización 24/7; pruebas automáticas (CI/CD) y despliegues canary	Activación de modo manual (supervisión humana total); rollback de versión; plan de contingencia técnico.
Pérdida de confianza del personal	Automatización de confianza provoca dependencia y pérdida de juicio crítico.	Media	Media	Formación obligatoria inicial y continuada; métricas de uso para detectar "automation bias"; panel explicativo obligatorio antes de actuar.	Campañas de reciclaje; suspensión temporal del sistema para readiestramiento operativo; sanciones administrativas si hay negligencia reiterada.
Riesgo psicosocial (ansiedad por vigilancia)	Sensación de hipercontrol en internos y personal, afectando derechos de reinserción.	Media	Media	Comunicación clara de finalidad; limitar alcance a detección de incidentes; aplicar principios de minimización; evaluación psicosocial pre-y post-despliegue.	Ajustes operativos (reducir coberturas), medidas de acompañamiento psicológico, suspensión parcial si hay impacto significativo

Plausibilidad (2029)

- Tecnologías clave (esqueletos, XAI, multimodalidad) son plausibles y maduras para 2029, pero la **reidentificación absoluta** y ataques adversariales son riesgos tecnológicos que requieren supervisión continua y EIPD iterativas.

Líneas rojas (prohibiciones)

1. Nunca decisiones automáticas punitivas.
2. Nunca almacenar datos biométricos identificables (rostros/voz identificable).
3. Nunca usar resultados como prueba judicial/disciplinaria directa sin revisión humana independiente.
4. Nunca comercializar o transferir datos fuera de ámbito institucional sin salvaguardas legales.

Emergencias / paradas

- **Quién puede parar:** Director del centro, Responsable de Seguridad/DPO, AEPD/autoridad judicial.
- **Cuándo:** FP rate súbito >20%; brecha de seguridad; orden del comité ético/AEPD.
- **Cómo (procedimiento):** Activar *modo seguro* (detener alertas, modo sólo-lectura logs), informe 24h, análisis 72h, decisión colegiada para reactivar o retirar.

Rendición de cuentas

- **Canal de reclamación:** formulario interno / legal@prisonrisk-ai.es / buzón físico.
- **Plazos orientativos:** acuse 48h; respuesta provisional 15 días hábiles; resolución final 3 meses.
- **Revisión:** humana obligatoria; escala: funcionario → jefe servicio → comité mixto (IA+DPO+ético).
- **Reparación:** reversión/eliminación de registros, corrección de expediente, comunicaciones y medidas preventivas.
- **Responsables:** Ministerio del Interior (titular) y proveedor (corresponsable contractual).

Referencias

Akhtar, N., & Mian, A. (2018). *Threat of Adversarial Attacks on Deep Learning in Computer Vision: A Survey*. arXiv. https://arxiv.org/abs/1801.00553

Slack, D., Hilgard, S., Jia, E., Singh, S., & Lakkaraju, H. (2019). *Fooling LIME and SHAP: Adversarial Attacks on Post hoc Explanation Methods*. arXiv. https://arxiv.org/abs/1911.02508

IBM España. (s. f.). ¿Qué es el sesgo de la IA? https://www.ibm.com/es-es/think/topics/ai-bias

AICAD. (s. f.). *Falsos positivos de la IA*. https://www.aicad.es/falsos-positivos-de-la-ia

BDO España. (2024, 25 abril). *Riesgos de sesgo y discriminación en la inteligencia artificial (IA)*. Blog de BDO. https://www.bdo.es/es-es/blogs-es/coordenadas-bdo/riesgos-de-sesgo-y-discriminacion-en-inteligencia-artificial-%28ia%29

Centro de Estudios Internacionales de Barcelona (CIDOB). (2024). *La regulación europea de la IA ante los sesgos y riesgos de discriminación*. https://www.cidob.org/publicaciones/la-regulacion-europea-de-la-ia-ante-los-sesgos-y-riesgos-de-discriminacion

RGPD - https://europa.eu/youreurope/business/dealing-with-customers/data-protection/data-protection-gdpr/index_es.htm

LOPD-RGPD - https://protecciondatos-lopd.com/empresas/cumplimiento-lopd/

IA ACT - https://maldita.es/malditatecnologia/20240916/ley-inteligencia-artificial-union-europea/

https://artificialintelligenceact.eu/es/ai-act-explorer/