Enpath

INFORME · INVESTIGACIÓN

Riesgos de la IA en el sector público: lo que enseñan las fallas reales

Junio de 2024

Organización independiente dedicada a promover un uso responsable de la inteligencia artificial en las decisiones públicas y organizacionales. Cada publicación incluye su metodología, sus fuentes y sus limitaciones.

Enpath · Investigación · Investigación · Pilar: Gobernanza de la IA **Publicado:** junio de 2024 · **Corte de información:** 30 de junio de 2024 · **Idioma:** español

> La mejor literatura sobre riesgos de la inteligencia artificial pública no está en los manuales: está en los expedientes. Catorce fallas documentadas —de COMPAS a Robodebt, de Salta a los chatbots municipales— permiten construir una taxonomía de cuatro familias de riesgo y seis lecciones que ninguna entidad debería aprender de nuevo por experiencia propia.

Actualización.

Actualizado el: 18/07/2026. Esta sección resume acontecimientos posteriores a la fecha original de publicación y no forma parte del análisis histórico original.

- La Corte Constitucional de Colombia (T-323 de 2024, agosto de 2024) fijó criterios sobre el uso de IA generativa en la función judicial, directamente relevantes para la familia de riesgos de proceso descrita en este informe.
- El Reglamento de IA europeo entró en vigor (agosto de 2024); su lista de prácticas prohibidas y su categoría de alto riesgo codifican, en lo esencial, el catálogo de fallas aquí documentado.
- El monitor de incidentes de IA de la OCDE y las bases públicas de incidentes continuaron creciendo, confirmando la utilidad del reporte sistemático de fallas que este informe recomienda.

Resumen ejecutivo.

Este informe cataloga catorce fallas documentadas de sistemas automatizados y de inteligencia artificial en el sector público —con fuentes oficiales, judiciales o de investigación verificable— y las convierte en dos herramientas: una **taxonomía de cuatro familias de riesgo** (resultado, proceso, institucional, social) y **seis lecciones operativas** para entidades de Colombia y América Latina.

El hallazgo transversal incomoda a los dos bandos del debate. A los entusiastas: **ninguna de las fallas examinadas fue un accidente técnico imprevisible**; todas materializaron riesgos conocidos, señalados a tiempo por funcionarios, auditores, académicos o afectados que no fueron escuchados. A los escépticos: **ninguna falla prueba que la tecnología sea inutilizable**; prueban que la variable decisiva es institucional. El mismo tipo de sistema que en un contexto produjo un escándalo nacional opera en otro con salvaguardas y resultados aceptables.

Para América Latina el catálogo tiene un sesgo instructivo: los casos regionales documentados —la predicción de embarazo adolescente en Salta, el sistema Alerta Niñez en Chile, el reconocimiento facial suspendido en Buenos Aires— comparten un patrón: **sistemas de predicción o vigilancia aplicados**

sobre poblaciones vulnerables, con salvaguardas débiles y evaluación tardía. La región no necesita imaginar sus riesgos; ya los tiene documentados.

Hallazgos principales.

1. **Las fallas se concentran en dos familias: resultado y proceso.** En 12 de los 14 casos el sistema produjo decisiones equivocadas o sesgadas (resultado); en 13, la organización no podía detectar, explicar o corregir el error a tiempo (proceso). La combinación —error a escala más incapacidad de verlo— es la firma de los grandes desastres.
2. **El riesgo institucional es el menos visible y el más colombiano.** La dependencia de proveedores sin salvaguardas —ilustrada por el incidente de IFX Networks (2023), que interrumpió servicios de decenas de entidades públicas— rara vez aparece en los marcos éticos y explica una parte creciente de los incidentes reales.
3. **Los sistemas de "predicción social" son la categoría de mayor riesgo específico para la región.** Salta (2018) y Alerta Niñez (2021) muestran la combinación regional característica: datos de menores y poblaciones vulnerables, metodologías sin validación pública y promesas preventivas que no resisten escrutinio (LIAA–UBA, 2018; Derechos Digitales, 2021).
4. **La IA generativa añadió una familia de fallas de verificación, no de intención.** Las citas judiciales inventadas (Mata v. Avianca, 2023) y el chatbot municipal que aconsejaba incumplir la ley (Nueva York, 2024) fallaron por el mismo mecanismo: salidas fluidas tratadas como fuentes confiables sin verificación institucional (SDNY, 2023; The Markup, 2024).
5. **Todas las alertas tempranas existieron y fueron desestimadas.** En Robodebt, abogados internos advirtieron la ilegalidad desde el inicio; en Toeslagenaffaire, señales internas y externas se acumularon durante años; en Horizon, los afectados denunciaron el sistema durante dos décadas. El costo no estuvo en la falta de información sino en la arquitectura que la ignoró.

Por qué importa.

La decisión que este informe apoya es previa a cualquier proyecto: **¿qué riesgos debe evaluar una entidad antes de adoptar un sistema, y qué salvaguardas convierten un riesgo inaceptable en uno gestionable?** La respuesta con evidencia evita los dos errores simétricos de la política tecnológica regional: adoptar sin salvaguardas (y pagar en derechos y en confianza) o prohibir por miedo (y pagar en servicios que no mejoran). Cada caso de este catálogo costó a su jurisdicción más que todo lo que su prevención habría costado.

Antecedentes.

El estudio sistemático de incidentes es reciente. La base de datos de incidentes de IA (AI Incident Database) acumula reportes desde 2020; la OCDE lanzó en 2023 su monitor de incidentes (AIM) para seguimiento sistemático de eventos reportados en medios (OECD.AI); y los reguladores comienzan a exigir reporte de incidentes graves —el Reglamento de IA europeo, aprobado en mayo de 2024, lo

incluye para sistemas de alto riesgo—. Este informe aplica esa lógica al subconjunto que importa a nuestra región: fallas en el sector público con documentación verificable, seleccionadas por relevancia y calidad de fuente, no por notoriedad.

La taxonomía.

Figura 1. Taxonomía Enpath de riesgos de la IA en el sector público. Fuente: Enpath.

- **Riesgos de resultado.** El sistema decide mal: errores de exactitud, sesgos contra grupos protegidos, degradación silenciosa del desempeño cuando la realidad cambia (deriva).
- **Riesgos de proceso.** Nadie puede verificar ni corregir: opacidad, ausencia de impugnación efectiva, supervisión nominal, y la variante letal —la *ilegalidad automatizada*: una regla jurídicamente inválida ejecutada a escala perfecta (Robodebt es el arquetipo).
- **Riesgos institucionales.** La entidad pierde control de su función: dependencia del proveedor, atrofia de la capacidad interna, discontinuidad operativa, costos de operación no presupuestados.
- **Riesgos sociales.** El daño excede los casos individuales: vigilancia con efecto inhibitorio, exclusión de quienes no encajan en los datos, y el activo que más tarda en reconstruirse: la confianza institucional.

Situación actual.

El catálogo.

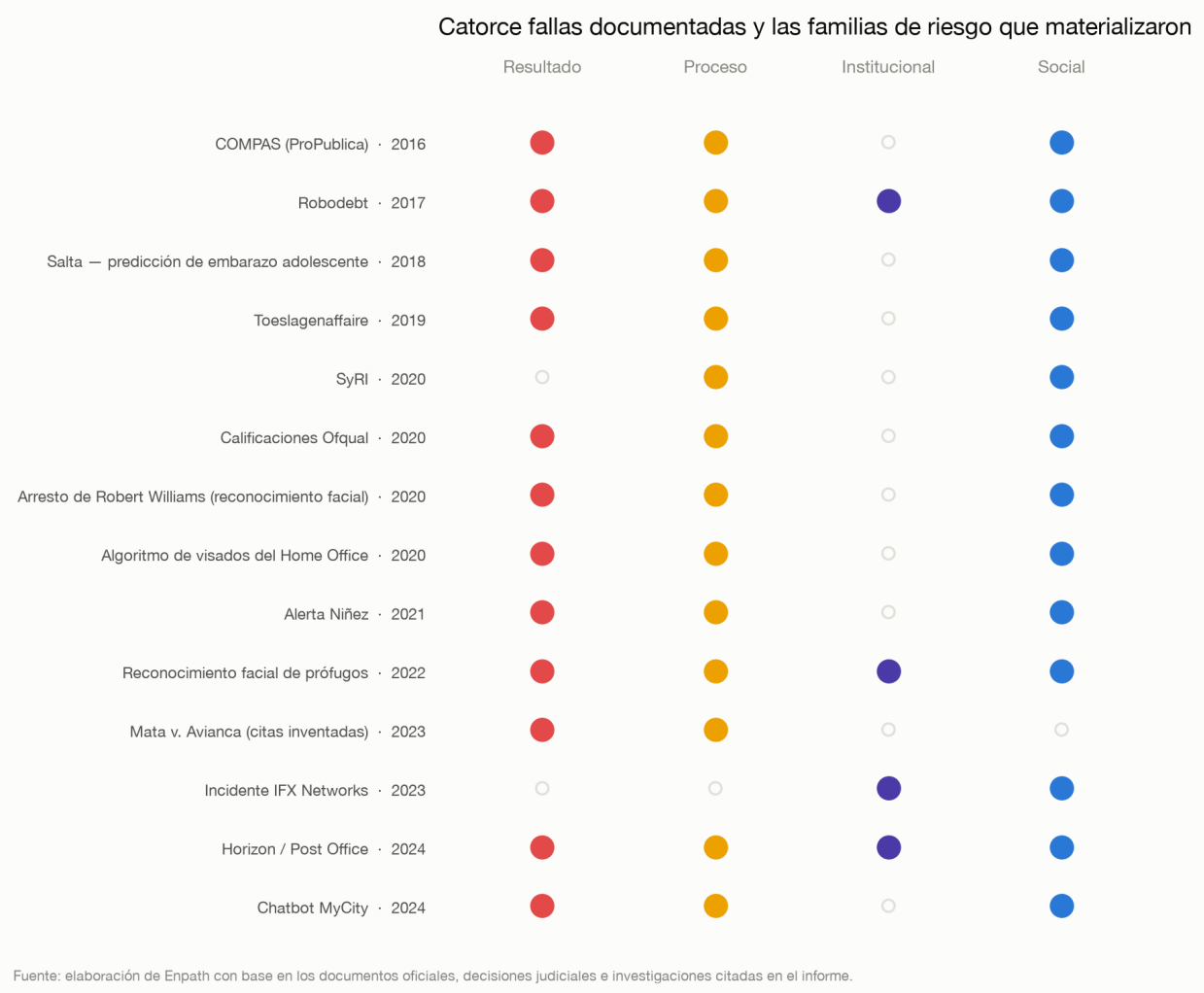

Figura 2. Catorce fallas documentadas (2016–2024) y las familias de riesgo que materializaron. Fuente: elaboración de Enpath; detalle y fuentes por caso en la tabla 1 y el archivo de datos adjunto.

Caso	Año	Qué falló	Fuente principal
COMPAS	2016	Disparidades raciales en falsos positivos de reincidencia	ProPublica (2016)
Robodebt	2017–2023	Cálculo de deudas ilegal ejecutado a escala	Comisión Real (2023)
Salta — predicción de embarazo	2018	Metodología inválida sobre datos de menores	LIAA–UBA (2018)
Toeslagenaffaire	2019–2021	Perfilamiento discriminatorio en subsidios	Parlamento de Países Bajos (2020)
SyRI	2020	Perfilamiento de barrios pobres contrario al CEDH	Rechtbank Den Haag (2020)

Caso	Año	Qué falló	Fuente principal
Ofqual	2020	39,1 % de calificaciones rebajadas por diseño del algoritmo	Ofqual (2020)
Arresto de R. Williams	2020	Falsa coincidencia de reconocimiento facial	ACLU (2020)
Visados del Home Office	2020	Clasificación por nacionalidad; retirado ante litigio	JCWI y Foxglove (2020)
Alerta Niñez	2021	Predicción de riesgo en niñez cuestionada por proporcionalidad	Derechos Digitales (2021)
Reconocimiento facial CABA	2022	Uso irregular de datos; suspensión judicial	ADC (2022)
Mata v. Avianca	2023	Jurisprudencia inventada por IA generativa en litigio	SDNY (2023)
Incidente IFX	2023	Interrupción de servicios públicos vía proveedor	Rama Judicial de Colombia (2023)
Horizon / Post Office	1999–2024	Presunción de fiabilidad de un sistema defectuoso	Gobierno del Reino Unido (2024)
Chatbot MyCity	2024	Asesoría municipal contraria a la ley	The Markup (2024)

Tabla 1. El catálogo con fuentes. Los detalles de codificación están en `data/casos_documentados.csv`. *Fuente: Enpath.*

Análisis de datos.

Tres patrones emergen de la codificación:

Primero: la falla típica es compuesta. Once de los catorce casos materializaron tres o más familias simultáneamente. El error técnico (resultado) casi nunca basta para el desastre; se requiere además la ceguera organizacional (proceso). Esto es una buena noticia operativa: las salvaguardas de proceso —impugnación, supervisión real, auditoría— protegen incluso cuando el sistema falla.

Segundo: el tiempo de detección es el multiplicador del daño. Ofqual se corrigió en una semana: el daño fue serio pero acotado. Horizon tardó dos décadas: el daño incluyó prisión, quiebras y suicidios documentados por la investigación pública británica. La variable no fue la gravedad del error técnico —el de Ofqual era mayor— sino la existencia de vías para verlo y revertirlo.

Tercero: los casos regionales fallan antes: en la pregunta. Salta y Alerta Niñez no son historias de modelos mal calibrados sino de proyectos que no debían existir en su forma propuesta: predicción individual de fenómenos sociales complejos, sobre datos de menores, sin evidencia de que la predicción mejorara la intervención. La primera línea de defensa regional no es la auditoría de modelos; es la evaluación de proporcionalidad del informe de enero de esta serie.

Colombia.

Colombia no aparece en el catálogo con un escándalo algorítmico propio de gran escala — y ese es exactamente el punto de este informe: **la ventana para prevenir sigue abierta, y los ingredientes ya están presentes.** El país tiene sistemas de decisión asistida en operación (asociación de casos penales, priorización de tutelas, asistentes tributarios), proyectos de predicción social en discusión, y el precedente institucional del incidente IFX: dependencia crítica de terceros sin salvaguardas equivalentes al riesgo.

Tres implicaciones directas para entidades colombianas:

- **El inventario es la primera salvaguarda.** Ninguna de las seis preguntas de control del informe de mayo puede responderse sobre sistemas que la entidad no sabe que tiene. La experiencia comparada muestra que los desastres crecen en sistemas que nadie listaba.
- **La regla de contradicción evita el patrón Horizon.** Definir por procedimiento qué ocurre cuando el ciudadano contradice al sistema —quién investiga, con qué carga de la prueba— es la protección más barata contra el riesgo de proceso.
- **Los proyectos de predicción social exigen el estándar máximo.** Cualquier sistema que pretenda predecir conductas o riesgos individuales de poblaciones vulnerables debe pasar la prueba de proporcionalidad con evidencia publicada, o no pasar.

América Latina.

El patrón regional tiene dos caras. La primera: los casos documentados (Salta, Alerta Niñez, reconocimiento facial en Buenos Aires) pertenecen a las categorías de mayor riesgo intrínseco — predicción social y biometría— y fueron corregidos por actores externos (académicos, sociedad civil, jueces), no por los controles internos de las entidades. La segunda: la corrección funcionó. Los tres casos fueron detenidos, suspendidos o abandonados; el litigio estratégico y la investigación académica regional demostraron capacidad real de contención.

La lección institucional es doble: fortalecer los controles internos para no depender del rescate externo, y proteger el espacio de quienes ejercen ese control externo — la transparencia activa (registros de sistemas, documentación pública) es la condición que lo hace posible. Los informes posteriores de esta serie sobre registros de algoritmos desarrollan el instrumento.

Comparación internacional.

La respuesta madura a las fallas no es la prohibición general sino la infraestructura de aprendizaje:

- **Reporte de incidentes.** El Reglamento de IA europeo exige notificar incidentes graves de sistemas de alto riesgo; la OCDE mantiene el monitor AIM; la base de incidentes de la comunidad técnica acumula el registro histórico. Aprender de fallas dejó de ser voluntario en las jurisdicciones líderes (OECD.AI).

- **Registros públicos de algoritmos.** Ámsterdam y Helsinki (2020) iniciaron los registros municipales; el Reino Unido estandarizó el reporte de transparencia algorítmica para su gobierno; los Países Bajos, tras sus escándalos, construyeron el registro nacional. La transparencia previa reduce el costo de la corrección posterior.
- **Investigaciones públicas ejemplares.** La Comisión Real australiana y la investigación británica de Horizon fijaron el estándar: nombres, causas, recomendaciones vinculantes. La región carece de un equivalente; sus fallas se cierran sin informe.

Oportunidades.

- **Prevenir es más barato que cualquier alternativa, y ya hay catálogo.** Una entidad que evalúe sus proyectos contra las cuatro familias y los catorce casos evita los errores más caros del repertorio conocido.
- **El "pre-mortem" institucional es de adopción inmediata.** Antes de cada despliegue: "supongamos que en dos años este sistema es nuestro Robodebt; ¿qué habrá fallado?" La técnica es gratuita y obliga a nombrar los riesgos incómodos.
- **El reporte de incidentes puede nacer regional.** Un mecanismo latinoamericano de reporte de incidentes de IA pública —anclado en CEPAL, la OEA o la red de autoridades de datos— convertiría las fallas de cada país en aprendizaje de todos.
- **La ventaja del segundo: la región puede saltarse los errores ya cometidos.** El costo de Robodebt, Toeslagen y Horizon ya fue pagado por otros; sus informes finales son bienes públicos gratuitos.

Riesgos.

- **El riesgo de leer este catálogo como argumento contra la IA.** No lo es: es el argumento contra la IA sin salvaguardas. La parálisis también tiene víctimas — los servicios que no mejoran, los fraudes que no se detectan, el tiempo de los funcionarios consumido en tareas automatizables.
- **La falla regional más probable no será exótica.** Será un sistema de focalización o detección de fraude, comprado a un tercero, sin inventario ni supervisión medida, fallando en silencio sobre población vulnerable durante años. Todo en este informe existe para evitar exactamente eso.
- **Normalización de la opacidad generativa.** La adopción acelerada de chatbots públicos sin verificación institucional replica el patrón MyCity a bajo costo de entrada y alto costo de confianza.
- **Aprendizaje sin memoria.** Sin registro de incidentes ni investigaciones públicas, cada gobierno nuevo repite los pilotos del anterior. La memoria institucional es una decisión, no un accidente.

Perspectivas.

Con corte a junio de 2024: la entrada en vigor del Reglamento de IA europeo (esperada en agosto) activará el primer régimen obligatorio de reporte de incidentes graves; la decisión pendiente de la Corte Constitucional colombiana sobre IA generativa en la justicia fijará doctrina regional; y la expansión de la IA generativa en atención ciudadana anticipa que la próxima cosecha de incidentes será conversacional. La predicción falsable de este informe: los países que instalen reporte de incidentes y registros de sistemas antes de 2026 tendrán fallas menores y públicas; los demás, fallas mayores y descubiertas tarde.

Recomendaciones.

1. **Evalúe cada proyecto contra las cuatro familias antes de aprobarlo**, con la pregunta de proporcionalidad primero: ¿debe existir este sistema, en esta forma, para este problema?
2. **Instituya el pre-mortem y documéntelo.** Los riesgos nombrados antes del despliegue son los únicos que generan salvaguardas presupuestadas.
3. **Cree el registro interno de incidentes y casi-incidentes** de sistemas automatizados, con revisión periódica del comité de gobernanza de datos (informe de mayo). Lo que no se registra no enseña.
4. **Aplique el estándar máximo a predicción social y biometría.** Para estas dos categorías: evidencia publicada de validez, evaluación externa previa, y prohibición de despliegue sobre menores sin autorización legal expresa.
5. **Gobierne el riesgo de terceros como riesgo propio.** Continuidad, respaldo, auditoría y salida: el incidente IFX define el mínimo contractual colombiano.
6. **Publique lo que aprende.** Un informe público breve por cada incidente cerrado —qué pasó, por qué, qué cambió— convierte el costo pagado en capital institucional. Es la práctica que separa a las administraciones que maduran de las que reinciden.

Metodología.

Catálogo construido por revisión de fuentes con corte al 30 de junio de 2024. Criterios de inclusión: (i) sistema automatizado o de IA usado por o para una función pública; (ii) falla o cuestionamiento documentado en fuente oficial, judicial, académica o de investigación periodística verificable; (iii) relevancia para decisiones de entidades latinoamericanas. Cada caso se codificó contra las cuatro familias de la taxonomía (materializada = 1) con base en las fuentes citadas; la codificación completa está en `data/casos_documentados.csv` y es discutible caso a caso — esa discusión es bienvenida. El catálogo no es exhaustivo ni una muestra estadística: es un conjunto de casos ejemplares con documentación sólida.

Limitaciones. El sesgo de documentación favorece a jurisdicciones con prensa, academia y justicia activas; la ausencia de casos documentados en un país no implica ausencia de fallas. La codificación binaria simplifica gradientes de daño muy distintos entre casos.

Referencias.

1. ACLU (2020). *Wrongfully Arrested Because Face Recognition Can't Tell Black People Apart*. https://www.aclu.org/news/privacy-technology/wrongfully-arrested-because-face-recognition-cant-tell-black-people-apart
2. ADC — Asociación por los Derechos Civiles (2022). *Documentación sobre la suspensión judicial del Sistema de Reconocimiento Facial de Prófugos (CABA)*. https://adc.org.ar/
3. Angwin, J., Larson, J., Mattu, S. y Kirchner, L. (2016). *Machine Bias*. ProPublica, 23 de mayo de 2016. https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing
4. Derechos Digitales (2021). *Análisis sobre el sistema Alerta Niñez (Chile)*. https://www.derechosdigitales.org/
5. Foxglove y JCWI (2020). *Home Office says it will abandon its racist visa algorithm*. Agosto de 2020. https://www.foxglove.org.uk/2020/08/04/home-office-says-it-will-abandon-its-racist-visa-algorithm-after-we-sued-them/
6. Gobierno del Reino Unido (2024). *New law to quash convictions of postmasters in Horizon scandal*. https://www.gov.uk/government/news/new-law-to-quash-convictions-of-postmasters-in-horizon-scandal
7. LIAA — Laboratorio de Inteligencia Artificial Aplicada, UBA (2018). *Sobre la predicción automática de embarazos adolescentes*. https://liaa.dc.uba.ar/
8. OECD.AI. *AI Incidents Monitor (AIM)*. https://oecd.ai/en/incidents
9. Ofqual (2020). *Summer 2020 grades for GCSE, AS and A level*. https://www.gov.uk/government/news/summer-2020-grades-for-gcse-as-and-a-level
10. Rechtbank Den Haag (2020). *Sentencia SyRI* (ECLI:NL:RBDHA:2020:865). https://uitspraken.rechtspraak.nl/inziendocument?id=ECLI:NL:RBDHA:2020:865
11. Royal Commission into the Robodebt Scheme (2023). *Report*. https://robodebt.royalcommission.gov.au/publications/report
12. Tweede Kamer (2020). *Ongekend onrecht*. https://www.tweedekamer.nl/kamerstukken/detail?id=2020D50046
13. U.S. District Court, SDNY (2023). *Mata v. Avianca, Inc.: Opinion and Order on Sanctions*. Junio de 2023. https://law.justia.com/cases/federal/district-courts/new-york/nysdce/1:2022cv01461/575368/54/
14. The Markup (2024). *NYC's AI Chatbot Tells Businesses to Break the Law*. 29 de marzo de 2024. https://themarkup.org/news/2024/03/29/nycs-ai-chatbot-tells-businesses-to-break-the-law

---

Enpath · Investigación aplicada para decisiones públicas basadas en evidencia. Cada publicación incluye su metodología, sus fuentes y sus limitaciones.