

Enpath

NOTA TÉCNICA · INVESTIGACIÓN

Diseño de sistemas públicos con supervisión humana: guía práctica de implementación

Septiembre 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.

"Habrà supervisi3n humana" es una frase de política. Este documento es lo que sigue: cómo se diseña esa supervisi3n en el flujo de trabajo, en la pantalla del revisor, en la planta de personal y en las métricas. Escrito para los equipos que implementan, no para los comités que declaran.

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.

- El Consejo Superior de la Judicatura de Colombia expidió en 2025 lineamientos para el uso de IA en la Rama Judicial, en cumplimiento del exhorto de la sentencia T-323 de 2024 analizada aquí.
- Las obligaciones de supervisi3n humana del Reglamento de IA europeo (art. 14) entraron en aplicaci3n para sistemas de alto riesgo en agosto de 2026, con orientaciones técnicas de la Oficina de IA europea que siguen el enfoque de diseño descrito en este documento.

Resumen ejecutivo.

Entre febrero y septiembre de 2024, la supervisi3n humana pasó en el mundo —y en Colombia— de principio a obligaci3n. El Reglamento de IA europeo entró en vigor con su artículo 14: los sistemas de alto riesgo deben diseñarse para que personas puedan "supervisarlos de manera efectiva", entender sus límites, interpretar sus salidas e intervenir o detenerlos. Y la Corte Constitucional colombiana, en la sentencia T-323 de 2024 (agosto), fijó la primera regla nacional: la inteligencia artificial puede apoyar tareas judiciales, pero no sustituir el razonamiento del juez, y su uso exige transparencia, responsabilidad y protecci3n de datos ([Corte Constitucional, 2024](#)).

Ninguna de las dos normas dice

cómo

. Este documento llena ese vacío con cinco componentes de diseño verificables:

1. **El flujo** — dónde entra la persona: tres patrones (revisi3n previa universal, cola de excepciones, muestreo posterior) asignados según la matriz de delegaci3n del informe de febrero.
2. **La interfaz** — qué ve el revisor: el caso completo, la recomendaci3n con su incertidumbre, la explicaci3n, los límites conocidos del sistema y acciones equivalentes sin opci3n preseleccionada. La evidencia experimental es inequívoca: la pantalla determina si la persona piensa o firma.

3. **La dotación** — cuántos revisores: la supervisión se presupuesta con aritmética simple (casos por día × fracción revisada × minutos por caso), y quien no puede pagar la dotación del patrón elegido debe cambiar de patrón o de proyecto, no simular la revisión.
4. **El registro** — qué queda documentado: cada intervención (acepta, modifica, rechaza, deriva) con su razón; de ahí salen las dos métricas de gestión (tasas de aceptación y derivación).
5. **El circuito de mejora** — cómo aprende el conjunto: revisión periódica de métricas, retroalimentación al modelo, y autoridad clara para suspender el sistema cuando los umbrales se cruzan.

El documento cierra con la **plantilla Enpath de protocolo de supervisión** —una página que cualquier entidad puede adoptar por acto administrativo— y con los anti-patronos que la evidencia identifica como las formas más comunes de simular supervisión.

Hallazgos principales.

1. **La regla colombiana ya existe; falta la ingeniería.** T-323 de 2024 prohíbe sustituir el razonamiento judicial y exige transparencia y responsabilidad en el uso de IA generativa; el exhorto a la Rama Judicial para expedir lineamientos convierte este diseño en tarea inmediata de implementación — no solo judicial: el estándar migrará al resto del Estado.
2. **La interfaz es política pública.** Los experimentos disponibles muestran que pequeñas decisiones de pantalla —mostrar o no la recomendación primero, exigir o no una razón para disentir, preseleccionar o no la aceptación— cambian las tasas de error y de dependencia del sistema más que la calidad del propio modelo ([Green y Chen, 2019](#); [Buçinca et al., 2021](#)).
3. **La supervisión sin presupuesto es la principal fuente de supervisión nominal.** Un sistema que procesa mil casos diarios con revisión universal de quince minutos exige más de cuarenta revisores de tiempo completo. Las entidades que declaran revisión universal sin esa planta están describiendo una ficción; la aritmética de dotación lo hace visible antes del despliegue.
4. **Los tres patrones cubren todo el espacio legítimo de diseño.** Revisión previa universal (alto impacto), cola de excepciones con umbrales (impacto medio, volumen alto) y muestreo posterior (bajo impacto, error reversible). Cualquier flujo que no encaje en ninguno es, en la práctica, decisión automatizada sin supervisión — y debe tratarse como tal.
5. **El registro convierte la supervisión en evidencia.** Sin bitácora de intervenciones no hay métricas; sin métricas, ni la entidad ni los órganos de control pueden distinguir supervisión real de teatro. La bitácora es, además, la defensa jurídica de la entidad y del revisor.

Por qué importa.

La decisión que este documento apoya es de los equipos de implementación: **cómo convertir la obligación de supervisión —constitucional, contractual o de política interna— en un diseño operable y demostrable.** El costo de no resolverla bien ya está documentado en esta serie: supervisión teatral que agrava los errores del sistema (febrero), fallas compuestas de resultado y proceso (junio) y

contratos que no exigen al proveedor la información que el revisor necesita (agosto). Este es el eslabón donde esas piezas se ensamblan o se rompen.

Antecedentes.

Dos hechos de 2024 definen el contexto. Primero, el artículo 14 del Reglamento de IA europeo —en vigor desde el 1 de agosto— convierte el diseño de supervisión en requisito de producto: los proveedores de sistemas de alto riesgo deberán entregarlos con capacidades de supervisión incorporadas, y las entidades usuarias asignarán personas competentes para ejercerla ([Unión Europea, 2024](#)). Segundo, la sentencia T-323 de 2024: a propósito del juez que usó ChatGPT en una tutela (el caso de Cartagena analizado en febrero), la Corte Constitucional colombiana estableció que la IA no puede reemplazar el razonamiento judicial, admitió usos de apoyo bajo criterios de transparencia, responsabilidad, privacidad y no sustitución, y exhortó a la Rama Judicial a expedir lineamientos ([Corte Constitucional, 2024](#)). Colombia tiene así, por vía jurisprudencial, lo que pocos países de la región tienen por ley: un estándar exigible de supervisión — sin manual de implementación. La evidencia experimental sobre sesgo de automatización (resumida en febrero) aporta el tercer insumo: el comportamiento del revisor es una función del diseño, no de su buena voluntad.

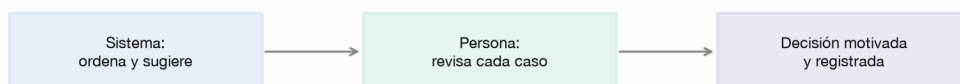
Los cinco componentes.

1 • El flujo.

Tres patrones de supervisión y cuándo usar cada uno

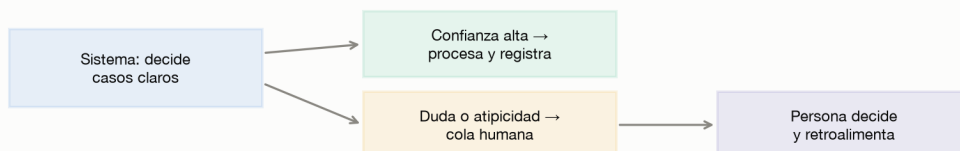
Patrón A • Revisión previa universal

Para decisiones de alto impacto (matriz: decisión humana plena o con apoyo)



Patrón B • Cola de excepciones

Automatización con umbrales: lo dudoso y lo atípico va a una persona



Patrón C • Muestreo posterior auditado

Solo para bajo impacto y error reversible; nunca si hay derechos en juego

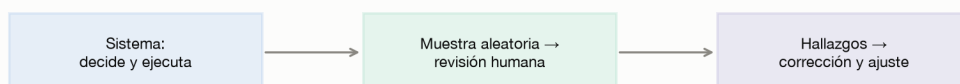


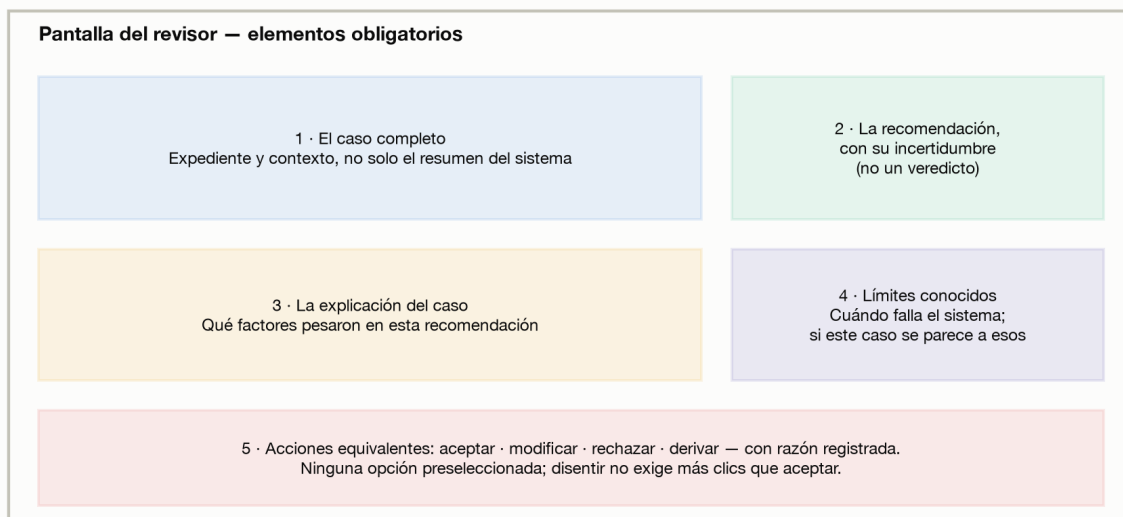
Figura 1. Tres patrones de supervisión y su ámbito de uso. Fuente: Enpath.

La elección del patrón no es discrecional: la fija la clasificación del sistema en la matriz de delegación (impacto sobre derechos × reversibilidad del error, informe de febrero). Dos reglas completan el diseño

del flujo. **Regla de derivación:** todo caso que el sistema marque con baja confianza, todo caso atípico y todo caso donde el ciudadano contradiga los datos van a la cola humana — la tasa de derivación esperada se pacta desde el diseño (si es cero, el flujo está mal diseñado). **Regla de contradicción:** el procedimiento define quién investiga y con qué carga de la prueba cuando ciudadano y sistema afirman cosas incompatibles.

2 · La interfaz.

La interfaz decide la calidad de la supervisión



Fuente: Enpath, con base en la evidencia sobre sesgo de automatización (Skitka et al., 1999; Green y Chen, 2019; Bućinca et al., 2021).

Figura 2. Elementos obligatorios de la pantalla del revisor. Fuente: Enpath.

Cinco elementos, cada uno anclado en evidencia: el **caso completo** (revisar solo el resumen del sistema es revisar al sistema, no al caso); la **recomendación con incertidumbre** (un porcentaje calibrado invita a ponderar; un veredicto invita a firmar); la **explicación del caso concreto** (qué factores pesaron); los **límites conocidos** (si este caso se parece a aquellos donde el sistema falla, el revisor debe saberlo); y **acciones equivalentes** — disentir no puede costar más clics, más formularios ni más tiempo que aceptar. Donde el riesgo lo justifica, el diseño puede añadir *fricción deliberada*: pedir al revisor una valoración propia antes de mostrar la recomendación, la técnica con mejor evidencia contra la aceptación refleja (Bućinca et al., 2021).

3 · La dotación.

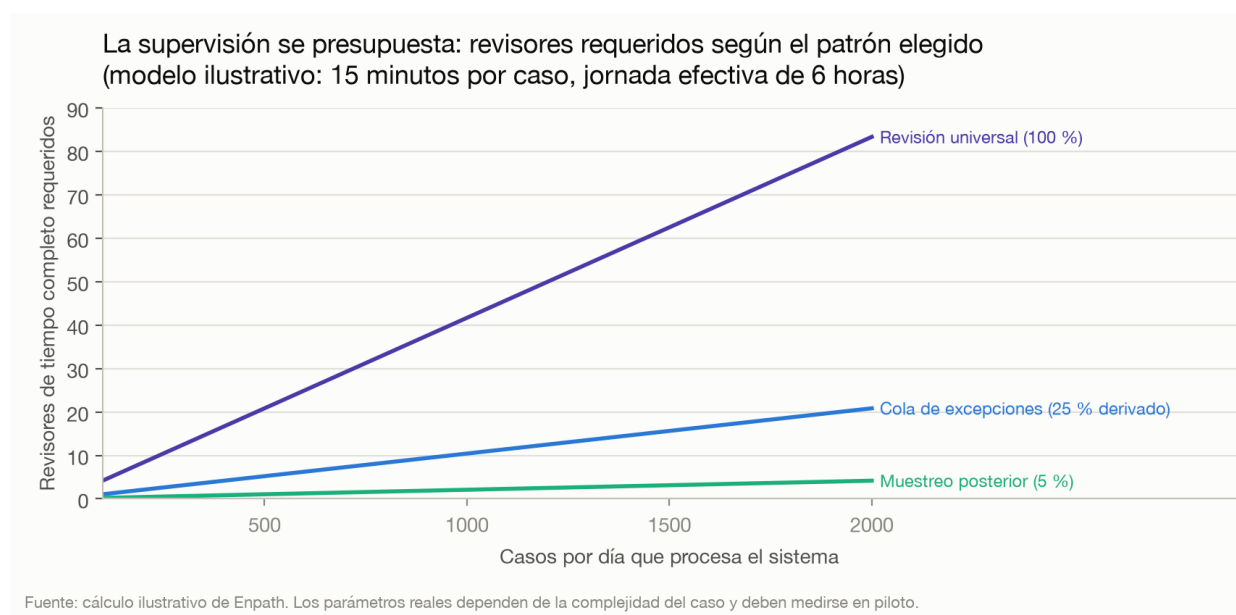


Figura 3. Revisores requeridos según patrón (modelo ilustrativo). Fuente: Enpath.

La fórmula es elemental — $\text{revisores} = (\text{casos/día} \times \text{fracción revisada} \times \text{minutos/caso}) / \text{minutos efectivos de jornada}$ — y sus consecuencias no lo son: **el patrón de supervisión es una decisión presupuesta**. Tres corolarios: los minutos por caso se miden en piloto, no se estipulan por deseo; la jornada efectiva de revisión sostenible es menor que la laboral (la fatiga de revisión degrada el juicio); y si la dotación resultante es impagable, la respuesta honesta es reducir el alcance del sistema o cambiar de patrón — nunca recortar los minutos por caso hasta que la cuenta cierre.

4 · El registro.

Cada intervención queda en bitácora: identificador del caso, revisor, acción (acepta / modifica / rechaza / deriva), razón breve y tiempo empleado. Del registro salen las métricas de gestión — **tasa de aceptación** (cercana al 100 % sostenido = supervisión nominal), **tasa de derivación** (cero sostenido = casos atípicos forzados dentro del sistema), distribución de tiempos (colapso hacia segundos = revisión imposible) — y la evidencia exigible por órganos de control. El registro protege también al revisor: documenta que ejerció juicio, no que estampó firmas.

5 · El circuito de mejora.

La supervisión produce el insumo más valioso del sistema: los desacuerdos entre revisor y modelo. El circuito mínimo trimestral: revisión de métricas por el comité de gobernanza (informe de mayo); análisis de desacuerdos (¿quién tenía razón? — la respuesta calibra tanto al modelo como a los revisores); ajuste de umbrales de derivación; y decisión documentada de continuar, ajustar o suspender. La autoridad de suspensión —quién detiene el sistema y con qué criterios— se designa por escrito antes del despliegue.

Situación actual.

En Colombia, ningún sistema del inventario de julio publica métricas de supervisión, y la única especificación pública de diseño de revisión es la de PretorlA (funciones informativas, decisión en los despachos). La sentencia T-323 cambia el tablero: la Rama Judicial deberá escribir lineamientos, y las entidades del Ejecutivo que usan sistemas comparables quedan notificadas del estándar que la jurisdicción constitucional aplicará cuando lleguen los casos. En la región, el patrón es idéntico: principios de supervisión en todos los marcos, protocolos operativos en ninguno. La ventana práctica de este documento es exactamente esa brecha.

Análisis de datos.

La aritmética de dotación merece tratarse como dato y no como detalle: es la variable que separa los diseños viables de las ficciones. Tres escenarios ilustrativos para un sistema de mil casos diarios (15 minutos por caso, jornada efectiva de 6 horas): revisión universal, ~42 revisores; cola de excepciones al 25 %, ~10; muestreo al 5 %, ~2. La lectura correcta no es "el muestreo es más barato": es que **cada patrón compra un nivel de protección distinto, y el nivel de protección lo fija el impacto del sistema, no el presupuesto disponible**. Cuando el presupuesto no alcanza para el patrón que el impacto exige, el proyecto está sobredimensionado — ese es el hallazgo, y llega mejor antes del contrato que después del escándalo. Los parámetros reales (minutos por caso, fracción derivable) se miden en piloto con la bitácora activada; el modelo aquí presentado es deliberadamente simple y sus supuestos están publicados para ser discutidos.

Colombia.

Implicaciones inmediatas de T-323 para los equipos colombianos, en el lenguaje de los cinco componentes: los usos generativos de apoyo (resúmenes, borradores, búsqueda) son legítimos si el flujo garantiza que el razonamiento y la decisión son humanos (patrón A para todo lo decisorio); la transparencia exigida por la Corte se implementa con el registro (declarar el uso y poder mostrarlo); la responsabilidad se implementa con la autoridad de suspensión y la bitácora; y la protección de datos remite al marco del informe de mayo (ningún dato de expediente en herramientas no contratadas — la regla contractual octava de agosto). Para el Ejecutivo, la lectura estratégica: los lineamientos que la Rama Judicial expida serán el primer protocolo oficial de supervisión del Estado colombiano; las entidades que adopten esta plantilla antes llegarán a ese estándar por decisión propia y no por orden judicial.

América Latina.

El activo regional es la experiencia acumulada en los casos ya analizados: PretorlA como patrón A documentado; los sistemas de predicción social cuestionados (Salta, Alerta Niñez) como recordatorio de que ningún diseño de supervisión salva un sistema desproporcionado; y la jurisprudencia colombiana

como primer estándar constitucional específico. La oportunidad de cooperación es de bajo costo: los protocolos de supervisión son documentos cortos y traducibles entre países con tradición jurídica común; un repositorio regional de protocolos publicados (vía CLAD, CEPAL o la red de escuelas judiciales) evitaría que cada país redacte el suyo desde cero.

Comparación internacional.

- **Unión Europea:** supervisión como requisito de producto (art. 14) — el proveedor debe entregar el sistema "supervisable". Lección contractual directa: exigir las capacidades de supervisión en el pliego (cláusulas 2 y 5 del informe de agosto), porque no pueden añadirse después.
- **Estados Unidos:** supervisión como salvaguarda mínima verificable (M-24-10) con plazo de diciembre de 2024 — la primera prueba a escala de si la supervisión declarada resiste auditoría.
- **Canadá:** supervisión graduada por nivel de impacto en la Directiva de Decisiones Automatizadas — el precedente del principio que este documento operacionaliza: el patrón lo fija el impacto.
- **Estándares de gestión:** la norma ISO/IEC 42001 (2023) de sistemas de gestión de IA ofrece el andamiaje organizacional en el que estos protocolos encajan como controles operativos ([ISO, 2023](#)).

Oportunidades.

- **El protocolo como ventaja jurídica.** Tras T-323, la entidad colombiana que documente flujo, interfaz, dotación y registro tiene la mejor defensa disponible ante tutelas y órganos de control; la que no, carga con la prueba imposible de una supervisión sin evidencia.
- **Piloto con bitácora: aprendizaje casi gratuito.** Activar el registro desde el piloto produce en semanas los parámetros reales de dotación y las primeras métricas — el insumo que ningún consultor puede vender.
- **Los desacuerdos como activo.** El análisis sistemático revisor-contramodelo es la fuente de mejora más subutilizada; las entidades que lo instalen mejorarán modelo y criterio humano a la vez.
- **Estandarización regional temprana.** Quien publique primero un protocolo nacional (la Rama Judicial colombiana tiene el mandato) fijará el molde regional.

Riesgos.

- **Cumplimiento de pantalla.** Añadir un botón de "revisar" sobre un flujo diseñado para aceptar: la forma más nueva de supervisión nominal. Las métricas de la bitácora son el antídoto — por eso el anti-patrón más revelador es no activarla.
- **Métricas convertidas en metas.** Si la tasa de aceptación se vuelve objetivo ("bajar al 80 %"), los revisores disenterán por cuota. Las métricas diagnostican; los objetivos se fijan sobre resultados de calidad, no sobre las tasas mismas.

- **Fatiga y rotación de revisores.** La revisión sostenida es trabajo cognitivo exigente; sin rotación, pausas y reconocimiento, la calidad cae en silencio. La dotación incluye el relevo, no solo la nómina.
- **El estándar judicial leído como techo.** T-323 fija un piso para la función judicial; tratarlo como techo ("solo aplica a jueces") dejaría al resto del Estado en el vacío exacto que la matriz de febrero cubre.

Perspectivas.

Con corte a septiembre de 2024: los lineamientos de la Rama Judicial colombiana (exhortados por T-323) definirán el primer protocolo oficial del país; el plazo estadounidense de diciembre de 2024 mostrará cuántas agencias pueden demostrar —no declarar— sus salvaguardas; y las orientaciones europeas de implementación del artículo 14 empezarán a estandarizar el diseño de supervisión como disciplina profesional. La predicción operativa de este documento: en dos años, "¿me muestra su bitácora de supervisión?" será la primera pregunta de todo auditor competente de sistemas públicos — y las entidades que la puedan responder serán minoría distinguida.

Recomendaciones.

La plantilla Enpath de protocolo de supervisión.

Una página, siete campos, adoptable por acto administrativo:

Campo	Contenido
1 · Sistema y clasificación	Nombre, función, posición en la matriz de delegación (impacto x reversibilidad)
2 · Patrón de flujo	A, B o C; umbrales de derivación; regla de contradicción
3 · Revisores	Perfil, número (aritmética anexa), formación en límites del sistema, rotación
4 · Interfaz	Los cinco elementos obligatorios; verificación de acciones equivalentes
5 · Registro	Bitácora: campos, retención, acceso; responsable de su integridad
6 · Métricas y umbrales	Aceptación, derivación, tiempos; valores que disparan revisión o suspensión
7 · Autoridad y circuito	Quién suspende el sistema; calendario del circuito de mejora; próximo informe público

Tabla 1. Plantilla de protocolo de supervisión. Fuente: Enpath.

Y cinco reglas de implementación: **piloto siempre con bitácora activada**; **medir minutos por caso antes de fijar la planta**; **ninguna interfaz con aceptación preseleccionada**; **publicar las métricas trimestralmente** (la transparencia disciplina más que la auditoría); y **revisar el protocolo cuando cambie el modelo** — todo cambio de versión del sistema reabre el diseño de supervisión (cláusula 7 del informe de agosto).

Metodología.

Nota técnica elaborada con corte al 30 de septiembre de 2024, sobre tres insumos: la evidencia experimental revisada por pares sobre interacción humano-IA y sesgo de automatización (citada en el texto), los requisitos normativos vigentes (Reglamento de IA europeo, art. 14; sentencia T-323 de 2024; memorando OMB M-24-10; Directiva canadiense) y los marcos de esta serie (matriz de delegación y condiciones de supervisión, febrero; taxonomía de riesgos, junio; cláusulas contractuales, agosto). El modelo de dotación es ilustrativo, con parámetros explícitos, y no sustituye la medición en piloto. La caracterización de la sentencia T-323 se basa en la información oficial disponible al cierre.

Limitaciones. La evidencia experimental citada proviene mayoritariamente de contextos de laboratorio o de jurisdicciones distintas a las latinoamericanas; la transferencia de sus magnitudes (no de sus direcciones) a entidades de la región es una hipótesis de trabajo. La plantilla no ha sido validada en despliegues reales; Enpath la someterá a prueba y revisión con las entidades que la adopten.

Referencias.

1. Buçinca, Z., Malaya, M. B. y Gajos, K. Z. (2021). *To Trust or to Think: Cognitive Forcing Functions Can Reduce Overreliance on AI in AI-assisted Decision-making*. PACM HCI, 5(CSCW1). <https://doi.org/10.1145/3449287>
2. Corte Constitucional de Colombia (2024). *Sentencia T-323 de 2024* (uso de IA generativa en la función judicial). Agosto de 2024. <https://www.corteconstitucional.gov.co/>
3. Gobierno de Canadá (2019, act. 2023). *Directive on Automated Decision-Making*. <https://www.tbs-sct.canada.ca/pol/doc-eng.aspx?id=32592>
4. Green, B. y Chen, Y. (2019). *Disparate Interactions: An Algorithm-in-the-Loop Analysis of Fairness in Risk Assessments*. FAT* 2019. <https://doi.org/10.1145/3287560.3287563>
5. ISO (2023). *ISO/IEC 42001:2023 — Artificial intelligence management system*. Diciembre de 2023. <https://www.iso.org/standard/81230.html>
6. Kleinberg, J., Lakkaraju, H., Leskovec, J., Ludwig, J. y Mullainathan, S. (2018). *Human Decisions and Machine Predictions*. Quarterly Journal of Economics, 133(1). <https://doi.org/10.1093/qje/qjx032>
7. OMB (2024). *Memorandum M-24-10*. <https://www.whitehouse.gov/wp-content/uploads/2024/03/M-24-10-Advancing-Governance-Innovation-and-Risk-Management-for-Agency-Use-of-Artificial-Intelligence.pdf>
8. Skitka, L. J., Mosier, K. L. y Burdick, M. (1999). *Does automation bias decision-making?* International Journal of Human-Computer Studies, 51(5). <https://doi.org/10.1006/ijhc.1999.0252>
9. Unión Europea (2024). *Reglamento (UE) 2024/1689, artículo 14 (supervisión humana)*. <https://eur-lex.europa.eu/eli/reg/2024/1689/oj>

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