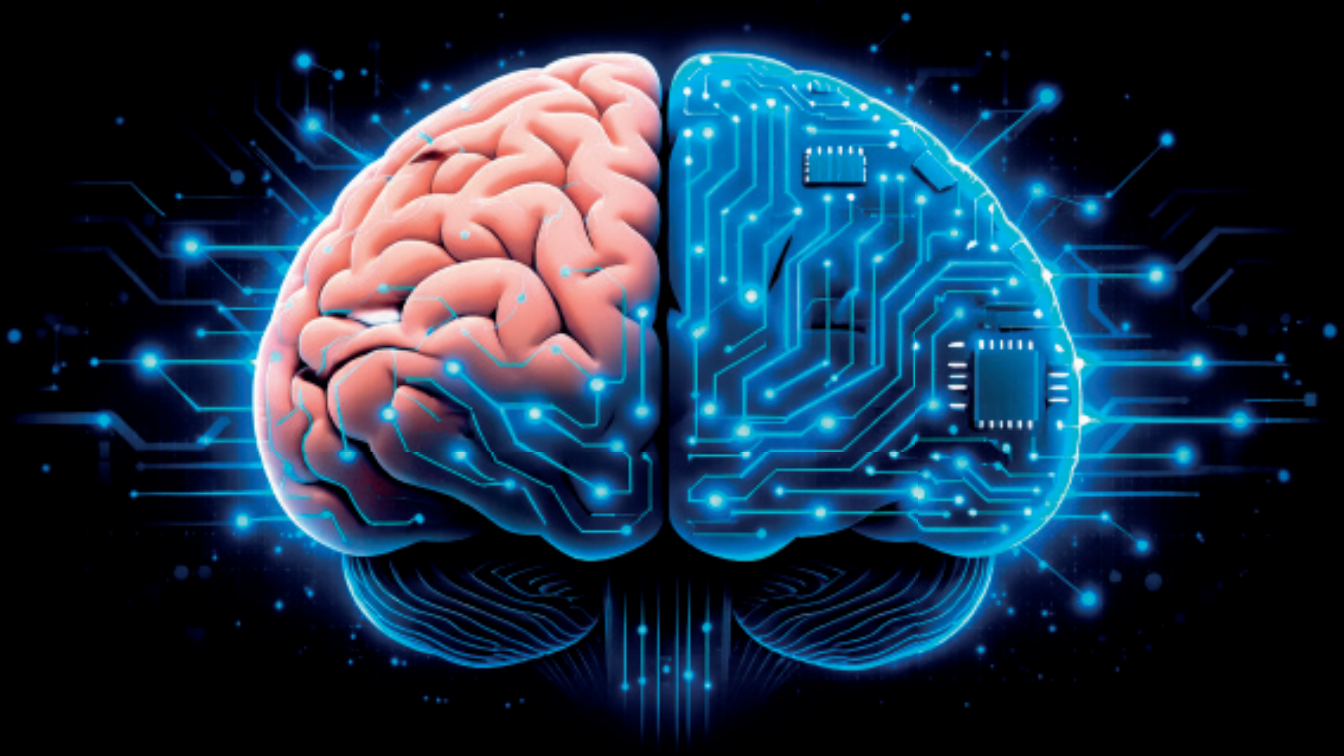
PROCURADURÍA
GENERAL DE LA NACIÓN
COLOMBIA



Principios Éticos y Marco de Gobernanza para la

INTELIGENCIA ARTIFICIAL

en el Ministerio Público de Colombia

DIÁLOGO PARA CONSTRUIR CONSENSOS

PROCURADURÍA EN LAS REGIONES

Paz Electoral

PROCURADURÍA
GENERAL DE LA NACIÓN
COLOMBIA

Principios Éticos y Marco de Gobernanza para la Inteligencia Artificial en el Ministerio Público de Colombia

Procuraduría General de la Nación
2026

Editor
Instituto de Estudios del Ministerio Público

Dirección: Carrera 5 # 15-80, Bogotá D.C., Colombia

Contáctenos:
https://www.procuraduria.gov.co

Carolina Hoyos Villamil
Directora Instituto de Estudios del Ministerio Público - IEMP

Coordinador Editorial
Gary Hernández Guerrero

Diseñador Editorial
Diego González Trujillo

Autores

Investigador principal
Joan Miguel Tejedor Estupiñán

Coinvestigadores
Carlos Mauricio Medina Fajardo
Luis Enrique Martínez Ballén

ISBN: 978-958-734-354-0

PROCURADURÍA
GENERAL DE LA NACIÓN
COLOMBIA

Gregorio Eljach Pacheco
Procurador General de la Nación

Julián Fernández
Viceprocurador General de la Nación

Carolina Hoyos Villamil
Directora Instituto de Estudios
del Ministerio Público

DIÁLOGO PARA CONSTRUIR CONSENSOS | PROCURADURÍA EN LAS REGIONES | Paz Electoral

Esta propuesta de Principios éticos y marco de gobernanza para una adopción responsable de la inteligencia artificial en el Ministerio Público de Colombia hace parte de la investigación titulada: Diseño de una estrategia integral para el uso ético de la inteligencia artificial en el Ministerio Público: fortalecimiento misional y lucha contra la desinformación. Elaborada por el Instituto de Estudios del Ministerio Público (IEMP) y tiene como referencia las "Directrices para el Uso Responsable de la Inteligencia Artificial en el Servicio Público" del Gobierno de Irlanda (Department of Public Expenditure, NDP Delivery and Reform, 2024). Los autores declaran el uso de modelos de lenguaje extenso como Gemini 3.0 y Manus 1.6 para la búsqueda de documentos, la traducción de artículos científicos y el análisis de información; todo esto bajo la respectiva supervisión humana durante todo el proceso de investigación.

Tabla de contenido

Presentación

La Procuraduría General de la Nación presenta a la Defensoría del Pueblo y a las personerías municipales y distritales, como instituciones que conforman el Ministerio Público, al Estado colombiano y a la ciudadanía en general, estos Principios éticos y marco de gobernanza para el uso responsable de la inteligencia artificial. Este documento representa un hito institucional en la transformación digital en el sector público con un enfoque centrado en las personas, los derechos humanos y el fortalecimiento del servicio a los ciudadanos.

El Ministerio Público, como órgano de control, garante del orden jurídico, custodio de los derechos humanos y promotor del servicio público eficiente, enfrenta el desafío histórico de incorporar la transformación digital en su dimensión relacionada con la inteligencia artificial, de manera que la tecnología potencie su misión constitucional sin comprometer los valores democráticos ni los principios éticos que fundamentan su razón de ser.

Dentro de las tecnologías que surgieron con la cuarta revolución industrial, la inteligencia artificial ofrece oportunidades sin precedentes para optimizar la vigilancia de la gestión pública, fortalecer la protección de derechos humanos, agilizar procesos disciplinarios, mejorar la atención ciudadana y combatir fenómenos emergentes como la desinformación. Sin embargo, su adopción también plantea riesgos que deben ser gestionados con prudencia: alucinaciones, sesgos algorítmicos, afectación a la privacidad, erosión de la rendición de cuentas y automatización de decisiones que deben preservar la dignidad humana.

Este marco para el uso ético de la IA se construye sobre cuatro pilares fundamentales:

- **Compromiso constitucional:** alineación estricta con los artículos 275-284 de la Constitución Política (1991) que definen las funciones del Ministerio Público.

- **Marco normativo robusto:** incorporación de la Ley 1581 de 2012 (protección de datos), Ley 1712 de 2014 (transparencia), Directiva Conjunta 007 de 2025 de

la Procuraduría General de la Nación y la Defensoría del Pueblo (transparencia algorítmica), Documento CONPES 3975 de 2019, Documento CONPES 4144 de 2025 (Política Nacional de IA), y demás normativa aplicable.

• **Estándares internacionales:** adopción de los Principios de IA de la OCDE, recomendaciones de UNESCO, directrices del Consejo de Europa y mejores prácticas internacionales.

• **Participación y cocreación:** construcción colaborativa con servidores públicos, expertos, sociedad civil y ciudadanía.

Este marco adapta y expande significativamente las "Directrices para el Uso Responsable de la Inteligencia Artificial en el Servicio Público" del Gobierno de Irlanda (Department of Public Expenditure, NDP Delivery and Reform, 2024), contextualizando sus principios y herramientas al marco normativo y misional del Estado colombiano. Este documento no es solo un simple marco normativo sino una brújula ética y metodológica que permitirá orientar a los funcionarios en la toma de decisiones sobre cuándo, cómo y para qué utilizar soluciones basadas en inteligencia artificial en sus funciones. Es un marco vivo, sujeto a actualización periódica conforme evolucione la tecnología, la regulación y las necesidades institucionales.

Invitamos a todos los servidores del Ministerio Público a apropiarse de estos principios, a participar activamente en la construcción de una institución digitalmente responsable y a mantener siempre presente que la tecnología debe servir al ser humano, nunca reemplazar el sentido de humanidad, el criterio profesional, el juicio ético ni la responsabilidad pública que nos caracteriza.

Gregorio Eljach Pacheco
Procurador General de la Nación

Glosario

Agencia humana: capacidad de las personas para ejercer control significativo sobre los sistemas de IA y sus decisiones.

Agente de IA: sistema de inteligencia artificial que puede percibir su entorno, tomar decisiones y ejecutar acciones de forma autónoma o semiautónoma mediante el uso de herramientas para alcanzar un objetivo específico.

Alucinación: fenómeno en el que un modelo de inteligencia artificial generativa produce respuestas que suenan lógicas o convincentes, pero que son factualmente incorrectas, inventadas o no están respaldadas por sus fuentes de datos.

Auditoría algorítmica: evaluación técnica y normativa, preferiblemente por terceros independientes, del diseño, funcionamiento e impacto de un sistema de IA para verificar su cumplimiento legal, mitigación de sesgos y rendimiento.

***AI Act* (Reglamento de IA de la UE):** regulación europea que establece normas armonizadas para sistemas de IA, clasificándolos por niveles de riesgo.

Algoritmo: conjunto de reglas y procedimientos lógicos que permiten resolver un problema de manera eficiente y automatizada.

Análisis de impacto algorítmico: evaluación sistemática de los posibles efectos de un sistema de IA sobre los derechos humanos y el bienestar social.

Aprendizaje por refuerzo: paradigma del aprendizaje automático mediante el cual el sistema aprende patrones en los datos por medio de ensayo-error.

Aprendizaje profundo (*deep learning*): redes neuronales con múltiples capas que permiten analizar información compleja, como procesar documentos mediante OCR de alta precisión y transcribir audiencias.

Aprendizaje no supervisado: paradigma del aprendizaje automático que permite a los sistemas aprender patrones a partir de datos históricos sin el uso de etiquetas.

Aprendizaje supervisado: paradigma del aprendizaje automático que utiliza ejemplos etiquetados de datos históricos para que el sistema aprenda patrones.

Ataques adversariales: manipulaciones intencionadas de las entradas de un sistema de IA para alterar su funcionamiento, tales como el jaqueo de instrucciones (*prompt hacking*), envenenamiento de datos (*data poisoning*) o la evasión de modelos (*model evasion*).

Automatización robótica de procesos (RPA): tecnología que permite automatizar tareas repetitivas mediante software que imita acciones humanas.

Bitácora de decisión: registro auditable que documenta tanto la recomendación emitida por un sistema de IA como la decisión final tomada por el supervisor humano, incluyendo la justificación en caso de que esta difiera de la sugerencia del algoritmo.

Caja negra (*Black box*): característica de modelos de IA complejos (como las redes neuronales profundas) cuyo procesamiento interno para llegar a un resultado es opaco, carece de explicabilidad y es difícil de interpretar para los seres humanos.

Canvas de IA responsable: herramienta estructurada que funciona como plantilla para talleres colaborativos, orientando a equipos multidisciplinarios a evaluar propuestas de IA desde perspectivas técnicas, éticas, jurídicas y organizacionales antes de su aprobación e implementación.

Ciclo de vida de la IA: serie de etapas por las que pasa un sistema de IA desde su concepción hasta su retiro.

Costos de infraestructura: costo del hardware, los recursos de computación en la nube y el almacenamiento necesarios para entrenar y ejecutar los modelos de IA.

Costos de datos: adquisición, preparación y etiquetado de datos para los modelos de IA.

Costos de desarrollo: cubre los salarios de los ingenieros de IA, científicos de datos y otros especialistas involucrados en la construcción y el perfeccionamiento de los modelos, así como el costo de las licencias y herramientas de software.

Costos operativos: gastos continuos o recurrentes asociados a la ejecución y el mantenimiento de los sistemas de IA, tales como el consumo de energía, el monitoreo y las actualizaciones.

***Data cards* (Fichas técnicas de datos):** documentación estandarizada sobre conjuntos de datos que incluye fuentes de origen, fechas de recolección, tamaño muestral, variables, métodos de recolección, transformaciones y políticas de retención.

Deriva de datos (*Data drift / Concept drift*): cambios en la distribución de los datos o en los comportamientos que se presentan durante la operación en condiciones reales del sistema, lo cual puede generar una degradación en las métricas de desempeño del modelo.

Datos de entrenamiento: conjunto de información histórica y ejemplos que se introducen en un modelo de aprendizaje automático durante su fase de desarrollo para que este aprenda a extraer características y reconocer patrones.

Datos personales: cualquier información concerniente a personas naturales identificadas o identificables (Ley 1581 de 2012).

Datos sensibles: información que afecta la intimidad o cuyo uso indebido puede generar discriminación (origen racial, orientación política, convicciones religiosas, datos de salud, biométricos, etc.).

***Deepfake*:** contenido sintético creado con IA que simula de manera realista la apariencia, voz o acciones de una persona real.

Desinformación: información falsa o engañosa creada y difundida deliberadamente con propósitos manipuladores.

Evaluación de Impacto en la Equidad (EIE): análisis previo a la implementación de un sistema de IA para identificar posibles efectos discriminatorios o impactos sobre la equidad y definir sus respectivas medidas de mitigación.

Evaluación de Impacto en la Protección de Datos (EIPD): análisis obligatorio antes de implementar sistemas de IA que traten datos personales a gran escala o datos sensibles, orientado a identificar riesgos de privacidad y establecer cómo mitigarlos.

Evaluación de Impacto sobre Derechos Humanos (EIDH): evaluación recomendada antes de desplegar IA en contextos sensibles que puedan vulnerar derechos humanos.

Evaluación de Impacto sobre Derechos de la Niñez (EIDN): evaluación obligatoria aplicable a sistemas de IA que procesen datos de niños, niñas y adolescentes, encargada de documentar riesgos específicos (exposición a contenidos inapropiados, ciberacoso, afectación cognitiva, etc.) y planes de monitoreo.

Explicabilidad: capacidad de un sistema de IA para proporcionar información comprensible sobre cómo llega a sus resultados.

Falsificación de pruebas: síntesis de audio o video (*deepfakes*) que crea evidencia falsa realista.

Gobernanza de datos: conjunto de políticas, procedimientos y controles para gestionar la calidad, seguridad, privacidad y uso de los datos.

IA agéntica (*Agentic AI*): enfoque de la inteligencia artificial centrado en sistemas con un alto grado de autonomía, proactividad y capacidad para razonar, planificar y ejecutar secuencias complejas de acciones minimizando la intervención humana continua.

IA de propósito general (GPAI): sistema de inteligencia artificial que puede realizar de manera competente una amplia gama de tareas distintas (como texto, imagen o código), a diferencia de la IA estrecha diseñada para un solo fin.

IA ética: componente de la IA fiable que garantiza el respeto de principios y valores éticos fundamentales a lo largo de todo el ciclo de vida de los sistemas.

IA fiable: enfoque de inteligencia artificial sustentado en tres componentes esenciales (según directrices europeas): debe ser lícita, ética y robusta.

IA lícita: componente de la IA fiable referido al cumplimiento estricto de todas las leyes y regulaciones aplicables.

IA robusta: componente de la IA fiable caracterizado por la solidez técnica y social que previene daños accidentales, incluso cuando existen buenas intenciones.

IA generativa (GenAI): tipo de inteligencia artificial capaz de generar contenidos nuevos (texto, imágenes, audio, video, código) basándose en datos de entrenamiento.
Inferencia: fase técnica en la que se aplica un modelo de inteligencia artificial (previamente entrenado) a nuevos datos para generar resultados o predicciones.

Ingeniería de instrucciones (Prompt engineering): técnica de formular, estructurar y optimizar las peticiones de entrada (prompts) dirigidas a un sistema de IA generativa para guiarlo y obtener el resultado más preciso y adecuado.

Inteligencia artificial (IA): sistema basado en máquinas que puede, para objetivos explícitos o implícitos, generar resultados como predicciones, recomendaciones o decisiones que influyen en entornos reales o virtuales (definición OCDE, 2024).

Interoperabilidad: capacidad de diferentes sistemas y organizaciones para trabajar conjuntamente e intercambiar información.

Irreproducibilidad: riesgo técnico en el que los modelos generativos producen resultados diferentes con la misma entrada.

***Large language model* (LLM):** modelo de lenguaje de gran escala entrenado con enormes cantidades de texto para comprender y generar lenguaje natural.

***Machine learning* (aprendizaje automático):** subcampo de la IA que permite a los sistemas aprender patrones a partir de datos sin ser programados explícitamente.

Manipulación de opinión: uso de bots que simulan movimientos sociales, amplificando narrativas o atacando instituciones.

Matriz de responsabilidades (Matriz RACI): herramienta que define y documenta claramente los roles de cada sistema de IA, identificando quién es el responsable, quién rinde cuentas, quién es consultado y quién es informado.

Ministerio Público: órgano de control del Estado colombiano conformado por la Procuraduría General de la Nación, la Defensoría del Pueblo y las personerías municipales y distritales (C.P., artículos 275-284).

***Model cards* (Fichas técnicas de modelos):** documentación técnica de un modelo de IA que detalla su arquitectura, hiperparámetros, proceso de entrenamiento, métricas de desempeño, así como sus limitaciones y casos de uso apropiados e inapropiados.

Modelo de IA: representación matemática y computacional que aprende patrones de los datos para realizar predicciones o tomar decisiones.

Patrones oscuros (*dark patterns*): diseños de interfaces o algoritmos que manipulan o confunden a los usuarios (especialmente menores de edad) respecto a las consecuencias de sus acciones en los entornos digitales.

Principios éticos de la IA: valores fundamentales que guían el desarrollo y uso responsable de la inteligencia artificial.

Privacidad desde el diseño (*Privacy by design*): principio normativo que exige integrar las medidas de protección de datos personales en la arquitectura tecnológica y organizativa de un sistema de IA desde su fase inicial de concepción.

RAG (Generación Aumentada por Recuperación): arquitectura que conecta un modelo de IA generativa a bases de datos documentales externas para fundamentar sus respuestas en información específica, verificable y actualizada, reduciendo el riesgo de alucinaciones.

Rendición de cuentas (*accountability*): responsabilidad clara sobre las decisiones y resultados de los sistemas de IA.

Robustez: capacidad de un sistema de IA para funcionar de manera confiable bajo diferentes condiciones y resistir ataques.

***Sandbox* regulatorio:** entorno de pruebas seguro y controlado por las autoridades públicas que permite a las organizaciones experimentar con nuevas tecnologías de IA bajo supervisión y con flexibilización normativa temporal.

Sesgo algorítmico: distorsión sistemática en los resultados de un sistema de IA que refleja o amplifica prejuicios presentes en los datos de entrenamiento.

Sistemas basados en reglas: sistemas que operan mediante reglas explícitas programadas (p. ej*. if-then statements*) que formalizan el conocimiento de manera precisa, pero que carecen de capacidad de aprendizaje automático.

Sistemas de apoyo a la decisión (*decision-support*): sistemas de IA que proveen recomendaciones, análisis predictivos o síntesis para facilitar el juicio humano, sin ejecutar acciones de manera autónoma.

Sistemas de automatización de tareas: sistemas que ejecutan funciones específicas y repetitivas (p. ej.. clasificación de documentos o transcripción) bajo un nivel de supervisión humana reducida.

Sistemas de decisión autónoma: sistemas que toman decisiones con un impacto directo sobre los derechos o la situación jurídica de las personas sin mediación humana sustantiva.

Sistema de toma de decisiones automatizadas (SDA): tecnología que asiste o sustituye el juicio humano en la toma de decisiones.

Supervisión humana (*human-in-the-loop*): intervención activa de personas en las decisiones críticas tomadas por sistemas de IA.

Supervisión humana sobre el circuito (*human-over-the-loop*): modelo de intervención aplicable a sistemas de riesgo moderado, en el cual el sistema opera pero una persona tiene la capacidad de intervenir y corregir el sistema en cualquier momento.

Transparencia algorítmica: disponibilidad de información sobre sistemas algorítmicos que permite conocer su operación y valorar su rendimiento.

Transparencia activa: divulgación proactiva de información sobre sistemas de IA sin que sea solicitada.

Transparencia pasiva: respuesta a solicitudes específicas de información sobre sistemas de IA.

Técnicas de mejora de la privacidad (PETs): conjunto de métodos y herramientas, como la anonimización, la seudonimización y el cifrado, empleados para minimizar los riesgos de identificación de personas y proteger la confidencialidad de los datos.

Verificación y validación: fase en el ciclo de vida donde los sistemas de IA son probados.

1. Resumen ejecutivo

1.1 Contexto y necesidad

El Estado colombiano, y en específico el Ministerio Público, enfrenta el imperativo estratégico de incorporar soluciones basadas en inteligencia artificial (IA) para fortalecer su capacidad misional en un contexto de sobrecarga procesal, desafíos de legitimidad institucional y proliferación de desinformación. En este documento se presentan unos Principios éticos y un marco de gobernanza los cuales establecen las directrices para una adopción responsable, ética y efectiva de la IA, alineada con el mandato constitucional del Ministerio Público de la República de Colombia como garante del orden jurídico y defensor de los derechos humanos.

1.2 Fundamentos

El presente marco se sustenta en tres componentes esenciales de la IA fiable, siguiendo las directrices europeas (High-Level Expert Group on AI, 2019):

- **IA lícita:** hace referencia al cumplimiento estricto de todas las leyes y regulaciones aplicables (Constitución Política, Ley 1581 de 2012, Ley 1712 de 2014, CONPES 3975 de 2019, CONPES 41,CONPES 4144 de 2025, Directiva Conjunta 007 de 2025).

- **IA ética:** se relaciona con la garantía del respeto de principios y valores éticos fundamentales a lo largo de todo el ciclo de vida de los sistemas.

- **IA robusta:** tiene que ver con la solidez técnica y social que previene daños accidentales incluso con buenas intenciones.

1.3 Ocho principios rectores

Este marco se fundamenta sobre la base de ocho (8) principios éticos identificados en la normativa internacional y nacional, que deben cumplirse simultáneamente a

la hora de planear, diseñar y desplegar cualquier proyecto o solución basada en IA:

i. Agencia y supervisión humana (*human agency and oversight*): los sistemas de IA deben empoderar a las personas y permitir supervisión humana efectiva. Las decisiones críticas que afecten derechos humanos recaen estrictamente sobre el control de los individuos (*human-in-the-loop*).

ii. Robustez técnica y seguridad: los sistemas deben ser seguros, precisos, confiables y resilientes ante ataques, errores o fallos durante todo su ciclo de vida.

iii. Privacidad y gobernanza de datos: cumplimiento riguroso de la Ley 1581 de 2012, protegiendo datos personales, implementando medidas de seguridad y respetando los derechos de los titulares.

iv. Transparencia y explicabilidad: los sistemas deben ser transparentes en su funcionamiento y capaces de explicar sus decisiones de manera comprensible, cumpliendo con la Directiva Conjunta 007 de 2025 sobre transparencia algorítmica.

v. Diversidad, no discriminación y equidad: prevención activa de sesgos, asegurando que los sistemas no perpetúen ni amplifiquen discriminaciones, y promoviendo acceso equitativo sin importar género, etnia, discapacidad u otra condición.

vi. Bienestar social y ambiental: los sistemas deben contribuir positivamente a la sociedad y minimizar impactos ambientales negativos, promoviendo el desarrollo sostenible.

vii. Rendición de cuentas (*accountability*): establecimiento de responsables y responsabilidades claras, mecanismos de auditoría, trazabilidad de decisiones y canales efectivos de recurso y reparación.

viii. Prevalencia de los derechos de niños, niñas y adolescentes: protección especial reforzada cuando los sistemas puedan afectar a población infantil y adolescente, siguiendo el interés superior del menor.

1.4 Herramientas prácticas

Esta guía aborda cuatro áreas clave que tienen como propósito apoyar a los servidores públicos en la adopción y el uso de la IA de manera responsable, las cuales son:

i. Ocho principios para una IA responsable: en esta sección aprenderás sobre los ocho (8) principios clave para construir y usar la IA de manera responsable.

ii. Marco de decisión: en esta sección se presenta una metodología estructurada para evaluar si la IA es la solución apropiada, qué tipo de sistema utilizar y cómo garantizar el uso responsable desde la fase de diseño.

iii. Canvas de IA responsable: ésta sección presenta una herramienta colaborativa para mapear proyectos de IA asegurando la incorporación de principios éticos en todas las dimensiones del sistema.

iv. Guía del ciclo de vida de la IA: en esta sección se desarrolla una orientación específica para aplicar los principios en cada fase del desarrollo e implementación de sistemas de IA (planificación, datos, construcción, validación, despliegue, monitoreo, retiro).

1.5 Compromiso institucional

El marco representa el compromiso del Ministerio Público con la innovación responsable, al adoptar tecnología de vanguardia sin comprometer valores éticos; con la transparencia, al informar proactivamente a la ciudadanía sobre el uso de IA; con la participación, al involucrar a funcionarios, expertos y ciudadanía en decisiones sobre IA; con la mejora continua, al actualizar el marco conforme evolucione la tecnología y la regulación, y con la protección de derechos, ya que busca garantizar que la IA fortalezca —y nunca erosione— los derechos humanos.

Este documento marca un hito en la transformación digital del Ministerio Público de Colombia, estableciendo estándares de excelencia ética que posicionan al país como referente regional en el uso responsable de inteligencia artificial.

2. Introducción y contexto

2.1 El Ministerio Público en la era digital

De conformidad con los artículos 117 y 118 de la Constitución Política de Colombia, el Ministerio Público es un órgano de control del Estado, integrado por tres entidades: la Procuraduría General de la Nación, la Defensoría del Pueblo y las personerías distritales y municipales.

A través de estas instituciones, el Ministerio Público cumple la función de proteger y promover los derechos humanos, salvaguardar el interés público y vigilar la conducta oficial de quienes ejercen funciones públicas. Si bien cada una de estas entidades cuenta con autonomía administrativa y financiera, así como con una estructura propia definida por la ley, la Constitución establece que el Procurador General de la Nación es el supremo director del Ministerio Público, encargado de ejercer la dirección general y la vigilancia superior de sus funciones, sin que ello desconozca la independencia funcional de la Defensoría del Pueblo y de las personerías en el ámbito de sus competencias.

Ahora bien, en el siglo XXI, estas instituciones enfrentan desafíos sin precedentes:

- **Desafíos de eficiencia interna:** (i) sobrecarga procesal con tiempos de respuesta que comprometen la justicia oportuna; (ii) gestión de volúmenes masivos de información proveniente de múltiples fuentes; (iii) necesidad de priorización estratégica de intervenciones con recursos limitados, y (iv) exigencia de mayor impacto en la lucha contra la corrupción y la mala gestión pública.

- **Desafíos de legitimidad externa:** (i) caída sostenida en la confianza ciudadana hacia las instituciones públicas (solo 32 % [OCDE, 2024]); (ii) proliferación de desinformación que erosiona la percepción pública sobre la labor institucional; (iii) ecosistema digital donde el 76.68 % de la desinformación se propaga por redes

sociales (Transparencia por Colombia, 2024), y (iv) necesidad de mejorar la comunicación y transparencia con la ciudadanía.

Desde el campo de la inteligencia artificial emergen herramientas que tienen el potencial para abordar y transformar ambos frentes, pero su adopción tiene que realizarse con extrema prudencia y responsabilidad ética, dado el poder y la sensibilidad de las funciones que ejerce el Ministerio Público.

De hecho, las herramientas de inteligencia artificial utilizadas en los procesos disciplinarios tienen como propósito fortalecer la eficiencia y la eficacia de la función disciplinaria, sin perder de vista la verificación rigurosa de la información y la confiabilidad de las fuentes empleadas. Su uso debe orientarse, ante todo, a garantizar el respeto por los derechos fundamentales y las garantías procesales de todos los sujetos que intervienen en el proceso. En este sentido, dichas herramientas deben contribuir al cumplimiento de los principios que rigen la función administrativa, consagrados en la Constitución Política y desarrollados en la ley, en particular en el Código General Disciplinario y sus reformas, tales como la economía procesal, la celeridad y la búsqueda de la verdad, entre otros.

2.2 Marco constitucional y legal

La adopción de IA en el Ministerio Público no ocurre en un vacío normativo, sino que está enmarcada por un robusto andamiaje legal que define tanto las oportunidades como los límites del uso de esta tecnología. Una compilación de los principales marcos legales identificados a nivel internacional y nacional se encuentra resumida en el Apéndice A.

Fundamento constitucional

Los artículos 118 y 275-284 de la Constitución Política de Colombia (1991) establecen la naturaleza, composición y funciones del Ministerio Público, definiendo que será ejercido por el Procurador General de la Nación, por el Defensor del Pueblo, por los procuradores delegados y los agentes del Ministerio Público, ante las autoridades jurisdiccionales, por los personeros municipales y por los demás funcionarios que determine la ley, correspondiéndole la guarda y promoción de los derechos humanos, la protección del interés público y la vigilancia de la conducta oficial de quienes desempeñan funciones públicas. El Ministerio Público en su calidad de garante de la efectividad de los derechos contenidos en la constitución y en la ley, es la institución

encargada de velar en particular de los siguientes artículos:

- **Artículo 15:** protege el derecho a la intimidad personal y familiar, y el derecho al *habeas data*, fundamentales para el uso ético de sistemas de IA que procesen datos personales.
- **Artículo 20:** garantiza la libertad de expresión e información y el derecho a fundar medios masivos de comunicación, relevante para abordar la desinformación y respetando libertades fundamentales.
- **Artículo 23:** consagra el derecho de petición, cuya eficiencia puede mejorarse mediante IA sin afectar la garantía de respuesta oportuna y completa.
- **Artículo 74:** este artículo tiene implicaciones con la transparencia sobre el uso de algoritmos y sistemas basados en IA en el sector público.
- **Artículo 209:** define los principios de la función administrativa (igualdad, moralidad, eficacia, economía, celeridad, imparcialidad y publicidad) que deben regir también el uso de IA.

Marco legal de protección de datos

La Ley 1581 de 2012, conocida como el Régimen general de protección de datos personales que regula el derecho al *habeas data* y establece los principios y reglas para la recolección, tratamiento, almacenamiento y circulación de datos personales. A su vez, el Decreto 1377 de 2013 reglamenta aspectos de la Ley 1581 de 2012, precisando el tratamiento de datos sensibles y estableciendo medidas de seguridad. De esta manera, los sistemas basados en IA que procesen datos personales deben cumplir rigurosamente con sus disposiciones.

Marco legal de transparencia y acceso a la información

Por su parte, la Ley de Transparencia y del Derecho de Acceso a la Información Pública Nacional (Ley 1712 de 2014) regula el derecho fundamental de acceso a información pública, los procedimientos para su ejercicio y las excepciones. Su artículo 23 establece que el Ministerio Público es responsable de velar por el cumplimiento de estas obligaciones, incluyendo la transparencia sobre sistemas de IA.

La Directiva Conjunta 007 de 2025 acerca de los estándares sobre transparencia algorítmica, emitida por el Procurador General de la Nación y la Defensoría del Pueblo en cumplimiento de la Sentencia T-067 de 2025 de la Corte Constitucional, establece estándares mínimos de transparencia activa y pasiva para sistemas algorítmicos utilizados por el Estado, definiendo qué información debe publicarse, cómo responder solicitudes ciudadanas y cuándo realizar análisis de impacto algorítmico.

Marco legal de función pública y régimen disciplinario

En primer lugar, el régimen disciplinario en Colombia se encuentra regulado principalmente por el Código General Disciplinario, contenido en la Ley 1952 de 2019 y sus reformas, el cual establece las normas aplicables a los servidores públicos y a los particulares que ejercen funciones públicas, administran recursos del Estado o cumplen labores de interventoría en obras públicas. Este marco normativo define las conductas sancionables, las autoridades competentes y los procedimientos para la investigación y el juzgamiento disciplinario.

De manera transitoria, aún se encuentra vigente el Código Disciplinario Único, Ley 734 de 2002 y sus modificaciones, para aquellos casos expresamente previstos por la ley. Tanto el régimen anterior como el actual no resultan incompatibles con el uso de herramientas tecnológicas que faciliten la labor de las autoridades disciplinarias; por el contrario, permiten su incorporación siempre que se respeten los principios y garantías que rigen el debido proceso.

En este contexto, el uso de herramientas de inteligencia artificial en los procesos disciplinarios debe orientarse a fortalecer la eficiencia y la eficacia de la función disciplinaria, deben contribuir al cumplimiento de los principios de la función administrativa, consagrados constitucionalmente y desarrollados en el Código General Disciplinario y sus reformas.

Y es así que, este marco de gobernanza en el uso de inteligencia artificial resulta fundamental para prevenir riesgos como la generación de normas inexistentes, la cita de jurisprudencia errónea o la incorporación de información falsa. Por ello, el uso responsable de estas herramientas debe estar siempre acompañado de controles humanos efectivos y del estricto respeto por las garantías procesales de los sujetos que intervienen en el proceso disciplinario.

Políticas nacionales de transformación digital e IA

La primera política nacional que reconoció a la IA como acelerador de la transformación digital fue el CONPES 3975 de 2019 sobre Política Nacional para la Transformación Digital e Inteligencia Artificial, que estableció acciones para desarrollar condiciones habilitantes (CONPES [DNP], 2019).

Posteriormente, el CONPES 4144 de 2025 sobre Política Nacional de Inteligencia Artificial, actualiza y profundiza la estrategia nacional, estableciendo seis ejes

estratégicos: (i) ética y Gobernanza; (ii) datos e infraestructura; (iii) investigación, desarrollo e innovación; (iv) desarrollo de capacidades y talento digital; (v) mitigación de riesgos, y (vi) uso y adopción de la IA (CONPES [DNP], 2025).

Así mismo, el Marco ético para la inteligencia artificial en Colombia, documento emitido por la Presidencia de la República que establece principios éticos aplicables a proyectos de IA: transparencia y explicación, privacidad, control humano de las decisiones, seguridad, responsabilidad, no discriminación, inclusión, prevalencia de derechos de NNA, y sostenibilidad ambiental (Departamento Administrativo de la Presidencia de la República [Dapre], 2021).

Por último, el Decreto 1263 de 2022 del Ministerio de Tecnologías de la Información y las Comunicaciones, acerca de los Lineamientos para la Transformación Digital de la Administración Pública, define estándares aplicables a la transformación digital del Estado, incluyendo el uso de tecnologías emergentes como IA.

Marco legal de propiedad intelectual y delitos con IA

Sobre derechos de propiedad intelectual, la Ley 2502 de 2025 relacionada con falsedad personal con inteligencia artificial, modifica el artículo 296 del Código Penal para tipificar y agravar la suplantación de identidad utilizando IA (deepfakes), estableciendo directrices para prevención y control de uso indebido.

A su vez, el Código de Procedimiento Penal (Ley 906, 2004) regula procedimientos para la persecución de delitos, incluyendo aquellos facilitados por IA. Los cuáles se extienden a los protocolos relacionados con la cadena de custodia y validez probatoria relevantes cuando se utilizan sistemas de IA en investigaciones.

Jurisprudencia constitucional relevante

Inicialmente un referente clave en el debate sobre el uso de la inteligencia artificial en la administración de justicia en Colombia es la Sentencia T-323 de 2024 de la Corte Constitucional, con ponencia del magistrado Juan Carlos Cortés González.

En esta decisión, considerada histórica por ser la primera en desarrollar de manera sistemática este tema, la Corte dejó claro que las herramientas de inteligencia artificial no pueden reemplazar al juez en la toma de decisiones judiciales. Si bien reconoció que estas tecnologías pueden ser útiles como apoyo en labores administrativas, documentales o de organización de la información,

enfatizó que la valoración de las pruebas, la interpretación jurídica y la adopción de decisiones corresponden exclusivamente a una autoridad humana.

El uso de la IA, señaló la Corte, debe realizarse de manera razonable y ponderada, con pleno respeto por los derechos fundamentales, bajo criterios de transparencia, responsabilidad, verificación de la información, protección de datos personales y control humano efectivo.

Por otro lado, la Sentencia T-067 de 2025 de la Corte Constitucional marca un antecedente jurisprudencial que establece que la transparencia algorítmica es componente esencial del derecho fundamental de acceso a información pública, ordenando la expedición de estándares sobre transparencia algorítmica.

Estándares internacionales adoptados por Colombia

Se identifican inicialmente los Principios de IA de la OCDE (2019, actualizados 2024) que Colombia adoptó formalmente en mayo de 2019, comprometiéndose con IA inclusiva, sostenible, centrada en derechos humanos, transparente, robusta, segura y responsable (Organisation for Economic Co-operation and Development [OECD, por sus siglas en inglés], 2019).

Igualmente se identificó la Recomendación sobre la ética de la inteligencia artificial de la UNESCO, a la cual Colombia se adhirió en noviembre de 2021, comprometiéndose con valores y principios éticos que integren evaluaciones de impacto, mecanismos de gobernanza y políticas éticas de datos (United Nations Educational, Scientific and Cultural Organization [UNESCO, por sus siglas en inglés], 2021).

Adicionalmente, las Directrices éticas para una IA fiable del Grupo de Expertos de Alto Nivel de la Comisión Europea, aunque no son vinculantes, constituyen referente técnico y ético adoptado globalmente, estableciendo los siete principios fundacionales de IA responsable (High-Level Expert Group on Artificial Intelligence, 2019).

Igualmente, el Reglamento de IA de la Unión Europea - AI Act (Reglamento UE 2024/1689) se establece como la primera regulación comprensiva de IA a nivel mundial, que clasifica sistemas por riesgo y establece obligaciones diferenciadas. Aunque no es ley colombiana, constituye referente de mejores prácticas regulatorias (Parlamento Europeo y Consejo de la Unión Europea, 2024).

2.3 Visión estratégica para la IA en el Ministerio Público

La adopción de la inteligencia artificial está impulsando una transformación de todas las instituciones y corporaciones del sector público en Colombia y el mundo. La visión del Ministerio Público para la adopción de inteligencia artificial se ha venido articulando en torno a tres ejes estratégicos:

• **Eje 1 - Gobernanza ética y responsable.** Establecer un marco institucional robusto que garantice que todo uso de IA respete los derechos humanos, los principios constitucionales y los valores democráticos:

- Principios éticos claros y operacionalizables.
- Mecanismos de supervisión y rendición de cuentas.
- Transparencia algorítmica proactiva.
- Evaluaciones de impacto previas a despliegues.
- Participación ciudadana y control social.
- Formación continua de funcionarios.

• **Eje 2 - Fortalecimiento de la capacidad misional.** La IA permite potenciar las capacidades de control disciplinario, vigilar la gestión pública, proteger derechos humanos y estar al servicio de la ciudadanía. Algunas de las aplicaciones de la IA en este campo son:

- Análisis inteligente de grandes volúmenes de información.
- Detección temprana de riesgos y patrones de irregularidad.
- Optimización de procesos administrativos y disciplinarios.
- Priorización estratégica de intervenciones.
- Mejora en tiempos de respuesta y calidad de servicio.

• **Eje 3 - Protección del ecosistema informativo**. La adopción de la IA en el Ministerio Público hará posible el desarrollo de capacidades institucionales que permitan combatir la desinformación que erosiona la confianza pública y el debate democrático. En este caso se identifican las siguientes aplicaciones:

- Monitoreo y detección temprana de narrativas falsas.
- Análisis de campañas coordinadas de desinformación.
- Verificación de hechos mediante herramientas automatizadas.

- Alfabetización digital de la ciudadanía.
- Transparencia proactiva sobre la gestión institucional.

2.4 Alcance y audiencia de este marco

Este Marco de Principios de Gobernanza y Ética aplica a:

- **Sistemas cubiertos:** (i) todos los sistemas de IA desarrollados, adquiridos, implementados o utilizados por la Procuraduría General de la Nación, la Defensoría del Pueblo y las Personerías; (ii) sistemas que utilicen técnicas de aprendizaje automático (machine learning), procesamiento de lenguaje natural (NLP), visión artificial o IA generativa; (iii) sistemas de toma de decisiones automatizadas que afecten a funcionarios o ciudadanos, y (iv) sistemas de análisis predictivo para priorización de casos o detección de riesgos.

- **Fases del ciclo de vida:** (i) planificación y diseño conceptual; (ii) recolección y procesamiento de datos; (iii) construcción y entrenamiento de modelos; (iv) validación y certificación; (v) despliegue e implementación; (vi) operación y monitoreo; (vii) actualización y mantenimiento, y (viii) retiro o desmantelamiento.

- **Modalidades de adquisición:** (i) desarrollo interno (in-house); (ii) adquisición de soluciones comerciales; (iii) colaboración con universidades o centros de investigación; (iv) asociaciones público-privadas, y (v) uso de modelos de IA de propósito general (como LLMs).

El marco está dirigido a múltiples audiencias dentro y fuera del Ministerio Público, en particular se identifican las siguientes:

- **Audiencias primarias (uso obligatorio):**

  - **Líderes institucionales:** Procurador General, Defensor del Pueblo, viceprocurador, vicedefensor y secretarios generales. Responsables de decisiones estratégicas sobre adopción de IA.
  - **Directivos de áreas misionales:** jefes de oficina, coordinadores de grupo y directores regionales. Responsables de identificar oportunidades de aplicación de IA en sus áreas.
  - **Profesionales de TI, datos y analítica:** oficinas de tecnología de la información y de planeación, y equipos de analítica de datos. Responsables del desarrollo técnico e implementación de sistemas.

- **Responsables de contratación:** áreas de contratación y adquisiciones. Deben incorporar requisitos éticos en pliegos y contratos de soluciones de IA.
- **Usuarios finales de sistemas de IA:** todo funcionario que utilice sistemas de IA en sus labores cotidianas debe comprender sus principios, limitaciones y responsabilidades asociadas.

- **Audiencias secundarias (consulta recomendada):**

  - **Proveedores y desarrolladores externos:** empresas, consultores e instituciones que desarrollen o suministren soluciones de IA al Ministerio Público deben conocer y adherirse a estos principios.
  - **Academia e investigadores:** universidades y centros de investigación que colaboren en proyectos de IA para el Ministerio Público.
  - **Ciudadanía y organizaciones de la sociedad civil:** últimos beneficiarios y sujetos de control social sobre el uso de IA en instituciones públicas.
  - **Otros organismos de control:** Contraloría General de la República, organismos de control territorial, autoridades de protección de datos. Para armonización de criterios éticos en el sector público.

3. La IA en el Ministerio Público

El presente capítulo describe las bases conceptuales y el contexto para comprender el rol transformador de la inteligencia artificial en el Ministerio Público. Tras explorar en capítulos anteriores el contexto institucional, ahora corresponde definir qué es la IA, sus aplicaciones en las funciones misionales, los beneficios que aporta y los riesgos que conlleva. Este capítulo fundamenta los principios de gobernanza y el marco ético desarrollados posteriormente.

3.1 ¿Qué es la inteligencia artificial?

La inteligencia artificial ha evolucionado de manera veloz de un concepto teórico a un campo de estudio de donde emergen una serie de tecnologías que están redefiniendo tanto el sector público como el privado en Colombia y el mundo. Para el presente marco ético, es fundamental presentar una definición clara, técnicamente rigurosa y normativamente alineada con instrumentos nacionales e internacionales.

Definiciones normativas

La OCDE define un sistema de IA como «(...) un sistema basado en máquinas que, para objetivos explícitos o implícitos, infiere, a partir de la información que recibe, cómo generar resultados tales como predicciones, contenidos, recomendaciones o decisiones que pueden influir en entornos reales o virtuales» (OECD, 2019). En este caso se destaca la capacidad de la IA para inferir patrones y producir salidas con impacto real, relevante para contextos de administración pública y justicia.

El Reglamento (UE) 2024/1689 (AI Act) define sistema de IA como un sistema automatizado «(...) diseñado para funcionar con distintos niveles de autonomía y que puede mostrar capacidad de adaptación tras el despliegue (...)» (Parlamento Europeo y Consejo de la Unión Europea, 2024, art. 3.1). En este caso se subraya la autonomía y adaptación como aspectos críticos sobre los cuales se deben evaluar riesgos.

En Colombia, el Marco ético para la IA la define como «... un campo de la informática dedicado a resolver problemas cognitivos comúnmente asociados con la inteligencia humana o seres inteligentes, entendidos como aquellos que pueden adaptarse a situaciones cambiantes. Su base es el desarrollo de sistemas informáticos, la disponibilidad de datos y los algoritmos» (Dapre, 2021, p. 8). El CONPES 4144 de 2025 subraya que «Es fundamental que los actores responsables del desarrollo, uso y apropiación de la inteligencia artificial como herramienta clave para el crecimiento económico y el desarrollo social sostenible, sigan una guía clara que oriente sus acciones en ese camino» (CONPES [DNP], 2025, p. 7) justificando de manera directa la existencia de este marco de gobernanza.

Tipos de inteligencia artificial

La IA es un campo del conocimiento como lo son el derecho o la economía, el cual se compone de diferentes subdisciplinas, entre las cuales podemos encontrar las siguientes:

- **Sistemas basados en reglas.** Operan mediante reglas explícitas programadas (*if-then statements*). Útiles donde el conocimiento se formaliza en reglas precisas, pero carecen de capacidad de aprendizaje.

- **Aprendizaje automático (*machine learning*).** Algoritmos que aprenden patrones de datos históricos mediante tres paradigmas: aprendizaje supervisado (ejemplos etiquetados), no supervisado (patrones sin etiquetas) y por refuerzo (ensayo-error) (Russell y Norvig, 2020).

- **Aprendizaje profundo (*deep learning*).** Redes neuronales con múltiples capas que han revolucionado el reconocimiento de imágenes, voz y procesamiento de lenguaje natural (LeCun et al., 2015). Permite analizar documentos mediante OCR de alta precisión, transcribir audiencias e identificar patrones complejos.

- **Inteligencia artificial generativa.** Modelos que crean contenido nuevo como texto, código, imágenes, audio y video. Los *Large Language Models* (LLMs) o Modelos de Lenguaje Extensos pueden generar resúmenes legales, redactar respuestas a peticiones y asistir en investigación jurídica (Brown et al., 2020). La IA generativa es «particularmente poderosa para analizar datos oscuros» como documentos no estructurados, audios y videos, pero presenta riesgos como «alucinaciones» y reproducción de sesgos.

• **Agentes de IA (*AI Agents*).** Son sistemas diseñados para percibir su entorno a través de sensores y actuar sobre él mediante efectores para alcanzar un objetivo específico. Un agente puede ser desde un programa simple que filtra correos electrónicos hasta un robot complejo que navega en un entorno físico. Su característica principal es la autonomía para tomar decisiones y ejecutar acciones sin intervención humana directa.

• **IA agéntica (*Agentic AI*).** Representa una evolución de los agentes de IA, donde el sistema no solo ejecuta tareas, sino que también puede razonar, planificar y descomponer un objetivo complejo en una secuencia de pasos ejecutables. Estos sistemas, a menudo impulsados por LLM, pueden interactuar con herramientas externas (como software o APIs) y aprender de sus interacciones para mejorar su desempeño futuro. La IA agéntica se considera un paso clave hacia la creación de sistemas de inteligencia artificial más generales y capaces de resolver problemas de manera proactiva y autónoma.

Conceptos técnicos fundamentales

• **Algoritmo:** Es un conjunto finito y ordenado de instrucciones o reglas bien definidas que un sistema sigue para realizar una tarea o resolver un problema. En IA, los algoritmos son la base para el aprendizaje y el razonamiento (Cormen et al., 2009).

• **Modelo:** Es la representación matemática de un proceso del mundo real, construida por un algoritmo de aprendizaje automático a partir de los datos. No es el algoritmo en sí, sino el resultado de su aplicación. Por ejemplo, una red neuronal entrenada para identificar imágenes es un modelo.

• **Datos (*Data*):** Son la materia prima para cualquier sistema de IA. Pueden ser estructurados (como tablas en una base de datos) o no estructurados (como textos, imágenes, audios y videos). La calidad, cantidad y representatividad de los datos son determinantes para el rendimiento del modelo.

• **Entrenamiento (*Training*):** Es el proceso de alimentar un algoritmo con un conjunto de datos (datos de entrenamiento) para que aprenda a identificar patrones o hacer predicciones. Durante este proceso, los parámetros internos del modelo se ajustan progresivamente para minimizar el error entre sus predicciones y los resultados correctos.

• **Inferencia (*Inference*):** Una vez que el modelo ha sido entrenado, la inferencia es el proceso de utilizarlo para hacer predicciones sobre datos nuevos y nunca antes vistos. Es la fase en la que el modelo aplica lo que ha "aprendido" para realizar una tarea práctica.

• **Parámetros e Hiperparámetros:** Los parámetros son las variables internas del modelo que se aprenden automáticamente durante el entrenamiento (por ejemplo, los pesos en una red neuronal). Los hiperparámetros son las configuraciones externas que se establecen antes del entrenamiento para controlar cómo aprende el algoritmo (por ejemplo, la tasa de aprendizaje o el número de capas en una red neuronal).

• **Sobreajuste (*Overfitting*) y Subajuste (*Underfitting*):** El sobreajuste ocurre cuando un modelo aprende tan bien los datos de entrenamiento que memoriza el ruido y los detalles específicos en lugar de los patrones generales. Esto provoca que funcione mal con datos nuevos. El subajuste es el caso contrario: el modelo es demasiado simple y no logra capturar los patrones subyacentes, resultando en un bajo rendimiento tanto en los datos de entrenamiento como en los nuevos.

• **Transparencia y Explicabilidad (XAI):** La transparencia se refiere a la disponibilidad de información sobre cómo funciona un modelo de IA (sus datos, su algoritmo, su arquitectura). La explicabilidad es la capacidad de un sistema para justificar y hacer comprensibles sus decisiones o predicciones específicas en un lenguaje humano (Arrieta et al., 2020). Ambos conceptos son cruciales en contextos jurídicos, donde las decisiones deben ser auditables y fundamentadas.

El tipo de sistema de IA determina beneficios, riesgos y salvaguardas necesarias. Para el Ministerio Público, cuyas funciones tienen implicaciones sobre garantías fundamentales, es imprescindible que cada adopción de una solución basada en IA esté precedida por una comprensión rigurosa del tipo de sistema, sus capacidades, limitaciones y riesgos potenciales.

3.2 Aplicaciones de la IA en las funciones misionales

Las funciones misionales del Ministerio Público abarcan vigilancia superior, poder disciplinario y defensa de derechos humanos (C.P., 1991, arts. 277-284). La IA ofrece capacidades técnicas para fortalecer estas funciones, siempre que su implementación esté guiada por principios éticos sólidos.

La OCDE ha clasificado tareas de IA aplicables al servicio público, adoptadas en las Directrices Irlandesas (Department of Public Expenditure, NDP Delivery and Reform, 2024). El portafolio de ideas para IA del Ministerio Público identifica soluciones concretas basadas en evidencia internacional (Tejedor et al., 2025b). A continuación, se presenta cada categoría adaptada al contexto colombiano:

- **Reconocimiento:** Uso de IA para identificar y categorizar datos, por ejemplo, en aplicaciones relacionadas con el análisis y clasificación documental de quejas, peticiones y expedientes; OCR avanzado para digitalizar documentos históricos, e identificación de patrones en pruebas audiovisuales. El sistema Prometea, implementado en Argentina, ha demostrado capacidad para procesar grandes volúmenes documentales (Estevez et al., 2020), si se implementara un sistema así, se reducirían los tiempos de análisis de tutelas de meses a minutos en la Corte Constitucional colombiana.

- **Detección de eventos y anomalías:** Identificación de patrones atípicos o detección de conductas irregulares en contratación pública mediante análisis de bases SECOP; monitoreo de desinformación en medios digitales, y alertas tempranas de vulneraciones de DD. HH. El sistema BSEAN identifica desinformación multimodal con precisión del 83-87 % (Xing et al., 2025). El diagnóstico identificó la lucha contra desinformación como prioritaria (Tejedor et al., 2025a).

- **Proyección y predicción.** Uso de datos históricos para predecir resultados como, por ejemplo, predicción de carga de trabajo para asignación de recursos e identificación de riesgos institucionales. Se advierte que el uso predictivo en justicia presenta riesgos de sesgo y discriminación (Barocas y Selbst, 2016). El Marco ético advierte que «Sin embargo la inteligencia artificial también puede generar efectos negativos en materia de desinformación, sesgos, discriminación,

seguridad y afectaciones a la privacidad» (Dapre, 2021, p. 9). Requiere evaluaciones rigurosas de impacto en derechos humanos.

• **Personalización.** Adaptación de servicios a necesidades individuales, como en la atención personalizada al ciudadano y rutas procesales adaptativas según características del caso.

• **Soporte a la interacción.** Mejora de comunicación entre usuarios y sistemas como en el caso de chatbots y asistentes virtuales disponibles 24/7 en múltiples idiomas, y traducción y accesibilidad para comunidades indígenas y personas con discapacidad.

• **Optimización orientada a objetivos.** Soluciones eficientes para problemas complejos, la cual permite una asignación óptima de recursos entre regionales y planificar efectivamente las rutas para despliegues territoriales.

• **Razonamiento con estructuras de conocimiento.** Uso de bases de conocimiento para derivar conclusiones, cuyas aplicaciones más comunes abarcan el análisis jurídico asistido mediante corpus normativos, e investigación jurisprudencial con búsqueda semántica. Por ejemplo, Prometea utiliza técnicas de inteligencia artificial para identificar automáticamente información relevante en expedientes judiciales, permitiendo a los operadores del sistema acceder sin dificultad a datos relevantes y consistentes (Estevez et al., 2020).

• **Generación de contenido.** Creación de material nuevo y redacción asistida de comunicaciones rutinarias y respuestas a peticiones; síntesis automática de expedientes; campañas pedagógicas de DD. HH. El sistema CPE con RAG permite clasificación explicable de contenido (Willats et al., 2025). La Administración Federal Suiza reconoce que los modelos de lenguaje generativo presentan riesgos específicos en aplicaciones administrativas, por lo que requiere una evaluación cuidadosa de su uso, particularmente en relación con la trazabilidad de los datos de entrenamiento, técnicas y algoritmos empleados (Federal Chancellery FCh., 2025).

En este contexto, al vincular cualquier tecnología basada en IA con las funciones misionales, cada aplicación debe evaluarse por su contribución a: (i) vigilancia superior (análisis de gestión pública, detección de anomalías); (ii) función disciplinaria (procesamiento de expedientes, análisis jurisprudencial), y (iii) defensa de DD. HH

(atención accesible, detección temprana de vulneraciones, pedagogía). El diagnóstico reveló adopción mayor en operaciones internas que en toma de decisiones (Tejedor et al., 2025a), reflejando una implementación prudente.

3.3 Beneficios de la IA para el Ministerio Público

El proceso riguroso de implementar una IA responsable en el Ministerio Público genera beneficios en tres dimensiones (Department of Public Expenditure, NDP Delivery and Reform, 2024; OECD, 2019, 2024):

Productividad, eficiencia y eficacia

• **Automatización de tareas repetitivas:** la IA puede automatizar clasificación, transcripción y búsqueda, liberando a funcionarios para actividades de mayor valor que requieren juicio humano.

• **Procesamiento masivo:** los sistemas pueden procesar grandes volúmenes de documentos, analizar pruebas y detectar patrones a escala imposible para equipos humanos

• **Reducción de tiempos:** Prometea redujo entre 76 % y 78 % el tiempo de procesamiento de procesos judiciales en Argentina, con reducciones que van de 167 a 38 días para procesos de requerimiento a juicio (Estevez, E., et al., 2020).

• **Gestión del conocimiento:** sistemas capturan y estructuran conocimiento institucional mediante bases inteligentes y sistemas de recomendación.

Capacidad de respuesta (*responsiveness*)

• **Disponibilidad 24/7:** chatbots ofrecen atención continua, superando limitaciones de horarios y geografía.

• **Personalización y accesibilidad:** adaptación dinámica de interfaces según características del usuario (discapacidad, lengua nativa, nivel educativo), promoviendo accesibilidad universal.

• **Anticipación de necesidades:** sistemas predictivos identifican patrones emergentes en demandas ciudadanas, permitiendo intervenciones preventivas.

• **Participación informada:** la IA puede traducir lenguaje jurídico complejo a lenguaje sencillo, superando barreras para participación ciudadana.

Rendición de cuentas (*accountability*)

• **Detección de fraude:** algoritmos analizan datos para identificar patrones de corrupción, fraude o mala administración.

• **Trazabilidad:** sistemas apropiadamente diseñados generan registros (l*ogs*) de cada paso decisional, fortaleciendo rendición de cuentas (Directiva Conjunta 007, 2025).

• **Análisis de patrones institucionales:** la IA ayuda a ejercer vigilancia superior mediante análisis sistemático de indicadores de gestión.

• **Monitoreo de cumplimiento:** sistemas monitorean el cumplimiento de obligaciones legales, alertando sobre incumplimientos en protección de datos, transparencia o tiempos de respuesta.

El diagnóstico reveló adopción temprana identificada en un 22.6 % con proyectos implementados y 34.9 % en planificación (Tejedor et al., 2025a). Guedes y Oliveira (2024) señalan que la IA es fundamental para «elevar el rendimiento de servicios públicos», consistente con literatura que documenta ganancias de eficiencia del 20-40 % en procesos automatizados.

Los beneficios requieren de una gobernanza de datos robusta. La adopción efectiva no comienza con algoritmos, sino con alfabetización en IA y gestión de datos, capacidades humanas y gestión del cambio (dado que la asociación de la IA con la reducción laboral actúa como freno) (Tejedor et al., 2025a). Todo esto enmarcado bajo lineamientos claros, pues a menudo el marco ético gubernamental es percibido como una barrera que genera parálisis regulatoria.

3.4 Riesgos asociados al uso de la IA

Al igual que con toda tecnología, la adopción de la IA conlleva riesgos que pueden afectar derechos humanos, integridad institucional y legitimidad decisional. Una gobernanza responsable exige identificar, evaluar y mitigar sistemáticamente estos riesgos, adoptando un enfoque basado en riesgo (Parlamento Europeo y Consejo de la Unión Europea, 2024).

Enfoque basado en riesgo

El AI Act establece cuatro niveles de riesgo:

- **Riesgo inaceptable (prohibidos):** manipulación subliminal, explotación de vulnerabilidades, puntuación social e identificación biométrica indiscriminada (Parlamento Europeo y Consejo de la Unión Europea, 2024, art. 5).

- **Alto riesgo:** sistemas que pueden causar perjuicios significativos a salud, seguridad o derechos humanos (identificación biométrica, infraestructuras críticas, educación, empleo, servicios esenciales, aplicación de la ley, justicia, migración) (Parlamento Europeo y Consejo de la Unión Europea, 2024, anexo III). Muchas aplicaciones del Ministerio Público caerían aquí, exigiendo controles estrictos.

- **Riesgo limitado:** requieren transparencia (p. ej. chatbots deben informar que son automatizados).

- **Riesgo mínimo:** sin regulación específica más allá de normas generales.

Sesgos algorítmicos y discriminación

- **El riesgo más documentado es el sesgo algorítmico:** tendencia a producir resultados desfavorables para grupos basados en raza, género u origen socioeconómico (Barocas y Selbst, 2016).

- **Origen:** sesgo en datos de entrenamiento (datos históricos discriminatorios), definición de variables (código postal como proxy de raza) y etiquetado humano con prejuicios.

• **Impacto:** un sistema que prioriza investigaciones podría focalizar desproporcionadamente ciertos perfiles, violando igualdad e imparcialidad (C.P. 1991, art. 13).

• **Discriminación:** El Marco ético establece que «Los sistemas deben adoptar un enfoque de neutralidad de género y se debe garantizar que el parámetro de género no sea utilizado como factor de discriminación. (Dapre, 2021, p. 30). Este marco establece que la IA no debe generar resultados «que atenten contra el bienestar de un grupo especifico o que limiten los derechos de poblaciones históricamente marginadas» (p. 30).

Falta de transparencia y explicabilidad

• **Sistemas de «caja negra»:** aquellos modelos que producen resultados precisos mediante procesos que no son del todo claros o más conocidos como opacos (Arrieta et al., 2020).

• **Incompatibilidad con el debido proceso:** las decisiones deben estar fundamentadas en razones explícitas y verificables. Si una decisión se basa en recomendaciones inexplicables, se vulnera el debido proceso y la defensa (C.P., 1991, arts. 29, 228).

• **Obstáculo para rendición de cuentas:** sin explicabilidad es imposible determinar si errores provienen de defectos técnicos, datos inadecuados o sesgos. La Directiva Conjunta 007/2025 exige divulgación de información sobre sistemas, incluyendo criterios de decisión, explicación técnica, resumen de impacto algorítmico y medidas de mitigación (arts. 9, 13, 14).

Mal uso deliberado

Más allá de riesgos involuntarios, la IA puede usarse deliberadamente para fines ilícitos. Ya en 2017, un estudio sobre la propaganda computacional identificaba varias tácticas maliciosas (Neudert, Kollanyi, y Howard, 2017):

• **Propaganda computacional y «noticias basura»:** Se identificó la distribución a gran escala de noticias falsas, sensacionalistas, conspirativas o extremistas, diseñadas para parecer noticias reales y engañar a los votantes.

• **Manipulación de la opinión pública:** El estudio analiza el uso de cuentas automatizadas (bots) para amplificar artificialmente ciertos mensajes políticos,

atacar a figuras públicas e instituciones y simular un apoyo popular inexistente, con el fin de influir en la conversación política.

• **Acoso y discurso de odio:** Se menciona que los bots son herramientas utilizadas para llevar a cabo actividades maliciosas como el acoso (*harassment*) y la difusión de discursos de odio a gran velocidad.

• **Falsificación de pruebas (*deepfakes*):** síntesis de audio/video crea evidencia falsa realista. La Ley 2502 de 2025 tipifica la falsedad personal mediante IA.

Para el Ministerio Público, estos riesgos son dobles: por un lado, debe proteger a las instituciones de estos ataques y, por otro, investigar las conductas reprochables de quienes ejercen funciones públicas relacionadas con el mal uso de la IA. Aunque el estudio es anterior a la popularización de los deepfakes y los LLM, describe las bases de la manipulación digital que hoy se ha vuelto más sofisticada.

Riesgos técnicos

• **Alucinaciones:** Los LLM generan información falsa con alta confianza, peligrosas en contextos jurídicos.

• **Irreproducibilidad:** los modelos generativos producen resultados diferentes con la misma entrada.

• **Vulnerabilidades:** los sistemas pueden sufrir ataques adversariales (manipulaciones de entradas o jaqueo de prompts).

• **Dependencia de proveedores:** la baja experticia técnica interna genera dependencia de proveedores externos con riesgos de falta de control, opacidad y discontinuidad.

Vulneración de derechos humanos

• **La operación de IA implica tratamiento de datos personales sensibles.** La Ley 1581 de 2012 y el Decreto de la Presidencia 1377 de 2013 requieren el consentimiento informado, finalidad específica, proporcionalidad y seguridad.

• **Los riesgos incluyen filtraciones, uso no autorizado y reidentificación de datos supuestamente anonimizados.** La Directiva 007 de 2025 reconoce el derecho a comprender decisiones automatizadas y objetarlas (arts. 2, 10). La clasificación incorrecta o falla en detectar riesgo puede generar denegación de justicia, vulnerando el acceso a la justicia y su administración efectiva (C.P., 1991, arts. 228, 229).

Deshumanización del servicio público

• **Pérdida de contacto humano:** en defensa de DD. HH., el contacto directo es esencial. La automatización excesiva genera servicios eficientes pero deshumanizados (Department of Public Expenditure, NDP Delivery and Reform, 2024, p. 20).

• **Inflexibilidad:** los sistemas operan dentro de parámetros de entrenamiento. Casos excepcionales requieren juicio prudencial del que los sistemas carecen.

• **Erosión de capacidades:** la delegación excesiva atrofia las capacidades analíticas, generando una dependencia tecnológica peligrosa.

Barreras organizacionales

El diagnóstico identificó: (i) baja experticia técnica; (ii) la ausencia de un marco ético como barrera (iii) resistencia al cambio; (iv) cultura conservadora, y (v) ausencia de liderazgo comprometido e incentivos configura un ecosistema hostil que muestra los retos que deben asumirse para una adopción responsable de la IA en el Ministerio Público (Tejedor et al., 2025a).

Necesidad de gestión proactiva

Los riesgos no son razones para rechazar la IA, sino para adoptarla responsablemente. El AI Act exige para situaciones de alto riesgo una gestión de riesgos durante el ciclo de vida, datos de calidad, documentación técnica, *logs* para trazabilidad, transparencia, supervisión humana, y garantías de precisión y ciberseguridad (Parlamento Europeo y Consejo de la Unión Europea, 2024, arts. 9-15). El Marco ético recomienda evaluaciones de impacto sobre derechos humanos (EIDH) antes de desplegar IA en contextos sensibles (Dapre, 2021). La Directiva 007 de 2025 exige análisis de impacto y medidas de mitigación (art. 14).

Síntesis del Capítulo

En este capítulo se establecieron las bases conceptuales para comprender la IA en el Ministerio Público. Se definió con precisión qué es la IA como campo de estudio, distinguiendo sistemas basados en reglas, *machine learning, deep learning* e IA generativa. Se identificaron ocho categorías de aplicaciones concretas en funciones misionales, ejemplificadas por soluciones como Prometea, BSEAN y CPE con RAG documentadas en el Portafolio de ideas.

Los beneficios incluyen un aumento en la productividad mediante automatización, y responsabilidad mediante servicios personalizados y mediante detección de irregularidades. El diagnóstico confirma apertura hacia la IA y valoración positiva de beneficios.

Los riesgos son sustanciales, entre ellos los sesgos que perpetúan discriminación, la falta de transparencia que dificulta el debido proceso, el mal uso para desinformación, vulnerabilidades técnicas, afectación a privacidad y deshumanización. El diagnóstico identificó barreras organizacionales críticas, especialmente baja experticia técnica y cultura organizacional. El desafío incluye tanto elegir entre innovación y protección de derechos, como construir una gobernanza que permita el desarrollo de estas dos dimensiones.

Los capítulos subsiguientes desarrollarán principios éticos (capítulo 4), marco de decisión (capítulo 5), herramientas operativas (capítulos 6-7) y mecanismos de control y supervisión (capítulo 8). El Ministerio Público tiene la oportunidad de posicionarse como líder en adopción responsable de la IA en la región, demostrando que es posible combinar innovación con protección de derechos, eficiencia con transparencia, y automatización con juicio humano reflexivo. Este marco busca compartir y promover esa visión altamente responsable.

4. Principios éticos para el uso responsable de la IA

El corazón de este marco de gobernanza reside en la adopción y aplicación rigurosa de ocho principios fundamentales identificados en los marcos normativos nacionales e internacionales sobre la IA. Estos principios se establecen como una serie de mandatos operativos que deben guiar cada etapa del ciclo de vida de los sistemas de IA que se pretendan adoptar en el Ministerio Público. Cada principio se presenta con su fundamento, directrices para su implementación, mecanismos de control para su verificación y la normativa que lo sustenta.

4.1 Principio 1: Supervisión y control humano

Los sistemas de inteligencia artificial son herramientas que potencian y amplifican las capacidades humanas, y en su etapa de evolución actual no reemplazan la agencia humana en lo concerniente a decisiones críticas. Se debe garantizar en todo momento una supervisión humana significativa, asegurando que la responsabilidad final sobre las decisiones que afectan los derechos de las personas recaiga siempre en un servidor público competente.

Este principio se alinea con la recomendación del Marco ético para la IA en Colombia (Dapre, 2021), que enfatiza la necesidad de «control humano de las decisiones propias de un sistema de inteligencia artificial» (p. 27), especialmente en las etapas iniciales de implementación.

Directrices

• **Intervención humana efectiva:** se deben implementar arquitecturas de sistema que permitan una intervención humana adecuada según el nivel de riesgo. Para sistemas de alto riesgo, se privilegiarán modelos de «humano en el circuito» (*human-in-the-loop*), donde cada decisión del sistema debe ser validada por una persona. Para riesgos moderados, se podrán usar modelos de «humano sobre el circuito» (*human-over-the-loop*), donde una persona puede intervenir y corregir el sistema en cualquier momento.

• **Prohibición de automatización completa en decisiones críticas:** queda prohibido delegar de forma completamente autónoma en un sistema de IA la toma de decisiones que impliquen la apertura de una investigación formal, la formulación de pliegos de cargos, la imposición de sanciones disciplinarias, o cualquier otra determinación que afecte sustancialmente los derechos humanos.

• **Roles y responsabilidades claras:** es necesario definir y documentar de manera explícita quién es el servidor público responsable de la supervisión de cada sistema de IA, así como los protocolos de escalamiento en caso de detectarse anomalías o resultados inesperados.

• **Capacitación para la supervisión:** los funcionarios encargados de supervisar sistemas de IA recibirán formación específica no solo sobre el funcionamiento del sistema, sino también sobre sus limitaciones, potenciales sesgos y los riesgos asociados.

Mecanismos de control

• **Comités de supervisión:** es imprescindible la creación de comités interdisciplinarios para revisar periódicamente el desempeño de los sistemas de IA de alto riesgo y el cumplimiento de los protocolos de supervisión humana.

• **Bitácoras de decisión:** todos los sistemas de IA que apoyen la toma de decisiones deben registrar tanto la recomendación del sistema como la decisión final tomada por el supervisor humano, incluyendo una justificación en caso de que la decisión humana difiera de la sugerencia del algoritmo. Estas bitácoras serán auditables.

• **Evaluaciones de impacto algorítmico:** antes del despliegue de un sistema, se debe evaluar el nivel de supervisión humana requerido en función del impacto potencial del sistema sobre los derechos de las personas.

Tabla 1. Normativa aplicable al principio de supervisión y control humano

Instrumento	Artículo / Sección relevante
Marco ético para la IA en Colombia (Dapre, 2021)	Principio 3: «Control humano de las decisiones propias de un sistema de inteligencia artificial».
Directrices éticas para una IA fiable (UE). (High-Level Expert Group on Artificial Intelligence, 2019)	Requisito 1: «Acción y supervisión humanas».
Reglamento 2024/1689 (AI Act). (Parlamento Europeo y Consejo de la Unión Europea, 2024)	Artículo 14: Requisitos de supervisión humana para sistemas de IA de alto riesgo.
Principios de IA de la OCDE (2019, 2024)	Principio 1.2: Durante el ciclo de vida de la IA, es vital respetar los derechos humanos, los valores democráticos y mitigar la desinformación. Para lograrlo, resulta imprescindible priorizar la agencia y supervisión humana mediante salvaguardas efectivas, asegurando así que las personas mantengan siempre el control, la autonomía y la capacidad de decisión final sobre estos sistemas.

Fuente. Elaboración de los autores.

4.2 Principio 2: Robustez técnica y seguridad

La confianza en los sistemas de inteligencia artificial depende directamente de su fiabilidad, precisión y resiliencia. Los sistemas de IA del Ministerio Público deben ser técnicamente robustos y seguros a lo largo de todo su ciclo de vida, garantizando que no causen daños involuntarios y que estén protegidos frente a vulnerabilidades y usos malintencionados que puedan comprometer la integridad de sus operaciones y la seguridad de los ciudadanos.

Este principio se basa en el principio 4, «seguridad», del Marco ético para la IA en Colombia (Dapre, 2021), que establece que «los sistemas de inteligencia artificial no deben generar afectaciones a la integridad y salud física y mental de los seres humanos con los que interactúan» (p. 28) y deben garantizar la confidencialidad e integridad de la información.

Directrices

- **Precisión, fiabilidad y reproducibilidad:** los sistemas de IA tienen que alcanzar un nivel de precisión adecuado y validado para casos de uso específico. Sus

resultados deben ser fiables y reproducibles bajo las mismas condiciones, y deben tener la capacidad de conocer y documentar sus márgenes de error.

• **Resiliencia y planes de contingencia:** los sistemas tienen que ser resilientes a errores y fallos. Se deben desarrollar planes de contingencia claros que definan los pasos a seguir en caso de mal funcionamiento, resultados inesperados o caídas del sistema, asegurando la continuidad de las funciones críticas del Ministerio Público.

• **Seguridad en todo el ciclo de vida:** la seguridad será un componente integral desde la fase de diseño (seguridad por diseño) y no un añadido posterior. Esto incluye la protección de los datos de entrenamiento, el modelo algorítmico y la infraestructura que lo soporta contra accesos no autorizados, manipulación o robo.

• **Protección contra ataques adversarios:** se implementarán medidas técnicas para proteger los sistemas contra ataques específicos de la IA, como el jaqueo de prompts (*promting hacking*), el envenenamiento de datos (*data poisoning*), la evasión de modelos (*model evasion*) o la extracción de información confidencial del modelo.

Mecanismos de control

• **Auditorías técnicas y de seguridad:** es necesaria la realización de auditorías periódicas, tanto internas como por parte de terceros independientes, para evaluar la robustez, precisión y seguridad del sistema, incluyendo pruebas de penetración y análisis de vulnerabilidades.

• **Entornos de prueba y validación (*sandbox*):** antes de su despliegue en un entorno de producción, todo sistema de IA debe ser probado exhaustivamente en un entorno de pruebas controlado (*sandbox*) que simule las condiciones reales de operación.

• **Monitoreo continuo del desempeño:** implementación de herramientas para el monitoreo en tiempo real del rendimiento y la precisión del sistema, con alertas automáticas en caso de degradación del desempeño o comportamiento anómalo.

• **Protocolos de gestión de incidentes de seguridad:** establecimiento de un protocolo claro para la notificación, gestión y resolución de incidentes de seguridad relacionados con los sistemas de IA, en línea con la política de seguridad de la información de la Procuraduría General de la Nación (PGN).

Tabla 2. Normativa aplicable al principio de robustez técnica y seguridad

Instrumento	Artículo / Sección relevante
Marco ético para la IA en Colombia (Dapre, 2021)	Principio 4: «Seguridad».
Reglamento 2024/1689 (AI Act). (Parlamento Europeo y Consejo de la Unión Europea, 2024)	Artículo 15: «Precisión, solidez y ciberseguridad».
Directrices éticas para una IA fiable (UE). (High-Level Expert Group on Artificial Intelligence, 2019)	Requisito 2: «Solidez técnica y seguridad».
Política de Seguridad de la Información Resolución 138 del 27 de junio de 2025 (PGN, 2025) (PGN, 2025)	Establece el marco institucional para la gestión de la seguridad de la información.

Fuente. Elaboración de los autores.

4.3 Principio 3: Privacidad y gobernanza de datos

El derecho fundamental a la intimidad y al *habeas data* (C.P., art. 15) es un pilar de la democracia y debe ser protegido con el máximo rigor en la era digital. Los sistemas de inteligencia artificial, por su naturaleza intensiva en el uso de datos, presentan altos riesgos para la privacidad. Por tanto, su diseño e implementación en el Ministerio Público deben estar regidos por una gobernanza de datos robusta y el cumplimiento estricto de la normativa de protección de datos personales.

Este principio se fundamenta en la Ley 1581 de 2012 y su decreto reglamentario, así como en el principio 2 del Marco ético para la IA en Colombia (Dapre, 2021), «privacidad», que exige la implementación de herramientas de gestión de riesgo para la privacidad y la realización de evaluaciones de impacto.

Directrices

• **Privacidad por diseño y por defecto:** la protección de la privacidad es un requisito esencial que debe resguardarse desde las primeras etapas de diseño de cualquier

sistema de IA. Por defecto, se deben aplicar las configuraciones más protectoras de la privacidad sin que el usuario tenga que realizar ninguna acción.

• **Minimización de datos:** solo se recolectarán, usarán y conservarán los datos que sean estrictamente necesarios, pertinentes y adecuados para la finalidad legítima para la cual fue diseñado el sistema de IA. Se prohíbe la recolección masiva e indiscriminada de datos.

• **Calidad y exactitud de los datos:** se tomarán todas las medidas razonables para asegurar que los datos utilizados para entrenar y operar los sistemas de IA sean precisos, completos y actualizados, con el fin de evitar decisiones erróneas basadas en información incorrecta.

• **Finalidad y consentimiento informado:** el uso de datos personales se limitará a la finalidad específica para la cual fueron recolectados, de acuerdo con las funciones del Ministerio Público. Cuando sea legalmente requerido, se deberá obtener el consentimiento libre, previo, expreso e informado del titular de los datos, explicándole de manera clara cómo serán utilizados por el sistema de IA.

• **Técnicas de mejora de la privacidad (PETs):** se promoverá el uso de técnicas como la anonimización, la seudonimización y el cifrado para reducir los riesgos de identificación de las personas y proteger la confidencialidad de la información.

Mecanismos de control

• **Evaluaciones de impacto en la protección de datos (EIPD):** será obligatorio realizar una EIPD antes de implementar cualquier sistema de IA que implique el tratamiento de datos personales a gran escala o de datos sensibles. Esta evaluación identificará los riesgos para la privacidad y definirá las medidas para mitigarlos.
• **Política de gobernanza de datos institucional:** se debe fortalecer y aplicar la Política de Gobierno de Datos de la PGN, asegurando que incluya un capítulo específico sobre el tratamiento de datos en sistemas de IA, definiendo roles, responsabilidades y procedimientos claros.

• **Oficial de protección de datos (DPO):** el DPO, más conocido como oficial de seguridad de la información de la entidad (o quien haga sus veces), tendrá un rol central en la supervisión del cumplimiento de la normativa de protección de datos en todos los proyectos de IA, debiendo ser consultado desde las fases iniciales (PGN, 2025).

• **Auditorías de datos:** realización de auditorías periódicas para verificar que el tratamiento de datos por parte de los sistemas de IA se ajusta a los principios de minimización, finalidad y calidad, y que las medidas de seguridad son efectivas.

Tabla 3. Normativa aplicable al principio de privacidad y gobernanza de datos

Instrumento	Artículo / Sección relevante
Constitución Política de Colombia (1991)	Artículo 15: derecho a la intimidad personal y familiar y al buen nombre, y derecho al habeas data.
Ley 1581 de 2012 Régimen General de Protección de Datos Personales.	Artículo 4, define los principios de legalidad, finalidad, libertad, veracidad, transparencia, acceso, seguridad y confidencialidad; Artículo 5, clasifica y define los datos sensibles; Artículo 6, regula el tratamiento de dichos datos; Artículo 7, prioriza los derechos de NNA; el Artículo 17, deberes de los responsables del tratamiento; el Artículo 18, enumera las obligaciones de los encargados; Artículo 25, crea el Registro Nacional de Bases de Datos, y el Artículo 26, prohíbe la transferencia internacional de información a países que no garanticen niveles adecuados de protección.
Decreto 1377 de 2013 de la Presidencia de la República	Artículo 5, prohíbe el uso de medios engañosos para la recolección de datos; el Artículo 7, obliga a los responsables a conservar la prueba de la autorización; el Artículo 10, define quiénes están legitimados para ejercer los derechos de los titulares; Artículo 13, exige la creación de políticas de tratamiento de información; Artículo 14, regula la implementación del aviso de privacidad; Artículo 15, detalla el contenido mínimo de dicho aviso; Artículo 21, establece requisitos para el tratamiento de datos de menores; Artículo 24, ordena la formalización de contratos de transmisión entre responsables y encargados, y el Artículo 26, consagra el principio de responsabilidad demostrada .
Marco ético para la IA en Colombia (Dapre, 2021)	Principio 2: «Privacidad».
Reglamento 2024/1689 (AI Act). (Parlamento Europeo y Consejo de la Unión Europea, 2024)	Artículo 10: «Datos y Gobernanza de Datos».
Política de Gobierno de Datos y funciones del Comité de Seguridad de la Información (PGN, 2024a; PGN, 2025)	Regula el ciclo de vida de la información bajo niveles de gobernanza estratégico, táctico y operativo. Su enfoque garantiza la seguridad, calidad y disponibilidad de los datos mediante la asignación de roles claros, el cumplimiento de estándares de interoperabilidad y el uso de indicadores para medir la eficiencia institucional (pp. 2, 17)

Fuente. Elaboración de los autores.

4.4 Principio 4: Transparencia y explicabilidad

La confianza ciudadana en las instituciones públicas se basa en la transparencia de sus actuaciones. En el contexto de la IA, este principio se materializa en la transparencia algorítmica, reconocida por la Corte Constitucional en la Sentencia T-067 de 2025 como un pilar del derecho de acceso a la información (C.P., art. 74). Las decisiones tomadas o apoyadas por sistemas de IA deben ser comprensibles y rastreables, permitiendo a los ciudadanos y a los propios funcionarios entender su lógica y, si es necesario, controvertirlas.

Este principio es el eje central de la Directiva Conjunta 007 de 2025, que establece estándares sobre transparencia algorítmica para el Estado colombiano, y se refuerza en el Principio 1 del Marco ético para la IA en Colombia (Dapre, 2021), «transparencia y explicación».

Directrices

- **Transparencia proactiva:** el Ministerio Público divulgará y mantendrá actualizado un inventario de los sistemas y soluciones de IA que utiliza (registro de sistemas algorítmicos), describiendo su finalidad, los datos que utiliza, su lógica general de funcionamiento y una evaluación de su impacto, en cumplimiento de la Ley 1712 de 2014 y la Directiva 007 de 2025.

- **Documentación exhaustiva:** todo sistema de IA contará con una documentación técnica y funcional completa que detalle su arquitectura, los conjuntos de datos de entrenamiento, los parámetros del modelo, las pruebas de validación y los resultados de las evaluaciones de rendimiento y sesgo.

- **Explicabilidad adaptada a la audiencia:** se desarrollarán mecanismos para proporcionar explicaciones sobre las decisiones algorítmicas que sean comprensibles para distintos tipos de público. Esto incluye explicaciones técnicas para auditores y expertos, y explicaciones en lenguaje claro y sencillo para los ciudadanos afectados por una decisión.

- **Trazabilidad de las decisiones:** los sistemas registrarán de forma inalterable cada operación y decisión, permitiendo reconstruir el proceso que llevó a un resultado específico. Esta trazabilidad es esencial para la auditoría y la rendición de cuentas.

Mecanismos de control

- **Registro público de sistemas algorítmicos:** se requiere la creación y mantenimiento de un registro público y de fácil acceso, que contenga la información no reservada de los sistemas de IA en uso en la entidad.

- **Informes periódicos de transparencia:** es necesario promover la publicación de informes periódicos sobre el uso de la IA, incluyendo estadísticas de rendimiento, resultados de auditorías de sesgo y las medidas tomadas para corregir errores o mitigar riesgos.

- **Derecho a la explicación:** se requiere establecer un procedimiento formal para que los ciudadanos puedan solicitar y recibir una explicación significativa sobre una decisión automatizada que les afecte.

- **Auditorías de transparencia y explicabilidad:** es necesaria la realización de auditorías para verificar que la información proporcionada sobre los sistemas es precisa y completa, y que los mecanismos de explicación son efectivos.

Tabla 4. Normativa aplicable al principio de transparencia y explicabilidad

Instrumento	Artículo / Sección relevante
Sentencia T-067 de 2025 de la Corte Constitucional	Define la transparencia algorítmica como derecho fundamental.
Directiva Conjunta 007 de 2025	Establece estándares mínimos de transparencia algorítmica para el Estado.
Ley 1712 de 2014	Ley de Transparencia y del Derecho de Acceso a la Información Pública Nacional.
Marco ético para la IA en Colombia (Dapre, 2021)	Principio 1: «Transparencia y Explicación».
Reglamento 2024/1689 (AI Act). (Parlamento Europeo y Consejo de la Unión Europea, 2024)	Artículo 13: «Transparencia y comunicación de información a los responsables del despliegue».

Fuente. Elaboración de los autores.

4.5 Principio 5: Equidad, no discriminación y justicia

El derecho a la igualdad, consagrado en el artículo 13 de la Constitución Política, es una piedra angular del Estado social de derecho. Si los sistemas de inteligencia artificial no se diseñan y supervisan cuidadosamente, tienen la capacidad de perpetuar e incluso amplificar sesgos históricos y sociales presentes en los datos, conduciendo a resultados discriminatorios. Por lo tanto, es un imperativo ético y legal asegurar que la IA en el Ministerio Público promueva la equidad, la no discriminación y la justicia para todas las personas.

Este principio se inspira en los principios 6 y 7 del Marco ético para la IA en Colombia sobre «No discriminación e Inclusión» (Dapre, 2021), los cuales exigen un análisis constante del impacto de la IA para evitar efectos discriminatorios y promover la equidad en el acceso a los sistemas de IA.

Directrices

- **Evaluación y mitigación de sesgos:** es necesario desarrollar análisis rigurosos de los conjuntos de datos de entrenamiento para identificar y mitigar posibles sesgos (políticos, de credo, género, raza, origen socioeconómico, etc.). Asimismo, la evaluación de los modelos algorítmicos es necesaria para detectar y corregir cualquier comportamiento discriminatorio antes y durante su despliegue.

- **Diseño inclusivo y participativo:** en el diseño de los sistemas de IA, especialmente aquellos que interactuarán con la ciudadanía o afectarán a grupos vulnerables, se fomentará la participación de equipos multidisciplinarios y de representantes de diversas poblaciones para asegurar que se tengan en cuenta múltiples perspectivas.

- **Monitoreo de impactos diferenciales:** una vez en operación, los sistemas serán monitoreados continuamente para evaluar si sus resultados están teniendo un impacto desproporcionado o adverso en determinados grupos poblacionales. Se establecerán umbrales de equidad y alertas para detectar estas situaciones.

- **Justicia procesal:** el uso de la IA no debe menoscabar las garantías del debido proceso. Las personas afectadas por decisiones algorítmicas tendrán el derecho a ser

informadas, a controvertir la decisión y a obtener una revisión humana significativa.

Mecanismos de control

- **Evaluaciones de impacto en la equidad (EIE):** antes de la implementación, se realizará una EIE para analizar los posibles efectos del sistema de IA sobre la equidad y la no discriminación, y para definir medidas de mitigación.

- **Auditorías de sesgo algorítmico:** contratación de auditorías externas e independientes para examinar los algoritmos y los datos en busca de sesgos ocultos y para validar la efectividad de las medidas de mitigación implementadas.

- **Métricas de equidad:** definición y seguimiento de métricas cuantitativas para medir la equidad del sistema, como la paridad demográfica o la igualdad de oportunidades, según el contexto de la aplicación.

- **Canales de denuncia de discriminación:** habilitación de canales claros y accesibles para que los ciudadanos puedan reportar posibles casos de discriminación por parte de un sistema de IA.

Tabla 5. Normativa aplicable al principio de equidad, no discriminación y justicia

Instrumento	Artículo / Sección relevante
Constitución Política de Colombia (1991)	Artículo 13: derecho a la igualdad y prohibición de la discriminación.
Marco ético para la IA en Colombia (Dapre, 2021)	Principio 6: «No discriminación». Principio 7: «Inclusión».
Directrices éticas para una IA fiable (UE). (High-Level Expert Group on Artificial Intelligence, 2019)	Requisito 5: «Diversidad, no discriminación y equidad».
Documento CONPES 4144 de 2025 (CONPES [DNP], 2025).	Aborda los desafíos éticos y sociales asociados con el desarrollo y la adopción de la IA.

Fuente. Elaboración de los autores.

4.6 Principio 6: Bienestar social y ambiental

La tecnología al servicio del Estado debe tener como fin último la promoción del bien común, el desarrollo sostenible y el mejoramiento de la calidad de vida de la población. Los sistemas de inteligencia artificial implementados por el Ministerio Público serán evaluados tanto por su eficiencia técnica como por su contribución positiva a la sociedad y su respeto por el medio ambiente, en armonía con los fines esenciales del Estado social de derecho.

Este principio se alinea con la visión de los Principios de IA de la OCDE (2019, 2024), que buscan un crecimiento inclusivo y un desarrollo sostenible, y con el principio 6 de las directrices de la UE, «bienestar social y medioambiental».

Directrices

• **Evaluación del impacto social:** antes de su implementación, se evaluará el impacto social más amplio del sistema de IA, considerando sus efectos potenciales sobre el empleo, la cohesión social, el acceso a los servicios públicos y el ejercicio de los derechos democráticos.

• **Sostenibilidad ambiental:** se realizarán evaluaciones del impacto ambiental del ciclo de vida de los sistemas de IA, desde el consumo energético de los centros de datos para el entrenamiento de los modelos hasta la gestión de los residuos electrónicos. Se preferirán aquellas soluciones que sean energéticamente eficientes.

• **Promoción del interés público:** el desarrollo y uso de la IA en el Ministerio Público estará orientado a resolver problemas públicos relevantes y a generar valor social, como la lucha contra la corrupción, la protección de poblaciones vulnerables o la agilización de la justicia.

• **Participación ciudadana:** se fomentarán espacios de diálogo y participación ciudadana en la definición de las prioridades para el uso de la IA y en la evaluación de su impacto, asegurando que la tecnología responda a las necesidades y valores de la sociedad.

Mecanismos de control

• **Evaluaciones de impacto social y ambiental:** inclusión de capítulos específicos sobre el impacto social y ambiental en las evaluaciones previas al despliegue de sistemas de IA de alto impacto.

• **Consultas públicas y diálogos con la sociedad civil:** realización de consultas públicas para proyectos de IA de gran envergadura y establecimiento de mesas de trabajo permanentes con la academia, la sociedad civil y el sector privado.

• **Indicadores de impacto positivo:** definición de indicadores para medir la contribución del sistema de IA a metas sociales específicas, como la reducción de la impunidad o la mejora en la protección de líderes sociales (World Justice Project, 2024).

• **Informes de sostenibilidad:** inclusión de información sobre el impacto social y ambiental de la IA en los informes de gestión y sostenibilidad de la entidad.

Tabla 6. Normativa aplicable al principio de bienestar social y ambiental

Instrumento	Artículo / Sección relevante
Constitución Política de Colombia (1991)	Artículos 79, 80 (derecho a un ambiente sano y la promoción del desarrollo sostenible), y fines esenciales del Estado (art. 2).
Marco ético para la IA en Colombia (Dapre, 2021)	Principio 9: «Beneficio Social».
Directrices éticas para una IA fiable (UE). (High-Level Expert Group on Artificial Intelligence, 2019)	Requisito 6: «Bienestar social y ambiental».
Objetivos de Desarrollo Sostenible (ODS)	Marco global para el desarrollo sostenible, en particular ODS 16 (Paz, Justicia e Instituciones Sólidas).

Fuente. Elaboración de los autores.

4.7 Principio 7: Rendición de cuentas y responsabilidad

La delegación de tareas en sistemas de inteligencia artificial no puede suponer una dilución de la responsabilidad. Es necesario establecer protocolos claros de rendición de cuentas, donde sea posible identificar a las personas y entidades responsables del diseño, implementación y supervisión del sistema de IA. Se establecerán los mecanismos efectivos para la reparación de los daños que un sistema pueda causar, en línea con el principio de responsabilidad de la función pública (C.P., arts. 6 y 90).

Este principio se basa en el principio 7 de las Directrices de la UE, «rendición de cuentas», y se refuerza en la normativa disciplinaria colombiana, que exige responsabilidad individual de los servidores públicos por sus acciones y omisiones relacionadas con el correcto funcionamiento de los sistemas de IA.

Directrices

- **Asignación explícita de responsabilidad:** para cada sistema de IA, se designará formalmente a un «propietario» del sistema, que será un servidor público de nivel directivo responsable de su correcto funcionamiento y de su alineación con este marco ético.

- **Trazabilidad para la rendición de cuentas:** la trazabilidad de las operaciones del sistema (*logs*, bitácoras de decisión) es un requisito indispensable para poder investigar incidentes, auditar el sistema y determinar responsabilidades.

- **Mecanismos de recurso y reparación:** es necesario definir y establecer canales claros y efectivos para que los ciudadanos puedan presentar quejas, solicitar la revisión de una decisión automatizada y, en caso de demostrarse un daño, acceder a mecanismos de reparación.

- **Responsabilidad compartida en la cadena de suministro:** en el caso de sistemas de IA adquiridos a terceros, los contratos incorporarán cláusulas claras de responsabilidad y obligaciones de transparencia y colaboración por parte de los proveedores en caso de incidentes o auditorías.

Mecanismos de control

- **Matriz de responsabilidades:** creación de una matriz de roles y responsabilidades (matriz *RACI)* para cada sistema de IA, que defina claramente quién es responsable, quién rinde cuentas, quién es consultado y quién es informado.

- **Protocolos de auditoría forense:** desarrollo de protocolos para realizar auditorías forenses en caso de incidentes graves, que permitan reconstruir los hechos y determinar las causas del fallo del sistema.

- **Informes de rendición de cuentas:** inclusión de un capítulo sobre el desempeño y la supervisión de los sistemas de IA en los informes de rendición de cuentas que la entidad presenta a la ciudadanía y a los organismos de control.

- **Simulacros de incidentes:** realización de simulacros periódicos para probar la efectividad de los protocolos de gestión de incidentes y la claridad de la cadena de responsabilidad.

Tabla 7. Normativa aplicable al principio de rendición de cuentas y responsabilidad

Instrumento	Artículo / Sección relevante
Constitución Política de Colombia (1991)	Artículo 6 (responsabilidad de los servidores públicos) y artículo 90 (responsabilidad patrimonial del Estado).
Ley 1952 de 2019	Código General Disciplinario, que establece las faltas y sanciones de los servidores públicos.
Directrices éticas para una IA fiable (UE). (High-Level Expert Group on Artificial Intelligence, 2019)	Requisito 7: «Rendición de cuentas».
Principios de IA de la OCDE (2019, 2024)	Principio 1.5: «Responsabilidad».

Fuente. Elaboración de los autores.

4.8 Principio 8: Prevalencia de los derechos de niños, niñas y adolescentes

Los sistemas de inteligencia artificial tienen que reconocer, respetar y privilegiar de manera especial y reforzada los derechos humanos de niños, niñas y adolescentes (NNA). En ningún caso está justificada la implementación de un sistema de IA que vaya en detrimento del interés superior del menor, principio constitucional establecido que reconoce la prevalencia de los derechos de la niñez sobre los derechos de los demás. Los NNA constituyen un grupo poblacional en situación de especial vulnerabilidad debido a su desarrollo físico, psicológico y emocional en curso, lo que exige salvaguardas adicionales más allá de las aplicables a la población adulta.

Este principio se fundamenta en el artículo 44 de la Constitución Política, el Código de la Infancia y la Adolescencia (Ley 1098 de 2006), el Marco ético para la IA en Colombia (Dapre, 2021) que establece este principio como uno de los nueve rectores, la adopción de la Convención sobre los Derechos del Niño de las Naciones Unidas (Ley 12 de 1991), y las Observaciones Generales del Comité de los Derechos del Niño (2021) sobre entornos digitales. Además, se alinea con el enfoque de la Unión Europea que considera sistemas de IA que afecten a menores como de «alto riesgo» que requieren requisitos estrictos (Parlamento Europeo y Consejo de la Unión Europea, AI Act, 2024, anexo III, numeral 3).

Directrices

- **Evaluación específica del impacto sobre NNA:** todo sistema de IA que pueda afectar directa o indirectamente a niños, niñas y adolescentes debe ser objeto de una evaluación de impacto sobre los derechos de la niñez (EIDN), que analice riesgos específicos como exposición a contenidos inapropiados, manipulación comercial, perfilamiento invasivo, afectación al desarrollo cognitivo y emocional, ciberacoso, discriminación algorítmica, y cualquier amenaza a su dignidad, integridad o desarrollo armónico.

- **Prohibición de uso perjudicial:** se prohíbe expresamente el desarrollo, adquisición o implementación de sistemas de IA que puedan causar daño físico, psicológico o emocional a NNA, incluyendo: sistemas de manipulación del

comportamiento infantil con fines comerciales o políticos; algoritmos de perfilamiento que clasifiquen o etiqueten a menores de manera estigmatizante; sistemas de vigilancia que vulneren su privacidad sin justificación imperiosa de protección, y cualquier tecnología que explote vulnerabilidades propias de la edad o el desarrollo infantil.

• **Datos de NNA - restricciones reforzadas:** los datos personales de niños, niñas y adolescentes únicamente pueden ser recolectados, procesados y utilizados cuando sea estrictamente necesario para actividades que se relacionen directamente con su interés superior y que tengan un impacto exclusivamente positivo sobre ellos. Se prohíbe el uso de datos de NNA para fines comerciales, publicitarios o de marketing automatizado. Los sistemas deben implementar verificación de edad para prevenir el tratamiento inadecuado de datos de menores, y cuando sea técnicamente posible, utilizar técnicas de privacidad por diseño como anonimización, seudonimización y minimización de datos. El consentimiento para el tratamiento de datos de menores de 14 años debe ser otorgado por los representantes legales, según lo establecido por la Ley 1581 de 2012 y la doctrina de la Superintendencia de Industria y Comercio.

• **Diseño comprensible y apropiado para la edad:** los algoritmos y las interfaces de los sistemas de IA que interactúen con NNA deben ser diseñados de manera que sean comprensibles para ellos, adaptados a su nivel de desarrollo cognitivo. La información sobre el funcionamiento del sistema, sus propósitos y sus consecuencias debe ser comunicada en lenguaje claro, sencillo y apropiado para diferentes grupos de edad. Los sistemas deben evitar el uso de «patrones oscuros» (*dark patterns*) que manipulen o confundan a los menores sobre las consecuencias de sus acciones en entornos digitales.

• **Protección activa contra riesgos digitales:** los sistemas de IA deben incorporar salvaguardas técnicas y organizativas para proteger a NNA de riesgos específicos del entorno digital, incluyendo la detección y prevención de ciberacoso (cyberbullying) mediante análisis de patrones de comunicación hostil; identificación proactiva de contenido inapropiado o dañino (violencia, contenido sexual explícito, apología de conductas de riesgo); alertas ante signos de explotación, captación para fines ilícitos o situaciones de vulneración de derechos, y mecanismos de reporte fácil y confidencial para que NNA o sus cuidadores puedan denunciar situaciones problemáticas.

• **Participación significativa de NNA:** cuando se diseñen o implementen sistemas de IA que afecten a niños, niñas y adolescentes, se debe garantizar su participación activa y significativa en el proceso, en una lógica de cocreación, respetando su derecho a ser escuchados consagrado en el artículo 44 de la Constitución y en el Código de la Infancia y la Adolescencia. Esta participación debe ser: (i) informada, los NNA deben comprender qué es el sistema, para qué se usará y cómo les afectará; (ii) voluntaria, sin coerción ni presión; (iii) adaptada, utilizando metodologías apropiadas para su edad; (iv) respetuosa, valorando sus opiniones y considerándolas seriamente en la toma de decisiones, y (v) sin sobrecarga de responsabilidad, su participación no debe implicar transferirles responsabilidades que corresponden a adultos e instituciones. Se deben establecer mecanismos específicos para que NNA puedan evaluar el impacto que los sistemas tienen sobre ellos y sobre su comunidad.

• **Educación digital y alfabetización en IA para NNA, familias y educadores:** el Ministerio Público, en coordinación con los Ministerios de Educación Nacional y el Ministerio de Tecnologías de la Información y las Comunicaciones, y entidades competentes, debe promover programas de educación digital y alfabetización en IA dirigidos a: (i) niños, niñas y adolescentes para que comprendan qué es la IA, cómo funciona, cómo les afecta, cuáles son sus derechos en entornos digitales y cómo ejercer su agencia y protección; (ii) padres, madres y cuidadores para que puedan acompañar y orientar a los menores en el uso de tecnologías con IA, identificar riesgos y activar mecanismos de protección; (iii) docentes y orientadores para integrar la educación en IA y ética digital en los currículos escolares y detectar situaciones de vulneración de derechos facilitadas por tecnología. Estos programas deben enfatizar la formación ética, el pensamiento crítico ante la información automatizada y el desarrollo de competencias digitales responsables.

• **Enfoque dinámico y contextualizado:** el cumplimiento de este principio debe ser dinámico, adaptándose a los avances tecnológicos constantes y siendo relevante para las generaciones presentes y futuras de niños, niñas y adolescentes. Las medidas de protección tienen que contextualizarse según las diferentes percepciones, impactos y realidades que enfrentan los NNA en diversas comunidades, sectores sociales, zonas geográficas (urbanas, rurales, remotas) y grupos étnico-culturales. Las políticas institucionales deben actualizarse periódicamente para responder a riesgos emergentes y evolución de usos tecnológicos por parte de NNA.

• **Responsabilidad clara y mecanismos de denuncia accesibles:** las instituciones del Ministerio Público designarán responsables específicos (puntos focales o referentes institucionales) encargados de velar por el cumplimiento de este principio en todos los proyectos de IA que afecten a NNA. Es necesario establecer canales de denuncia accesibles, confidenciales y adaptados a NNA, sus familias y educadores para reportar situaciones en las que un sistema de IA pueda estar vulnerando los derechos de menores, y garantizar respuesta oportuna, investigación y adopción de medidas correctivas.

Mecanismos de control

• **Evaluación de impacto sobre derechos de la niñez (EIDN):** implementación obligatoria de EIDN para cualquier sistema de IA que procese datos de NNA o cuyos resultados puedan afectarlos. Esta evaluación debe seguir metodologías reconocidas internacionalmente (como las propuestas por UNICEF o el Comité de los Derechos del Niño) y debe realizarse antes del despliegue del sistema. La EIDN debe documentar: (i) descripción del sistema y su propósito; (ii) identificación de los NNA afectados y características de la población; (iii) análisis de impactos potenciales positivos y negativos sobre sus derechos; (iv) medidas de mitigación de riesgos; (v) mecanismos de participación de NNA en el diseño o evaluación, y (vi) plan de monitoreo continuo. La EIDN debe ser revisada y actualizada periódicamente o cuando cambien las circunstancias del sistema.

• **Auditorías especializadas en protección de NNA:** los sistemas de IA que afecten a NNA deben ser objeto de auditorías independientes realizadas por expertos en derechos de la niñez, psicología del desarrollo, educación y ética de la IA. Estas auditorías deben verificar el cumplimiento de las directrices establecidas en este principio, evaluar la efectividad de las salvaguardas implementadas e identificar riesgos emergentes no previstos inicialmente.

• **Comité de protección de NNA en entornos digitales:** se sugiere la creación de un comité especializado dentro de la estructura de gobernanza del Ministerio Público (que puede ser un subcomité del comité de ética de IA o una instancia independiente), integrado por representantes de la Procuraduría Delegada para la Defensa de los Derechos de la Infancia, la Adolescencia y la Familia, la Defensoría Delegada para los Derechos de la Niñez y la Juventud, almenos 3 personerías municipales o distritales, expertos en desarrollo infantil, representantes de organizaciones de la

sociedad civil especializadas en derechos de NNA, y, cuando sea apropiado, representantes de NNA. Este comité debe: (i) revisar y aprobar las EIDN; (ii) supervisar el cumplimiento de las directrices de este principio; (iii) emitir recomendaciones sobre políticas y prácticas; (iv) promover la educación y alfabetización en IA para NNA, y (v) conocer y resolver casos de vulneración de derechos por sistemas de IA.

• **Protocolos específicos de respuesta ante incidentes:** desarrollo e implementación de protocolos específicos para la gestión de incidentes que involucren afectación a los derechos de NNA por parte de sistemas de IA. Estos protocolos deben establecer procedimientos claros para: (i) recepción y clasificación de denuncias; (ii) suspensión inmediata del sistema si existe riesgo inminente; (iii) investigación expedita con enfoque de derechos de la niñez; (iv) activación de redes de protección (ICBF, comisarías de familia, autoridades competentes); (v) reparación integral del daño causado, y (vi) adopción de medidas preventivas para evitar recurrencia.

• **Monitoreo continuo de sistemas que afecten a NNA:** los sistemas de IA que involucren o afecten a niños, niñas y adolescentes deben estar sujetos a monitoreo continuo más frecuente e intensivo que los sistemas que afectan solo a población adulta. Este monitoreo debe incluir: (i) análisis periódico de métricas de desempeño con desagregación por grupos de edad; (ii) revisión de logs y registros de interacciones para detectar patrones problemáticos; (iii) recolección de retroalimentación de NNA, familias y educadores sobre su experiencia con el sistema; (iv) análisis de incidentes y quejas, y (v) evaluación del impacto real del sistema sobre el bienestar y desarrollo de NNA comparado con los impactos proyectados en la EIDN.

• ***Sandbox* regulatorio para innovación segura con NNA:** cuando se desarrollen sistemas innovadores de IA que involucren a niños, niñas y adolescentes, se puede utilizar el mecanismo de «*sandbox* regulatorio» o ambiente controlado de prueba, donde el sistema opera bajo supervisión estricta, con un número limitado de usuarios, y con salvaguardas reforzadas que permitan experimentación responsable sin exponer a NNA a riesgos innecesarios. Los aprendizajes del *sandbox* deben informar decisiones sobre escalamiento o ajustes antes del despliegue amplio.

• **Transparencia reforzada con familias y comunidad educativa:** cuando un sistema de IA pueda afectar a NNA, la obligación de transparencia algorítmica establecida en la Directiva Conjunta 007 de 2025 debe ser reforzada mediante comunicación proactiva, clara y accesible dirigida específicamente a padres, madres, cuidadores,

docentes y directivos de instituciones educativas. Esta comunicación debe explicar en lenguaje sencillo qué hace el sistema, qué datos procesa, cómo protege a los menores, qué beneficios busca generar, qué riesgos existen y cómo se gestionan, y cómo pueden ejercer sus derechos de objeción, acceso, rectificación y remoción de datos.

Tabla 8. Normativa aplicable al principio de prevalencia de los derechos de niños, niñas y adolescentes

Instrumento	
Constitución Política de Colombia (1991)	Artículo 44 (derechos fundamentales de los niños: «Los derechos de los niños prevalecen sobre los derechos de los demás») y artículo 45 (protección y formación integral de los adolescentes).
Ley 1098 de 2006	Código de la Infancia y la Adolescencia, que desarrolla el principio de interés superior del menor, establece derechos fundamentales de NNA y obligaciones de protección del Estado, la familia y la sociedad.
Ley 1581 de 2012	Artículo 7 (derechos de los titulares: especial protección a datos de menores) y artículo 12 (requisitos especiales para datos de menores: consentimiento de representantes legales).
Convención sobre los Derechos del Niño (ONU, 1989; Ley 12, 1991)	Ratificada por Colombia mediante la Ley 12 de 1991. Artículo 3 (interés superior del niño), artículo 12 (derecho a ser escuchado) y artículo 16 (protección de la vida privada).
Observación general n.° 25 (Comité de los Derechos del Niño de la ONU, 2021)	Sobre los derechos de los niños en relación con el entorno digital, que establece directrices específicas sobre IA, algoritmos, protección de datos y participación de NNA en entornos tecnológicos.
Marco ético para la IA en Colombia (Dapre, 2021)	Principio 8: «Prevalencia de los derechos de niños, niñas y adolescentes», que establece directrices sobre ética de datos, ética de algoritmos y ética de prácticas aplicables a NNA».
Reglamento 2024/1689 - AI Act. (Parlamento Europeo y Consejo de la Unión Europea, 2024)	Anexo III, numeral 3: considera sistemas de IA en ámbito educativo o que afecten a menores como de «alto riesgo», requiriendo evaluaciones de conformidad, supervisión humana y transparencia.
Principios de IA de la OCDE (2019, 2024)	Principio 1.1: «Crecimiento inclusivo, desarrollo sostenible y bienestar», que incluye consideración especial de impactos sobre grupos vulnerables como NNA.

Fuente. Elaboración de los autores.

5. Marco de decisión para adoptar la IA

La adopción responsable de inteligencia artificial (IA) en el Ministerio Público de la República de Colombia requiere un marco estructurado de decisión que abarque la innovación tecnológica, necesidades misionales, cumplimiento normativo y protección de derechos humanos. Este capítulo presenta un modelo de evaluación secuencial diseñado para guiar a funcionarios y decisores de la institución en la valoración crítica de cada propuesta de implementación de sistemas de IA, desde la identificación del problema hasta la definición de métricas de éxito y mecanismos de supervisión humana (Department of Public Expenditure, NDP Delivery and Reform, 2024; Dapre, 2021).

El marco se estructura en tres diferentes componentes clave. Cada componente aborda dimensiones técnicas, éticas, jurídicas y organizacionales específicas, garantizando que solo las iniciativas que superen todos los criterios de viabilidad y pertinencia avancen hacia etapas de implementación piloto o escalamiento institucional.

5.1 ¿Es la IA la mejor solución?

La primera pregunta del marco exige un análisis riguroso de la pertinencia tecnológica. La IA no constituye una solución universal, su adopción solo se justifica cuando posee ventajas demostrables sobre alternativas más simples, predecibles y económicas, en este sentido los sistemas de IA solo deben implementarse cuando soluciones no automatizadas resulten inadecuadas, ineficientes o inviables para el logro de los objetivos institucionales.

Antes de considerar la IA, los equipos técnicos deben evaluar tres alternativas: (i) optimización de procesos actuales mediante mejora continua, rediseño organizacional o capacitación del personal; (ii) automatización basada en reglas deterministas (RPA, sistemas expertos, flujos de trabajo digitales) cuando las decisiones sigan una lógica

explícita y predecible, y (iii) herramientas analíticas convencionales como estadística descriptiva, consultas estructuradas o visualización de datos (Department of Public Expenditure, NDP Delivery and Reform, 2024). El diagnóstico interno del Ministerio Público desarrollado por el IEMP, reveló que múltiples necesidades operativas (gestión de correspondencia, seguimiento de términos, consolidación de informes) pueden resolverse mediante sistemas de información tradicionales sin requerir componentes de aprendizaje automático (Tejedor et al., 2025a).

La justificación de IA debe fundamentarse en criterios objetivos documentados: (i) complejidad de patrones en los datos (relaciones no lineales, dependencias contextuales, variables latentes) que exceden la capacidad de modelado estadístico convencional; (ii) volumen, variedad y velocidad de información que imposibilitan el procesamiento manual o semiautomatizado; (iii) necesidad de personalización adaptativa en tiempo real según características específicas de usuarios o casos, y (iv) evidencia empírica de superioridad de desempeño (precisión, eficiencia, costo-beneficio) demostrada mediante prototipos, pilotos o estudios de casos comparables en jurisdicciones similares (Department of Public Expenditure, NDP Delivery and Reform, 2024).

Es esencial evitar la adopción tecnológica por presión mediática, modas administrativas o decisiones impulsadas exclusivamente por el efecto de estrategias de mercadeo y proveedores comerciales sin evaluación crítica interna. La experiencia internacional documenta casos de inversiones significativas en sistemas de IA que fueron posteriormente abandonados por no agregar valor sobre sistemas heredados (Department of Public Expenditure, NDP Delivery and Reform, 2024).

Para considerar si la IA es la mejor solución, es necesario trabajar con un equipo multidisciplinario de expertos (incluyendo especialistas con buen conocimiento de los datos y expertos en la materia que conozcan el entorno donde se implementará el modelo) y considerar factores como los siguientes (Department of Public Expenditure, NDP Delivery and Reform, 2024, p. 38):

- ¿Qué soluciones alternativas tenemos disponibles para resolver este problema y cuáles son las ventajas y desventajas de cada solución? También debemos considerar cómo se compara el costo previsto de cada solución con nuestro presupuesto. Por ejemplo, ¿se pueden utilizar métodos más sencillos que generen resultados de la misma calidad en menos tiempo o a un menor costo?

• ¿Qué datos tenemos? ¿Serán estos datos precisos, representativos y lo suficientemente completos para ser utilizados? ¿Tenemos un conjunto de datos lo suficientemente grande para entrenar un modelo de IA?

• ¿Podemos utilizar estos datos de manera responsable y estamos autorizados para usarlos bajo las directrices y normas nacionales e internacionales?

• ¿Superarán los beneficios del sistema de IA cualquier posible consecuencia negativa?

• ¿Beneficiará por igual a todos los usuarios o solo ayudará desproporcionadamente a algunos, a costa de otros?

• ¿Tenemos a nuestra disposición las competencias necesarias para poder implementar la solución de IA?

• ¿Resolverá el problema? ¿Qué métricas son importantes para evaluar esta hipótesis y cómo las mediremos?

5.2 ¿Qué tipo de sistema de IA es más apropiado?

Una vez validada la pertinencia de adoptar alguna solución basada en IA, el segundo paso tiene que ver con seleccionar el tipo de sistema que mejor satisfaga las necesidades identificadas, considerando el nivel de autonomía, complejidad técnica, requisitos de datos, capacidades explicativas y categorización de riesgo según marcos regulatorios aplicables.

El diagnóstico realizado por el IEMP identificó como más viables y pertinentes los sistemas de apoyo a la decisión, específicamente: (i) asistentes de análisis de tutelas basados en recuperación aumentada por generación (RAG) (Willats et al., 2025), modelados según Prometea (Estevez, et al. 2020). (ii) herramientas de búsqueda semántica sobre bases de jurisprudencia y doctrina; (c) transcriptores automáticos de audiencias con verificación humana obligatoria, y (d) sistemas de detección de patrones anómalos en contratación pública para alertas tempranas de riesgo de corrupción (Department of Public Expenditure, NDP Delivery and Reform, 2024, p. 39).

Con el fin de ayudar a determinar qué tipo de solución de IA es la más adecuada, todo el equipo multidisciplinario debe reflexionar profundamente sobre las siguientes preguntas:

- ¿De qué manera espera el usuario final utilizar o interactuar con el sistema de IA?
- ¿Necesitamos poder obtener el mismo resultado cada vez que ejecutamos el modelo?
- ¿Qué nivel de precisión necesitamos?
- ¿Hasta qué punto (o con qué nivel de detalle) será necesario explicar el modelo al usuario final?
- ¿Cuánto tiempo debe tardar el modelo en generar los resultados?
- ¿Reflejan los datos de entrenamiento las situaciones del mundo real que el modelo abordará?
- ¿En qué categoría de riesgo se clasificaría según el Reglamento de IA (AI Act) de 2024?

5.3 Consideraciones sobre IA gratuita vs. licenciada

La elección entre herramientas de IA de acceso gratuito (modelos de lenguaje grandes públicos, APIs sin costo, software de código abierto) y soluciones licenciadas de proveedores comerciales plantean dilemas técnicos, financieros, jurídicos y éticos que requieren análisis caso por caso (Department of Public Expenditure, NDP Delivery and Reform, 2024; Dapre, 2021).

Por un lado, las soluciones gratuitas ofrecen ventajas aparentes, sin costos directos de licenciamiento, flexibilidad de personalización en modelos de código abierto e innovación acelerada mediante comunidades de desarrollo distribuidas. Sin embargo, presentan riesgos críticos para entidades públicas: (i) ausencia de garantías contractuales sobre disponibilidad, rendimiento o continuidad del servicio (los proveedores

pueden modificar términos o descontinuar servicios unilateralmente); (ii) responsabilidad legal difusa en caso de fallos que generen daños (ausencia de acuerdos de nivel de servicio, seguros de responsabilidad profesional o cláusulas de indemnización); (iii) riesgos de privacidad y seguridad cuando datos institucionales se procesan en servidores externos sin control jurisdiccional (especialmente crítico para información clasificada o datos personales de ciudadanos); (iv) dependencia de infraestructura externa que puede ser inaccesible por interrupciones tecnológicas, decisiones comerciales del proveedor o incluso restricciones geopolíticas, y (v) opacidad sobre prácticas de entrenamiento de modelos (posible uso de datos de usuarios para mejora de algoritmos propietarios sin consentimiento explícito) (Department of Public Expenditure, NDP Delivery and Reform, 2024).

Por otro lado, las soluciones licenciadas con contratos comerciales formales proveen: (i) acuerdos de nivel de servicio con garantías de disponibilidad, tiempos de respuesta y soporte técnico especializado; (ii) asignación contractual clara de responsabilidades entre proveedor y entidad contratante; (iii) opciones de despliegue *on-premise* o en nubes privadas que mantienen control sobre datos sensibles; (iv) documentación técnica exhaustiva, auditorías de terceros y certificaciones de seguridad, y (v) compromisos de cumplimiento normativo (GDPR, estándares ISO, certificaciones de ciberseguridad). Sus desventajas incluyen costos significativos, dependencia de proveedores específicos (*vendor lock-in*) y menor transparencia algorítmica cuando los modelos son cajas negras propietarias.

La evidencia empírica recomienda explorar modelos híbridos: uso de grandes modelos de lenguaje *open-source* para capacidades generales, combinados con ajuste fino (*fine-tuning*) y RAG sobre bases documentales propias en infraestructura controlada institucionalmente (Department of Public Expenditure, NDP Delivery and Reform*, 2024*).

Al determinar si se debe adquirir o desarrollar una solución de IA, es necesario desarrollar una evaluación de ambas opciones para definir el mejor enfoque. Como parte de esta evaluación, es necesario considerar los siguientes factores:

- ¿Cuáles son las ventajas y desventajas de ambas soluciones? ¿Qué solución tendrá un mejor rendimiento o resultará en una mejor solución para el usuario final?

- ¿Cuáles son los costos y el retorno de la inversión de desarrollar frente a comprar

a lo largo de todo el ciclo de vida de la IA? Por ejemplo, ¿tendrá un costo mayor durante la fase de desarrollo? ¿Implicará la solución adquirida una tarifa de licencia recurrente?

• ¿Con qué rapidez necesitamos que se implemente el sistema de IA?

• ¿Podría la solución desarrollada o adquirida utilizarse para otros casos de uso en el servicio público?

• ¿Tenemos a nuestra disposición las competencias necesarias para desarrollar una solución o, de hecho, para operar una solución adquirida? Esto incluye implementar la solución, operarla y mantenerla después del despliegue.

• ¿Ofrece una de las opciones una mejor seguridad de los datos?

• ¿Ofrece una de las opciones una mejor compatibilidad con los sistemas existentes?

• ¿Qué capacitación y soporte técnico se necesitarían para ambas opciones?

A diferencia de las soluciones tecnológicas tradicionales, la integración y administración de la inteligencia artificial representan costos notablemente más elevados. Su adopción demanda requerimientos superiores en cuanto a procesamiento computacional, cantidad de datos y nivel de preparación institucional. Asimismo, es indispensable prever que las inversiones en desarrollo, implementación y soporte técnico tienden a multiplicarse una vez que el sistema entra en su fase operativa.

Desde una perspectiva financiera, la fuerte dependencia de la IA hacia la computación en la nube transforma el perfil de gastos y riesgos de la institución. Esto puede impactar la continuidad del proyecto y aumentar la dependencia de infraestructuras críticas. Por esta razón, resulta indispensable realizar un análisis riguroso de costo-eficiencia (*value for money*) desde el inicio, aplicando evaluaciones ágiles y un monitoreo financiero constante a lo largo de todo el ciclo de vida del proyecto. Este enfoque estratégico garantiza que la decisión de adoptar IA esté bien fundamentada, que se maximice el valor de la implementación y que el sistema sea sostenible y beneficioso a largo plazo.

Según el *Department of Public Expenditure, NDP Delivery and Reform*, (2024), existen diversos factores clave para el análisis de costos que deben integrarse en el caso de la planeación inicial y revisarse periódicamente para demostrar la eficiencia de la inversión. Entre estas consideraciones se incluyen:

- **Costos de infraestructura:** el costo del hardware, los recursos de computación en la nube y el almacenamiento necesarios para entrenar y ejecutar los modelos de IA.

- **Costos de datos:** la adquisición, preparación y etiquetado de datos para los modelos de IA puede representar un gasto significativo, especialmente en proyectos complejos que requieren grandes conjuntos de datos.

- **Costos de desarrollo:** esto cubre los salarios de los ingenieros de IA, científicos de datos y otros especialistas involucrados en la construcción y el perfeccionamiento de los modelos de IA. También incluye el costo de las licencias y herramientas de software.

- **Costos operativos:** son los gastos continuos o recurrentes asociados a la ejecución y el mantenimiento de los sistemas de IA, tales como el consumo de energía, el monitoreo y las actualizaciones.

- **Costos imprevistos:** los proyectos de IA pueden enfrentarse a desafíos no previstos, como la necesidad de reentrenar el modelo o la depuración de errores (*debugging*), lo cual puede incrementar el costo total.

Como en cualquier estudio de viabilidad, el punto de partida siempre debe ser evaluar la opción de mantener el *statu quo* (no implementar cambios). Tanto esta alternativa como la elección del momento oportuno para la adopción requieren una consideración minuciosa. Esto se debe a que las soluciones de IA, al igual que otras tecnologías emergentes, suelen implicar costos elevados para quienes las adoptan en etapas tempranas, precios que tienden a descender conforme nuevos proveedores dinamizan el mercado.

Asimismo, es indispensable evaluar el despliegue de la IA en los servicios públicos, analizando sus procesos, impacto y la relación costo-beneficio. Es necesario comprender cómo se comparan los sistemas de IA frente a los métodos actuales, buscando mejorar las intervenciones existentes, fundamentar futuras adopciones y garantizar la responsabilidad en la ejecución del gasto público.

Si bien los principios fundamentales para evaluar el impacto de un programa gubernamental aplican también a la IA, esta tecnología introduce oportunidades y desafíos únicos, derivados principalmente de su naturaleza iterativa y evolutiva. En este sentido, las mejores prácticas emergentes para evaluar el impacto de la IA en el sector público abarcan lo siguiente:

• Considerar la evaluación lo antes posible en el proceso de despliegue de la IA y tener claridad sobre el propósito de dicho despliegue.

• Desarrollar una comprensión completa de la relación entre los insumos (*inputs*) y productos (*outputs*) propuestos, así como los resultados (*outcomes*) esperados del despliegue de la IA (lo que comúnmente se conoce como Modelo Lógico y Teoría del Cambio).

• Documentar y registrar todos los pasos planificados y ejecutados en el desarrollo y despliegue de la solución de IA, y señalar cualquier discrepancia entre lo planificado y lo que realmente ocurre.

• Estar preparados para adaptar y adoptar métodos de evaluación adicionales que sean adecuados para reflejar la naturaleza evolutiva del despliegue de la IA a lo largo del tiempo.

• Considerar y documentar cualquier diferencia en los resultados y el impacto para distintos grupos de población, tanto al planificar la evaluación como durante el transcurso del despliegue, a medida que estas se identifiquen.

• Pensar desde el principio en cómo establecer una línea de base claramente definida para respaldar la evaluación, considerando qué datos ya existen y cuáles podrían necesitar ser recolectados.

5.4 Inclusión y diversidad desde el inicio

Los sistemas de IA aprenden patrones de datos históricos que reflejan estructuras sociales, relaciones de poder y prácticas institucionales preexistentes. Cuando estos datos históricos contienen sesgos sistemáticos contra grupos poblacionales específicos (mujeres, personas de grupos étnicos minoritarios, personas con discapacidad, población LGBTIQ+, habitantes de zonas rurales), los algoritmos entrenados perpetúan

y amplifican discriminaciones existentes bajo apariencia de objetividad técnica (Department of Public Expenditure, NDP Delivery and Reform, 2024; Dapre, 2021; OECD, 2019, 2024). Los principios de no discriminación e inclusión reconocidos en el Marco ético nacional exigen que toda iniciativa de IA incorpore estrategias proactivas de inclusión y evaluación de sesgos desde las fases tempranas de diseño (Dapre, 2021).

La integración de perspectivas diversas requiere: (i) equipos multidisciplinarios que incluyan no solo perfiles técnicos (ingenieros, científicos de datos), sino también especialistas en derechos humanos, género, grupos étnicos, atención a población vulnerable, sociólogos y representantes de comunidades usuarias; (ii) consultas estructuradas con organizaciones de la sociedad civil, asociaciones de víctimas, grupos de defensa de derechos y beneficiarios directos de servicios institucionales; (iii) análisis desagregado de datos de entrenamiento y resultados algorítmicos por variables demográficas clave (sexo, edad, etnia, ubicación geográfica, nivel socioeconómico, condición de discapacidad) para detectar disparidades de desempeño; (iv) protocolos de corrección de sesgos mediante técnicas de balanceo de datos, ajuste de funciones de pérdida, posprocesamiento de predicciones y auditorías algorítmicas con métricas específicas de equidad (paridad demográfica, igualdad de oportunidades, suficiencia separativa), y (v) evaluaciones de impacto sobre derechos humanos (*human rights impact assessments*) previas a despliegue (Department of Public Expenditure, NDP Delivery and Reform, 2024).

Algunos sesgos adicionales se relacionan con sistemas de predicción de reincidencia disciplinaria que podrían discriminar contra funcionarios de regiones específicas si datos históricos reflejan patrones de fiscalización desigual; herramientas de análisis de tutelas que podrían subvalorar solicitudes de poblaciones rurales si el lenguaje de entrenamiento privilegia terminología urbana especializada; asistentes de atención ciudadana que podrían brindar respuestas de menor calidad a dialectos regionales si fueron entrenados mayoritariamente con español estándar capitalino. Estos riesgos obligan a implementar desde la fase de diseño medidas compensatorias específicas.

5.5 Siguientes pasos

Una vez el equipo haya completado el marco de decisión y tenga la capacidad de optar por proceder con una solución de IA determinada, el siguiente paso es diligenciar con los equipos multidisciplinarios, el Lienzo o Canvas de IA responsable en el marco de un ejercicio participativo y colaborativo de planificación.

6. Canvas de IA responsable para el Ministerio Público

El marco de decisión requiere herramientas operativas que traduzcan principios en procesos concretos. El Canvas de IA responsable, adaptado del modelo irlandés (*Department of Public Expenditure, NDP Delivery and Reform, 2024*) y complementado con el Marco ético colombiano (Dapre, 2021), opera como plantilla estructurada para talleres colaborativos donde equipos multidisciplinarios evalúan sistemáticamente diferentes propuestas desde perspectivas técnicas, éticas, jurídicas y organizacionales, garantizando un análisis prospectivo riguroso antes de aprobar inversiones en soluciones basadas en IA.

6.1 Descripción general del Canvas

El Canvas o lienzo de IA responsable es una herramienta práctica y estructurada, diseñada para apoyar a los servidores públicos en el desarrollo, implementación y supervisión de soluciones de inteligencia artificial que cumplan con los ocho principios de responsabilidad. Se recomienda encarecidamente utilizar esta herramienta durante la fase de planificación de cualquier proyecto de IA.

El diligenciamiento del Canvas debe ser un esfuerzo colaborativo realizado por un equipo multidisciplinario que incluya, aunque no se limite, a proveedores de servicios, equipos técnicos, líderes de producto, asesores legales y socios externos. Este proceso debe realizarse en alineación con los estándares de diseño gubernamentales.

El objetivo principal del Canvas es facilitar un diálogo estructurado sobre la aplicación efectiva de los ocho principios. A través de preguntas clave alineadas con

las directrices, la herramienta guía a los funcionarios en pasos esenciales como la identificación de las partes interesadas y la definición precisa del problema a resolver. Asimismo, la herramienta abre la discusión sobre el cumplimiento normativo, abordando regulaciones como el RGPD (2018) y el Reglamento de IA de la UE (Parlamento Europeo y Consejo de la Unión Europea, 2024), y fomenta una gestión de riesgos proactiva para evaluar y mitigar posibles amenazas desde el inicio. El Canvas de IA responsable está disponible para consulta en el apéndice B y será desplegado en las plataformas web del Ministerio Público.

El Canvas se organiza en doce bloques secuenciales que cubren el ciclo completo de diseño, implementación y gobernanza:

(I) Información general: identifica nombre, liderazgo, función misional y objetivos específicos.

(II) Partes interesadas: mapea actores afectados, aliados internos/externos y grupos de impacto especial (NNA, comunidades étnicas, población vulnerable).

(III) Categorización de riesgo: clasifica según el EU AI Act (Parlamento Europeo y Consejo de la Unión Europea, 2024) en categorías de riesgo mínimo, limitado, alto o inaceptable, determinando nivel de escrutinio regulatorio.

(IV) Declaración del problema y justificación de IA: exige evidencia cuantitativa, justificación de superioridad sobre alternativas, tipo de solución apropiada y documentación de alternativas descartadas.

(V) Inclusión y diversidad: mecanismos para integrar perspectivas diversas en diseño y validación, análisis desagregado y corrección de sesgos (principio 3).

(VI) Agencia humana y supervisión: puntos de intervención humana, criterios de escalamiento y evaluación de afectación a derechos humanos (principio 1).

(VII) Privacidad y gobernanza de datos: tipos de datos procesados, bases jurídicas, medidas técnicas de protección (cifrado, anonimización) y cumplimiento de Ley 1581 de 2012 (principio 2).

(VIII) Transparencia, explicabilidad y robustez: información pública, mecanismos de explicación, estándares de ciberseguridad y planes de contingencia (principios 4 y 5).

(IX) Diversidad, no discriminación y equidad: estrategias técnicas de detección de sesgos (paridad demográfica, igualdad de oportunidades), protocolos de mitigación y accesibilidad universal (WCAG 2.1 AA).

(X) Bienestar social, ambiental y marco normativo: evaluación de impactos positivos/negativos, marco normativo aplicable e instituciones relevantes (principio 6).

(XI) Responsabilidad, comunicación y ciclo de vida: matriz RACI por fase del ciclo, estrategia de comunicación y mecanismos de rendición de cuentas (principio 7).

(XII) Aprobaciones: firmas de líder técnico, supervisor jurídico y director, garantizando escrutinio multinivel.

El Canvas traduce los cinco interrogantes del marco de decisión (capítulo 5) y los ocho principios de gobernanza (capítulo 4) en formato operativo: el bloque IV operacionaliza las preguntas 5.1 y 5.2 (pertinencia y tipo de IA); el bloque V responde a la pregunta 5.4 (inclusión); el bloque VII aborda parcialmente la pregunta 5.3 (licenciamiento); el bloque once desarrolla la pregunta 5.5 (siguientes pasos), y los bloques del VI al X operacionalizan los ocho principios del marco ético.

Esta integración garantiza que ningún proyecto avance sin evidencia documentada de cumplimiento de todos los criterios de pertinencia, viabilidad técnica, conformidad normativa y aceptabilidad ética. El Canvas no debe entenderse como una lista de verificación burocrática, sino como un instrumento de reflexión crítica colectiva que externaliza razonamientos implícitos, expone supuestos cuestionables y promueve justificaciones explícitas y sustentadas con evidencia real y empírica.

6.2 ¿Cómo utilizar el Canvas?

El Canvas de IA responsable se utiliza en las etapas iniciales de cualquier proyecto de inteligencia artificial. Su aplicación ideal es dentro de talleres de planificación

o sesiones de design thinking, facilitando que los equipos aborden e integren los principios de IA responsable desde el comienzo. A continuación, se detallan los pasos clave para aprovechar esta herramienta de manera efectiva:

• **Conformar un equipo multidisciplinario:** este puede incluir, entre otros, a los equipos técnicos, los responsables del producto (*product owners*) y los equipos jurídicos. También puede contar con la participación de partes interesadas externas relevantes. Esto aporta una amplia gama de perspectivas y permite identificar desde el inicio los riesgos potenciales y los desafíos de implementación.

• **Trabajar en cada sección del Canvas:** se debe debatir cada pregunta, pero el Canvas no debe considerarse una lista exhaustiva o cerrada. Se alienta activamente a los equipos a añadir preguntas adicionales y a adaptar la herramienta para que se ajuste al caso de uso específico.

• **Gestión de riesgos:** el Canvas fomenta una gestión proactiva para evaluar y mitigar los riesgos potenciales desde el principio. No obstante, deben iniciarse planes que permitan realizar evaluaciones de riesgos continuas a lo largo de todo el ciclo de vida de la IA.

• **Estrategia de comunicación:** vale la pena considerar, desde el inicio, la estrategia de comunicación sobre cómo se utilizará el sistema de IA, ya que este factor puede influir en la propia solución.

• **Monitoreo continuo:** el Canvas está diseñado para ser utilizado durante la etapa de planificación del ciclo de vida de la IA. Sin embargo, los equipos deben establecer mecanismos de gobernanza que cubran la totalidad del ciclo de vida de la inteligencia artificial. (Department of Public Expenditure, NDP Delivery and Reform, 2024; Dapre, 2021).

Alcance y limitaciones del Canvas de IA responsable

Si bien el Canvas o lienzo de IA responsable constituye un instrumento de gran valor para armonizar los proyectos con los lineamientos establecidos, es importante aclarar que su mera utilización no garantiza automáticamente que la solución resultante sea responsable ni que cumpla con la normativa vigente.

El objetivo del Canvas es plantear interrogantes estratégicos alineados con estas directrices para guiar la reflexión dentro del Ministerio Público de la República de Colombia. En última instancia, recae sobre las partes responsables la obligación de garantizar que la solución de IA sea lícita en su totalidad (de extremo a extremo) y que sus procesos de desarrollo, despliegue y mantenimiento se ejecuten con responsabilidad.

6.3 Siguientes pasos

Luego de desarrollar los talleres de cocreación y diligenciar el Canvas de IA responsable, el capítulo 7 sirve como una hoja de ruta para identificar las actividades principales que se deben llevar a cabo a lo largo del ciclo de vida de la IA.

7. Ciclo de vida de la IA responsable

7.1 Introducción al ciclo de vida de la IA

El compromiso institucional con la adopción responsable de inteligencia artificial va más allá del cumplimiento formal de obligaciones normativas. Si bien el acatamiento del marco jurídico colombiano (Constitución Política, Ley 1581 de 2012, Ley 1712 de 2014) y de estándares internacionales aplicables (Reglamento de IA de la Unión Europea, Principios de la OCDE) establece una base regulatoria esencial, la verdadera adopción responsable requiere la integración sistemática de los ocho principios de gobernanza ética desarrollados en el capítulo 4 en cada etapa del ciclo de vida operativo de los sistemas de IA (Department of Public Expenditure, NDP Delivery and Reform, 2024; Dapre, 2021)

Este capítulo provee orientación práctica a funcionarios y equipos técnicos del Ministerio Público sobre acciones concretas, consideraciones críticas y mejores prácticas que deben aplicarse durante cada fase del desarrollo, despliegue, operación y eventual retiro de sistemas de IA. La estructura del ciclo de vida adoptada sigue la definición establecida por la Organización para la Cooperación y el Desarrollo Económicos (OCDE, 2019, 2024) y alinea cada fase con los principios éticos institucionales, garantizando coherencia metodológica con marcos internacionales de referencia.

Este capítulo presenta un modelo de ciclo de vida para la IA responsable, estructurado en siete fases secuenciales pero interconectadas. Este enfoque permite al Ministerio Público aplicar los principios de gobernanza de manera sistemática, identificar riesgos de manera temprana y garantizar que los sistemas de IA se desarrollen y operen de forma lícita, ética y robusta. Las siete fases del ciclo de vida son: (i) planificación y diseño; (ii) recolección y procesamiento de datos; (iii) construcción del modelo; (iv) verificación y validación; (v) despliegue; (vi) operación y monitoreo, y (vii) retiro, desmantelamiento o actualización (ver tabla 9).

Es fundamental reconocer que estas fases no constituyen un proceso lineal e inmutable, sino un ciclo iterativo y ágil donde las etapas pueden repetirse, sobreponerse o ejecutarse de manera no secuencial según las necesidades específicas del proyecto (OCDE 2019, 2024). La decisión de retirar un sistema puede ocurrir en cualquier momento de la operación, y nuevos casos de uso pueden requerir reiniciar el ciclo con modificaciones significativas.

La gobernanza de la IA debe aplicarse de manera consistente a lo largo de todo el ciclo de vida del sistema. A continuación, se detallan las consideraciones y acciones clave para cada fase, asegurando que los principios éticos se integren desde el inicio hasta el final.

Tabla 9. Ciclo de vida de la IA

Fase del ciclo de vida	Actividades y entregables clave
(i) Planificación y diseño	- Estudio de viabilidad y análisis de necesidad. - Clasificación inicial de riesgo. - Realización de evaluaciones de impacto (derechos humanos, privacidad, equidad). - Definición de requisitos funcionales y no funcionales. - Diseño de la arquitectura del sistema.
(ii) Recolección y procesamiento de datos	- Recolección, limpieza y preparación de datos. - Documentación del linaje y características de los datos (p. ej., *datasheets for datasets*). - Aseguramiento de la calidad y representatividad de los datos.
(iii) Construcción del modelo	- Entrenamiento, prueba y validación del modelo algorítmico. - Documentación del modelo (p. ej., *model cards*). - Implementación de técnicas de mitigación de sesgo.
(iv) Verificación y validación	- Pruebas independientes de robustez, seguridad y sesgo. - Auditorías de cumplimiento normativo y ético. - Validación del desempeño en entornos controlados.
(v) Despliegue	- Implementación en entorno de producción. - Monitoreo inicial del rendimiento y la equidad. - Implementación de protocolos de supervisión humana.
(vi) Operación y monitoreo	- Auditorías periódicas (técnicas, de cumplimiento, de sesgo). - Revisión continua del desempeño del sistema. - Recopilación de retroalimentación de usuarios. - Informes de transparencia y rendición de cuentas.
(vii) Retiro, desmantelamiento o actualización	- Definición de criterios para el retiro del sistema. - Plan de transición para minimizar disrupciones. - Gestión segura de datos históricos y del modelo. - Documentación de lecciones aprendidas.

Fuente. Elaboración de los autores.

Esta naturaleza iterativa exige flexibilidad metodológica, pues conforme los sistemas evolucionan, cambian los contextos de aplicación, emergen riesgos no anticipados o se actualizan marcos normativos; las prácticas relevantes para cada fase deben adaptarse. Los equipos institucionales deben cultivar una capacidad de respuesta ágil ante cambios, manteniendo siempre la adherencia a principios éticos fundamentales (Department of Public Expenditure, NDP Delivery and Reform, 2024).

7.2 Aplicación de principios de gobernanza en el ciclo de vida

Las siguientes secciones presentan un enfoque estructurado con acciones sugeridas que los equipos del Ministerio Público pueden implementar para operacionalizar los ocho principios de gobernanza ética (capítulo 4) en cada fase del ciclo de vida. Es importante enfatizar que estas prácticas no constituyen una lista exhaustiva ni son de aplicación obligatoria universal. Su pertinencia y profundidad de implementación deben calibrarse según el contexto específico del sistema, su clasificación de riesgo conforme al marco del Reglamento de IA de la Unión Europea (Parlamento Europeo y Consejo de la Unión Europea, 2024), y la naturaleza de los derechos humanos potencialmente afectados.

Diferentes casos de uso presentan perfiles de riesgo heterogéneos como sistemas de apoyo administrativo de bajo riesgo requieren salvaguardas menos intensivas que sistemas de alto riesgo que afecten procesos disciplinarios, decisiones sobre derechos de ciudadanos o análisis de evidencia en investigaciones penales. Por tanto, no todos los sistemas exigirán la aplicación de todas las prácticas disponibles para considerarse seguros y responsables. Los equipos deben ejercer juicio profesional informado, complementando estas orientaciones con acciones adicionales específicas al contexto y garantizando salvaguardas proporcionales a los riesgos identificados.

Estas orientaciones no sustituyen ni reemplazan los mecanismos de gobierno corporativo y cumplimiento regulatorio existentes en materia de protección de datos, seguridad de la información y contratación pública. Su función es complementaria y tiene como objetivo ayudar a formular las preguntas correctas en cada etapa y fortalecer el compromiso institucional con la IA responsable. Cuando se combinan con estructuras de gobernanza establecidas, estas prácticas potencian la capacidad institucional para gestionar la IA de manera ética, transparente y efectiva (Dapre, 2021).

Siguiendo la metodología del *High-Level Expert Group* on AI de la Comisión Europea (2019), cada fase del ciclo de vida identifica áreas de «enfoque prioritario» (*primary focus*) que requieren atención intensificada debido a su relevancia e impacto particular en determinada etapa. Esta priorización orienta a los profesionales sobre dónde concentrar esfuerzos sin disminuir la importancia de los demás principios, que permanecen como consideraciones esenciales en todas las fases (Department of Public Expenditure, NDP Delivery and Reform, 2024).

7.3 Primera fase: Planificación y diseño

Esta fase se enfoca en establecer los fundamentos éticos, técnicos y organizacionales del sistema; definición de arquitectura de supervisión humana, y evaluación preliminar de riesgos sobre derechos humanos.

Tabla 10. Principios y acciones fase de planificación y diseño

Principio	Acciones
Agencia y supervisión humana	• Definir explícitamente los puntos del flujo operativo donde se integrará supervisión y control humano, especialmente en procesos de toma de decisiones que afecten derechos o la situación jurídica de personas. • Establecer directrices claras sobre intervención humana para garantizar que las salidas algorítmicas no anulen la capacidad de juicio profesional de funcionarios. • Asignar responsabilidades específicas por rol para la supervisión desde el inicio mediante la matriz Raci (responsable, aprobador, consultado, informado). • Identificar y consultar formalmente a partes interesadas clave: expertos temáticos, asesores jurídicos, representantes de grupos potencialmente afectados, organizaciones de la sociedad civil. • Evaluar si el sistema pudiera producir efectos jurídicos o afectar significativamente a usuarios; en estos casos, los ciudadanos tienen derecho constitucional a no ser sometidos a decisiones basadas únicamente en procesamiento automatizado (C.P., 1991, art. 29). • Utilizar el Canvas de IA responsable (capítulo 6) como herramienta de evaluación estructurada (Department of Public Expenditure, NDP Delivery and Reform, 2024).
Robustez técnica y seguridad	• Verificar que el sistema propuesto se alinee con políticas institucionales de seguridad de la información y estándares de ciberseguridad aplicables (Esquema Nacional de Seguridad colombiano, ISO/IEC 27001). • Realizar evaluaciones de riesgo técnico previas al despliegue, identificando vulnerabilidades potenciales de seguridad, privacidad y desempeño. • Establecer planes de contingencia y respaldo ante fallos, incluyendo protocolos de respuesta a incidentes, procedimientos de recuperación de datos y estrategias de mitigación de ataques cibernéticos. • Involucrar formalmente al equipo de seguridad de TI institucional en todas las etapas del ciclo de vida para garantizar supervisión continua de procedimientos de seguridad.

Privacidad y gobernanza de datos	• Asegurar la existencia de procesos robustos de protección de privacidad y datos personales; realizar Evaluaciones de Impacto sobre Protección de Datos (EIPD) cuando el tratamiento implique alto riesgo para derechos de titulares, conforme a la Ley 1581 de 2012 y Decreto 1377 de 2013. • Revisar y documentar procedimientos de acceso a datos; integrar protocolos de minimización (recolectar solo datos estrictamente necesarios) y seguridad desde el diseño (privacy by design); planificar técnicas de anonimización o pseudonimización cuando sea técnicamente factible. • Identificar explícitamente campos de información personal identificable (PII) que el sistema no debe procesar bajo ninguna circunstancia, estableciendo controles técnicos de restricción (Department of Public Expenditure, NDP Delivery and Reform, 2024; DAPRE, 2021).
Transparencia y explicabilidad	• Establecer el nivel requerido de transparencia y explicabilidad para el sistema según su clasificación de riesgo y derechos potencialmente afectados. • Definir procesos alternativos que permitan interacción humana directa y explicaciones comprensibles de decisiones algorítmicas a los afectados para sistemas de alto riesgo o que afecten derechos humanos.
Equidad, no discriminación y justicia	• Desarrollar un plan de participación de partes interesadas que garantice inclusión de perspectivas diversas durante todo el ciclo de vida. • Realizar evaluaciones de impacto sobre derechos humanos si existe potencial de afectación negativa a grupos poblacionales específicos. • Evaluar accesibilidad del sistema para personas con discapacidad, adoptando principios de diseño universal. • Conformar equipos de desarrollo multidisciplinarios y diversos en términos de género, origen geográfico, formación profesional y experiencia, garantizando que las personas correctas participen desde las etapas iniciales.
Bienestar social y ambiental	• Realizar evaluaciones de impacto social para garantizar que el sistema contribuya positivamente a la misión institucional y al bienestar ciudadano, identificando potenciales externalidades negativas (desplazamiento laboral, ampliación de brechas digitales, erosión de confianza pública, deterioro del medio ambiente).
Rendición de cuentas y responsabilidad	• Establecer marcos de responsabilidad desde el inicio, cubriendo todas las áreas de riesgo como el diseño, datos, algoritmos, desempeño, terceros proveedores y cumplimiento normativo. • Utilizar herramientas de evaluación de impacto (red teaming, evaluaciones algorítmicas de impacto, análisis de escenarios adversos) para minimizar impactos negativos. • Especificar roles responsables de detectar y corregir desviaciones del comportamiento esperado, y documentar formalmente cadenas de responsabilidad.

Prevalencia de los derechos de niños, niñas y adolescentes	• Realizar evaluación específica del impacto sobre derechos de la niñez (EIDN) cuando el sistema pueda afectar directa o indirectamente a niños, niñas y adolescentes (NNA), ya sea como usuarios, sujetos de decisiones administrativas o beneficiarios de servicios institucionales. • Identificar explícitamente si el sistema procesará datos de NNA y establecer restricciones reforzadas de tratamiento conforme al principio del interés superior del niño consagrado en el artículo 44 de la Constitución Política y en el Código de la Infancia y la Adolescencia (Ley 1098, 2006). • Prohibir categóricamente usos perjudiciales del sistema que puedan generar daño psicológico, explotación, manipulación o discriminación contra NNA. • Diseñar mecanismos de participación significativa y apropiada según edad y madurez de NNA en la evaluación del sistema cuando sean población directamente afectada, garantizando que sus voces sean escuchadas en decisiones que les conciernen. • Establecer salvaguardas especiales en el diseño para proteger la vulnerabilidad particular de esta población, incluyendo supervisión humana reforzada y umbrales de explicabilidad más exigentes cuando las decisiones algorítmicas puedan afectar sus derechos (Dapre, 2021).

Fuente: Elaboración de los autores

7.4 Segunda fase: Recolección y procesamiento de datos

El objetivo de esta fase es determinar los lineamientos para una correcta recolección, gestión y protección de datos con énfasis en privacidad, calidad de datos y equidad representativa.

Tabla 11. Principios y acciones fase de recolección y procesamiento de datos.

Principio	Acciones
Agencia y supervisión humana	• Validar que los procesos de procesamiento de datos incluyan puntos críticos de intervención humana. • Monitorear continuamente para prevenir accesos no autorizados o uso indebido de información sensible.
Robustez técnica y seguridad	• Garantizar manejo seguro y confiable de datos mediante controles de acceso estrictos, cifrado en tránsito y en reposo, y registros de auditoría de todas las operaciones. • Auditar fuentes de datos y procesos de transformación para verificar integridad, completitud y trazabilidad.

Privacidad y gobernanza de datos	• Limitar la recolección a campos estrictamente necesarios para los propósitos declarados, y evitar información personal identificable (PII) salvo que sea indispensable y exista base jurídica clara para su tratamiento. • Abordar sistemáticamente problemas de calidad de datos (valores faltantes, duplicados, inconsistencias, errores de medición) y sesgos inherentes a las fuentes (subrepresentación de grupos poblacionales, errores sistemáticos de registro, sesgos históricos en procesos administrativos). • Documentar exhaustivamente todos los pasos de preparación de datos: limpieza, transformación, agregación, derivación de variables, técnicas de imputación aplicadas (Department of Public Expenditure, NDP Delivery and Reform, 2024).
Transparencia y explicabilidad	• Documentar conjuntos de datos utilizados mediante data cards o fichas técnicas estandarizadas que incluyan: fuentes de origen, fechas de recolección, tamaño muestral, variables incluidas, métodos de recolección, transformaciones aplicadas, procesos de etiquetado cuando corresponda, políticas de retención y eliminación. • Mantener registros completos de linaje de datos (data lineage) que permitan trazabilidad desde fuentes originales hasta datasets procesados.
Equidad, no discriminación y justicia	• Asegurar que los datos sean representativos de la población objetivo, evitando subrepresentación sistemática de grupos específicos. • Identificar, documentar y mitigar sesgos en datos mediante técnicas de balanceo muestral, sobremuestreo de grupos minoritarios o reponderación estadística cuando sea metodológicamente apropiado. • Involucrar a partes interesadas en la evaluación de representatividad y calidad de datos.
Bienestar social y ambiental	• Evitar que las prácticas de recolección y manejo de datos generen impactos sociales o externalidades negativas, como invasiones desproporcionadas de privacidad o recolección excesiva que erosione confianza pública.
Rendición de cuentas	• Mantener registros de auditoría (audit trails) completos de todas las operaciones sobre datos. • Continuar evaluaciones de impacto iniciadas en la fase de diseño, actualizándolas conforme se dispone de información real sobre características de los datos.
Rendición de cuentas	• Aplicar restricciones reforzadas en la recolección y procesamiento de datos de niños, niñas y adolescentes, limitándola estrictamente a casos donde sea absolutamente indispensable para el cumplimiento de funciones misionales y exista base jurídica específica. • Obtener consentimiento de representantes legales conforme a la Ley 1581 de 2012, explicando en lenguaje claro y adaptado a distintas edades el propósito del tratamiento, los derechos de los titulares y los mecanismos de protección implementados. • Adoptar técnicas de anonimización o pseudonimización robusta de datos de NNA siempre que sea técnicamente factible sin comprometer la funcionalidad esencial del sistema.

Rendición de cuentas	• Establecer períodos de retención de datos más restrictivos para información de NNA, eliminándola tan pronto como se cumplan los propósitos declarados. • Garantizar que los datos de NNA se almacenen con las más altas medidas de seguridad técnica y organizacional, segregados de otros conjuntos de datos institucionales y con acceso restringido mediante controles de autorización multinivel. • Verificar que los datasets no contengan sesgos que puedan perpetuar discriminación contra NNA de grupos vulnerables (población rural, comunidades étnicas, niños y niñas con discapacidad, víctimas de violencia) mediante análisis desagregado específico por estas categorías (Department of Public Expenditure, NDP Delivery and Reform, 2024; Dapre, 2021).

Fuente: Elaboración de los autores

7.5 Tercera fase: Construcción de modelos

El objetivo de esta fase es establecer modelos que se caractericen por su robustez técnica, seguridad y equidad en el desarrollo algorítmico, y garantía de capacidad de explicación y supervisión humana sobre el entrenamiento.

Tabla 12. Principios y acciones fase de construcción de modelos

Principio	Acciones
Agencia y supervisión humana	• Diseñar modelos de manera que permitan evaluación humana experta de salidas durante el entrenamiento. • Implementar mecanismos de inspección de predicciones en muestras de desarrollo para identificar patrones problemáticos antes del despliegue.
Robustez técnica y seguridad	• Asegurar que el entorno de desarrollo esté protegido contra accesos no autorizados y exfiltración de modelos o datos. • Entrenar modelos considerando diversidad de condiciones operativas para garantizar estabilidad y generalización (manejo robusto de valores atípicos, casos edge, distribuciones de datos cambiantes). • Implementar técnicas de validación cruzada, conjuntos de prueba independientes y evaluación en distribuciones *out-of-sample.*
Privacidad y gobernanza de datos	• Mantener protocolos de protección de datos durante el entrenamiento, y considerar técnicas de aprendizaje que preserven privacidad (*federated learning, differential privacy*) cuando se manejen datos altamente sensibles.
Transparencia y explicabilidad	• Asegurar que el modelo cumpla con requisitos de explicabilidad establecidos en la fase de diseño. • Para modelos complejos tipo caja negra donde la explicabilidad intrínseca es limitada, documentar exhaustivamente las razones técnicas que justifican esta opción y compensar mediante mayor supervisión humana y técnicas de explicabilidad *post-hoc (SHAP values, LIME, attention mechanisms).*

	• Producir model cards o fichas técnicas de modelos que documenten: arquitectura, hiperparámetros, proceso de entrenamiento, métricas de desempeño, limitaciones conocidas y casos de uso apropiados e inapropiados (*Department of Public Expenditure, NDP Delivery and Reform, 2024)*.
Equidad, no discriminación y justicia	• Realizar pruebas continuas de sesgo durante el desarrollo mediante análisis desagregado de métricas de desempeño por grupos demográficos. • Utilizar métricas de equidad algorítmica (paridad demográfica, igualdad de oportunidades, igualdad predictiva, calibración por grupos) para evaluar si el modelo trata equitativamente a diferentes subpoblaciones. • Implementar técnicas de mitigación de sesgo en entrenamiento (*re-weighting, adversarial debiasing, constraint-based optimization*) cuando se detecten disparidades inaceptables.
Bienestar social y ambiental	• Considerar impactos ambientales del entrenamiento de modelos, particularmente para modelos de aprendizaje profundo que requieren recursos computacionales intensivos. • Evaluar huella de carbono del entrenamiento y seleccionar arquitecturas más eficientes cuando sea factible sin sacrificar desempeño crítico. • Evaluar impactos potenciales sobre bienestar físico y mental de usuarios y funcionarios que interactuarán con el sistema, así como implicaciones sociales más amplias del uso de la tecnología.
Rendición de cuentas	• Seguir rigurosamente el marco de responsabilidad establecido en diseño. • Mantener registros de auditoría de experimentos de entrenamiento (versiones de datos, configuraciones de modelos, resultados de evaluaciones). • Continuar evaluaciones de impacto con información actualizada. • Documentar procesos para reportar problemas detectados y mecanismos de respuesta ante hallazgos de sesgo, bajo desempeño o comportamientos no deseados (OECD, 2019, 2024).
Prevalencia de los derechos de NNA	• Cuando el sistema pueda afectar decisiones sobre NNA, implementar auditorías especializadas de equidad algorítmica que evalúen específicamente disparidades de desempeño según edad, género, condición socioeconómica, discapacidad, pertenencia étnica u otras características protegidas de esta población. • Establecer umbrales de equidad más exigentes que para población adulta, reconociendo la mayor vulnerabilidad de NNA ante decisiones erróneas o sesgadas. • Diseñar el modelo para maximizar transparencia y explicabilidad cuando afecte derechos de NNA, evitando arquitecturas de caja negra excesivamente complejas salvo que exista justificación técnica robusta y se compense con supervisión humana intensificada. } • Documentar exhaustivamente en model cards las consideraciones específicas implementadas para proteger derechos de NNA, incluyendo métricas de equidad desagregadas por grupos etarios y análisis de potenciales impactos sobre desarrollo integral, educación, salud, integridad y vida familiar de esta población. • Prohibir el uso de técnicas de perfilamiento predictivo sobre NNA que puedan generar estigmatización, etiquetamiento permanente o limitación de oportunidades futuras basadas en características o comportamientos de la niñez (Parlamento Europeo y Consejo de la Unión Europea, 2024, art. 5; Dapre, 2021).

Fuente: Elaboración de los autores.

7.6 Cuarta fase: Verificación y validación

El objetivo de esta fase es garantizar que los modelos sean sólidos, seguros y confiables para su propósito previsto antes del despliegue en producción.

Tabla 13. Principios y acciones fase de verificación y validación

Principio	Acciones
Agencia y supervisión humana	• Probar que exista involucramiento humano significativo y control efectivo sobre las operaciones del sistema. • Validar que los mecanismos de anulación humana de decisiones algorítmicas funcionen correctamente.
Robustez técnica y seguridad	• Realizar pruebas de penetración y análisis de vulnerabilidades para identificar aplicaciones no previstas, potencial de abuso malicioso o vectores de ataque. • Garantizar que el sistema no cause daños bajo condiciones adversas o inputs adversariales. • Implementar pruebas en entornos piloto controlados que simulen condiciones operativas reales antes del despliegue completo. • Verificar confiabilidad (*reliability*) mediante pruebas de estrés, reproducibilidad de resultados bajo condiciones idénticas, y consistencia de comportamiento en el tiempo. • Seguir procedimientos de seguridad establecidos institucionalmente, incluyendo revisiones de código por pares y auditorías de seguridad independientes para sistemas de alto riesgo (*Department of Public Expenditure, NDP Delivery and Reform, 2024).*
Privacidad y gobernanza de dato	• Mantener procesos de protección de datos durante la validación. • Verificar que no existan fugas inadvertidas de información sensible a través de salidas del modelo (model inversion attacks, membership inference).
Transparencia y explicabilidad	• Probar exhaustivamente las capacidades de explicabilidad del modelo mediante evaluación con usuarios finales representativos (funcionarios que usarán el sistema, ciudadanos potencialmente afectados). • Documentar métodos de prueba y resultados de validación de manera accesible para *stakeholders downstream* (responsables de gobernanza, auditores, supervisores jurídicos).
Transparencia y explicabilidad	• Realizar pruebas rigurosas de sesgo en datasets de validación independientes que reflejen diversidad poblacional real. • Verificar que métricas de equidad se mantengan dentro de umbrales aceptables establecidos en la fase de diseño. • Involucrar a representantes de grupos potencialmente afectados en la validación del sistema.

	• Asegurar que el diseño cumpla con principios de accesibilidad universal, realizando pruebas de usabilidad con personas con distintos tipos de discapacidad (visual, auditiva, motora, cognitiva) conforme a estándares WCAG 2.1 nivel AA.
Bienestar social y ambiental	• Evaluar si los beneficios sociales anticipados del sistema superan potenciales daños identificados durante la validación. • Documentar impactos sobre bienestar de individuos y comunidades, incluyendo efectos sobre autonomía, dignidad, salud mental, cohesión social y medio ambiente.
Rendición de cuentas	• Mantener registros completos de pruebas y validaciones realizadas. • Actualizar evaluaciones de impacto con resultados de validación. • Verificar que existan mecanismos funcionales de reporte de problemas y respuesta ante hallazgos críticos. • Establecer umbrales de desempeño mínimos que deben cumplirse obligatoriamente antes de autorizar el despliegue (Dapre, 2021).
Prevalencia de los derechos de NNA	• Realizar pruebas específicas de validación con casos que involucren NNA, asegurando que el sistema opere con los más altos estándares de precisión, equidad y seguridad para esta población. • Involucrar a especialistas en derechos de la infancia, psicólogos infantiles, pedagogos y representantes de organizaciones de defensa de derechos de NNA en la evaluación del sistema. • Conducir pruebas de usabilidad y accesibilidad específicas para distintos rangos etarios cuando NNA sean usuarios directos, garantizando que interfaces y comunicaciones sean apropiadas a su nivel de desarrollo cognitivo y madurez. • Validar rigurosamente que el sistema no genere riesgos de daño psicológico, manipulación, explotación o exposición a contenidos inapropiados. • Verificar el funcionamiento efectivo de todos los mecanismos de protección especial diseñados para NNA, incluyendo controles de privacidad reforzados, supervisión humana intensificada y procesos de escalamiento inmediato ante detección de situaciones de riesgo. • Realizar análisis de escenarios adversos específicos que evalúen qué podría ocurrir si el sistema es usado de manera no prevista por o contra NNA, estableciendo salvaguardas ante estos riesgos (*Department of Public Expenditure, NDP Delivery and Reform, 2024)*.

Fuente: Elaboración de los autores.

7.7 Quinta fase: Despliegue

El propósito de esta fase es garantizar operacionalización apropiada del sistema, gestión del cambio organizacional, capacitación de usuarios, documentación exhaustiva y supervisión continua post-despliegue.

Tabla 14. Principios y acciones fase de despliegue

Principio	Acciones
Agencia y supervisión humana	• Implementar los niveles de supervisión humana definidos en la fase de diseño, garantizando que funcionarios capacitados puedan monitorear, intervenir y anular decisiones algorítmicas cuando sea necesario. • Proveer información clara a usuarios finales sobre cómo funciona el sistema y ofrecer procesos alternativos de decisión basados en evaluación humana directa cuando se afecten derechos humanos. • Ofrecer capacitación integral y soporte técnico continuo a usuarios finales para garantizar uso competente y crítico del sistema, evitando automatización excesiva o dependencia acrítica de recomendaciones algorítmicas (*Department of Public Expenditure, NDP Delivery and Reform, 2024*).
Robustez técnica y seguridad	• Monitorear el desempeño del sistema en condiciones operativas reales, detectando degradación de métricas, cambios en distribución de datos (*data drift, concept drift*) o comportamientos anómalos. • Activar planes de contingencia ante fallos técnicos, garantizando continuidad de servicios mediante procedimientos manuales alternativos. • Producir documentación técnica de despliegue que incluya: requisitos de infraestructura, procedimientos de instalación y configuración, dependencias de sistemas, protocolos de *backup* y recuperación.
Privacidad y gobernanza de datos:	• Garantizar que el despliegue mantenga medidas de seguridad de datos establecidas: cifrado, controles de acceso basados en roles, segregación de ambientes. • Restringir acceso a información sensible según principio de mínimo privilegio (*least privilege*).
Transparencia y explicabilidad	• Implementar mecanismos de trazabilidad que documenten automáticamente decisiones algorítmicas individuales con suficiente detalle para auditorías posteriores. • Comunicar claramente a usuarios la influencia del modelo en procesos de toma de decisiones y la racionalidad detrás de recomendaciones específicas. Informar explícitamente a ciudadanos cuando estén interactuando con sistemas de IA, cumpliendo con obligaciones de transparencia establecidas en el Marco ético nacional y la Ley 1712 de 2014 de Transparencia y Acceso a la Información Pública. • Ofrecer alternativas de interacción humana cuando usuarios prefieran no utilizar sistemas automatizados (Dapre, 2021).
Equidad, no discriminación y justicia	• Monitorear continuamente decisiones post-despliegue para detectar patrones de inequidad que no fueron evidentes en validación. • Establecer procesos de revisión periódica de métricas de equidad desagregadas. • Mantener canales de participación activa con partes interesadas afectadas para recibir retroalimentación sobre experiencias de uso y potenciales impactos discriminatorios.

Bienestar social y ambiental	• Verificar mediante evaluación preliminar que los beneficios sociales del despliegue superen los riesgos y costos potenciales. • Establecer mecanismos de monitoreo de impactos sociales y ambientales no anticipados.
Rendición de cuentas	• Mantener estructuras de supervisión, actualización continua de evaluaciones de impacto, y procesos de reporte y respuesta ante incidentes. • Proveer guías de usuario que incluyan consideraciones éticas y limitaciones conocidas del sistema. • Designar puntos de contacto institucionales responsables de atender consultas éticas y quejas relacionadas con el funcionamiento del sistema (*Department of Public Expenditure, NDP Delivery and Reform, 2024; OECD, 2019, 2024*).
Prevalencia de los derechos de NNA	• Implementar protocolos de supervisión humana reforzada cuando el sistema opere en contextos que afecten NNA, garantizando que funcionarios especializados con formación en derechos de la infancia revisen todas las decisiones algorítmicas de alto impacto antes de su aplicación. • Establecer canales de comunicación específicos para que NNA, sus familias o representantes puedan ejercer derechos de información, acceso, rectificación y oposición sobre tratamiento de datos, utilizando lenguaje claro, accesible y apropiado para distintas edades. • Proveer capacitación especializada a funcionarios que utilizarán el sistema en casos que involucren NNA, enfatizando la obligación constitucional y legal de garantizar prevalencia del interés superior del niño sobre consideraciones de eficiencia administrativa o recomendaciones algorítmicas. • Implementar mecanismos de alerta temprana que señalen automáticamente casos que involucren NNA en situaciones de especial vulnerabilidad (víctimas de violencia, riesgo de reclutamiento, explotación sexual, trabajo infantil, desplazamiento forzado) para activar protocolos de protección urgente. • Informar claramente a NNA y sus familias cuando interactúen con sistemas de IA, explicando en lenguaje comprensible cómo funciona, qué decisiones apoya, cuáles son sus limitaciones y cómo acceder a revisión humana de decisiones (Dapre, 2021; Ley 1098, 2006).

Fuente: Elaboración de los autores.

7.8 Sexta fase: Operación y monitoreo

Esta fase se enfoca en el monitoreo continuo, la participación sostenida de partes interesadas, el mantenimiento proactivo con énfasis en cumplimiento normativo permanente y el proceso de mejora continua.

Tabla 15. . Principios y acciones fase de operación y monitoreo

Principio	Acciones
Agencia y supervisión humana	• Monitorear sistemáticamente que los mecanismos de control humano continúen funcionando efectivamente. • Revisar y actuar sobre retroalimentación de usuarios y ciudadanos afectados. • Realizar auditorías periódicas de casos donde se ejerció anulación humana de decisiones algorítmicas para identificar patrones de problemas recurrentes.
Robustez técnica y seguridad	• Implementar monitoreo automatizado de métricas de desempeño (precisión, recall, F1-score, AUC-ROC, según el caso de uso) para detectar degradación temprana. • Monitorear indicadores de contaminación de datos, drift distribucional y cambios en patrones de input que puedan señalar problemas. Verificar continuamente ausencia de daños imprevistos o efectos adversos sobre usuarios y ciudadanos.
Privacidad y gobernanza de datos:	• Mantener cumplimiento continuo de regulaciones de protección de datos. • Realizar auditorías periódicas de acceso a datos y uso de información personal. • Actualizar evaluaciones de impacto sobre protección de datos cuando cambien las condiciones de procesamiento.
Transparencia y explicabilidad	• Verificar que usuarios finales comprendan efectivamente su interacción con el sistema de IA. • Adaptar estrategias de comunicación y materiales de capacitación según retroalimentación recibida y problemas de comprensión identificados.
Equidad, no discriminación y justicia	• Realizar pruebas post-despliegue periódicas para verificar que se mantengan estándares de equidad. • Monitorear métricas de accesibilidad y uso por distintos grupos poblacionales. • Sostener procesos de participación de partes interesadas, incluyendo mecanismos de quejas y reparación cuando se identifiquen impactos discriminatorios (Dapre, 2021).
Bienestar social y ambiental	• Monitorear impactos sociales y ambientales continuos del sistema, incluyendo efectos sobre empleo, acceso a servicios, confianza ciudadana y cohesión social. • Evaluar periódicamente si los beneficios originalmente anticipados se están materializando en la práctica.
Rendición de cuentas	• Garantizar auditabilidad completa del sistema mediante registros exhaustivos de operaciones. • Para aplicaciones que afecten derechos humanos o sean críticas para la seguridad, permitir y facilitar auditorías independientes por terceros especializados. • Mantener canales de comunicación abiertos con entidades de control (Superintendencia de Industria y Comercio, Ministerio Público [Defensoría del Pueblo, Procuraduría General, Personerías]) para atender requerimientos de información y colaborar en investigaciones cuando sea necesario (*Department of Public Expenditure, NDP Delivery and Reform, 2024*).

Prevalencia de los derechos de NNA	• Establecer monitoreo continuo específico de casos que involucren NNA, analizando desagregadamente métricas de desempeño, equidad y satisfacción para esta población. • Mantener el funcionamiento del comité de protección de NNA en IA como instancia de supervisión especializada que revise periódicamente el impacto del sistema sobre derechos de la infancia y adolescencia, con participación de defensores de derechos de NNA, representantes de organizaciones especializadas y, cuando sea apropiado, adolescentes que puedan aportar su perspectiva. • Realizar auditorías especializadas semestrales que evalúen específicamente el cumplimiento de salvaguardas para NNA, incluyendo revisión de casos donde se ejerció anulación humana de decisiones algorítmicas, análisis de quejas o reclamos presentados por familias, y evaluación de impactos no anticipados sobre esta población. • Mantener actualizado el sandbox regulatorio como espacio de prueba controlada para innovaciones que afecten NNA, permitiendo experimentación responsable con máxima protección de derechos. • Documentar y analizar sistemáticamente lecciones aprendidas de la operación del sistema en casos de NNA, retroalimentando estos hallazgos a las fases de diseño y desarrollo de futuras actualizaciones. • Garantizar que cualquier cambio significativo en el sistema que pueda afectar NNA sea sometido nuevamente a evaluación de impacto sobre derechos de la niñez actualizada antes de su implementación (*Department of Public Expenditure, NDP Delivery and Reform, 2024; Dapre, 2021*).

Fuente: Elaboración de los autores

7.9 Séptima fase: Retiro, desmantelamiento o actualización

Esta fase se enfoca en el manejo seguro, responsable y transparente del retiro o actualización de sistemas de IA, garantizando protección de datos y continuidad de servicios.

Tabla 16. Principios y acciones fase de retiro, desmantelamiento o actualización

Principio	Acciones
Agencia y supervisión humana	• Someter la decisión de desmantelamiento o actualización mayor a revisión humana experta y aprobación por autoridades institucionales competentes. • Evaluar impactos del retiro sobre usuarios, ciudadanos y operaciones institucionales.
Robustez técnica y seguridad	• Seguir estrictamente políticas institucionales de seguridad de datos durante el proceso de desmantelamiento. • Preparar procedimientos para desconexión segura del sistema, garantizando que no se generen vulnerabilidades de seguridad o pérdidas de información crítica durante la transición.
Privacidad y gobernanza de datos:	• Implementar manejo seguro de datos post desmantelamiento, incluyendo eliminación certificada de información personal conforme a políticas de retención establecidas y obligaciones del Régimen General de Protección de Datos (Ley 1581, 2012). • Cuando la normativa exija conservación de datos para fines de archivo histórico, investigación o rendición de cuentas, garantizar que el almacenamiento cumpla con medidas de seguridad reforzadas y acceso estrictamente controlado.
Transparencia y explicabilidad	• Documentar exhaustivamente la justificación para el desmantelamiento o actualización mayor del sistema, incluyendo razones técnicas, éticas, jurídicas u organizacionales. • Informar oportunamente a partes interesadas (usuarios finales, ciudadanos afectados, entidades aliadas) sobre el retiro del sistema y sugerir alternativas de servicio o soportes sustitutos (*Department of Public Expenditure, NDP Delivery and Reform, 2024*).
Equidad, no discriminación y justicia	• Garantizar que el proceso de desmantelamiento sea equitativo para todas las partes interesadas, sin generar cargas desproporcionadas sobre grupos específicos. • Asegurar que alternativas propuestas sean accesibles para todos los usuarios anteriores del sistema.
Bienestar social y ambiental	• Realizar el desmantelamiento de manera ambientalmente responsable, incluyendo disposición apropiada de hardware y gestión de residuos electrónicos conforme a normativa ambiental. • Evaluar impactos sociales del retiro y establecer medidas de mitigación cuando sea necesario.
Rendición de cuentas	• Designar una persona o equipo responsable del proceso de desmantelamiento con rendición de cuentas clara. • Actualizar toda la documentación del sistema para reflejar su estatus de retirado. • Archivar documentación técnica, evaluaciones de impacto, registros de auditoría y lecciones aprendidas para consulta futura y como insumo para mejora de procesos institucionales (Dapre, 2021; OECD, 2019, 2024).

Prevalencia de los derechos de NNA	• Someter la decisión de retiro o actualización mayor de sistemas que procesen datos de NNA a revisión específica del comité de protección de NNA en IA, evaluando impactos particulares sobre esta población. • Garantizar eliminación certificada y verificable de todos los datos de NNA conforme a estándares más exigentes que para población adulta, incluyendo destrucción segura de *backups*, registros de auditoría que contengan información identificable y cualquier dato derivado o agregado que permita re-identificación. • Cuando la normativa exija conservación de datos para fines de archivo histórico o investigación, implementar medidas de anonimización irreversible mediante técnicas criptográficas robustas que imposibiliten la re-identificación de NNA. • Informar oportunamente a familias y representantes legales sobre el retiro del sistema cuando este haya procesado datos de sus hijos o tutelados, explicando qué ocurrirá con la información recolectada y cómo se garantiza su eliminación o anonimización. • Asegurar que alternativas de servicio propuestas tras el desmantelamiento mantengan o mejoren los niveles de protección de derechos de NNA que proveía el sistema retirado. • Documentar exhaustivamente lecciones aprendidas sobre protección de derechos de NNA durante el ciclo de vida del sistema, generando conocimiento institucional que fortalezca el diseño de futuras iniciativas tecnológicas que afecten esta población. • Archivar toda la documentación relacionada con salvaguardas implementadas para NNA como evidencia de cumplimiento de obligaciones constitucionales y legales de protección reforzada de la infancia y adolescencia (Ley 1098, 2006; Dapre, 2021; Parlamento Europeo y Consejo de la

Fuente: Elaboración de los autores.

Las orientaciones presentadas en este capítulo constituyen un marco flexible para que los equipos del Ministerio Público las adapten según las necesidades específicas y niveles de riesgo de cada proyecto de IA. Si bien no son prácticas de aplicación obligatoria universal, ofrecen una hoja de ruta valiosa para alinear esfuerzos con principios de responsabilidad y fomentar la confianza pública. Estas prácticas buscan ayudar a gestionar la IA de manera responsable, manteniendo el bienestar ciudadano y la excelencia en el servicio público como prioridades fundamentales de toda iniciativa tecnológica institucional.

8. Orientaciones finales para usuarios de la IA generativa

La proliferación de sistemas de inteligencia artificial generativa (GenAI) accesibles públicamente plantea desafíos específicos de gobernanza, seguridad y ética para el sector público. Este capítulo ofrece orientaciones para funcionarios del Ministerio Público sobre requisitos de calidad de datos, transparencia, protección de información sensible, validación obligatoria, supervisión humana y mitigación de sesgos, permitiendo aprovechamiento responsable mientras se salvaguardan derechos humanos y confianza pública (Department of Public Expenditure, NDP Delivery and Reform, 2024; Dapre, 2021).

8.1 Advertencias preliminares y requisitos de aprobación

No debe incorporarse IA generativa en procesos institucionales sin que la propuesta haya sido formalmente aprobada conforme a los principios establecidos (Canvas de IA responsable, capítulo 6; marco de decisión, capítulo 5). No debe permitirse acceso generalizado de funcionarios a herramientas de GenAI hasta que las dependencias hayan realizado evaluaciones exhaustivas de riesgos, establecido políticas institucionales claras e implementado programas de capacitación sobre uso seguro. La experiencia internacional documenta fugas inadvertidas de información clasificada, violaciones de privacidad y generación de contenido sesgado derivados del uso no regulado, justificando controles preventivos institucionales (Department of Public Expenditure, NDP Delivery and Reform, 2024; National Cyber Security Centre, 2023).

8.2 Requisitos de calidad de datos

La efectividad de GenAI depende críticamente de la calidad de datos subyacentes. Datos de baja calidad producen salidas inexactas, sesgadas o engañosas. Los datos deben satisfacer: (i) ser apropiados para el propósito (completitud, relevancia, exclusión de información sensible por razones legales/éticas/regulatorias, vigencia temporal, exhaustividad, ausencia de sesgos históricos), y (ii) ser precisos (sin duplicaciones, correctamente etiquetados, valores exactos, identificadores consistentes). Debe asignarse responsabilidad formal de garantizar calidad desde el origen mediante procesos estructurados y rendición de cuentas. Líderes institucionales deben exigir aseguramiento de calidad mediante revisiones independientes periódicas (Department of Public Expenditure, NDP Delivery and Reform, 2024; Dapre, 2021).

8.3 Riesgos de herramientas de IA generativa de acceso público

Las herramientas gratuitas de GenAI públicamente accesibles carecen de gestión institucional y supervisión adecuadas, generando riesgos inaceptables, por ello es necesario ser conscientes de que cualquier información introducida puede utilizarse para entrenar modelos sin control sobre uso posterior. Se recomienda fundamentalmente no usar usar herramientas públicas de GenAI para procesar información institucional del Ministerio Público.

Esta prohibición obedece a que la información que la organización no desea hacer pública jamás debe introducirse en modelos públicos. Esto incluye: información clasificada (Ley 1712 de 2014, arts. 18, 19, 24); datos personales (Ley 1581 de 2012); información comercial sensible; deliberaciones sobre investigaciones en curso; estrategias institucionales no públicas, y cualquier información cuya divulgación pueda comprometer la misión institucional o derechos de personas (National Cyber Security Centre, 2023).

8.4 Mejores prácticas para usuarios autorizados de IA generativa

Cuando el Ministerio Público autorice formalmente herramientas de GenAI mediante contratación con garantías contractuales apropiadas (acuerdos de nivel de servicio, protección de datos, compromisos de no uso para entrenamiento, auditoría), los usuarios deben adherirse a las siguientes prácticas:

• **Transparencia y divulgación:** obligación de revelar explícitamente cuando el contenido fue generado por IA, especialmente en comunicaciones oficiales e interacciones con ciudadanos, preservando confianza pública y permitiendo evaluación crítica (principio 4; Ley 1712, 2014). Comunicar claramente limitaciones conocidas (inexactitudes factuales/alucinaciones, carencia de comprensión contextual, sesgos, restricciones temporales). La validación y verificación humana exhaustiva es obligatoria antes de uso en decisiones o comunicaciones (Department of Public Expenditure, NDP Delivery and Reform, 2024; Dapre, 2021).

• **Privacidad y sensibilidad de datos:** prohibido introducir en GenAI —incluso en sistemas contratados— datos personales sensibles (salud, orientación sexual, afiliación política, origen étnico, creencias religiosas, antecedentes penales), información clasificada, secreto profesional constitucionalmente protegido, información comercial propietaria o datos cuya filtración genere daños. Esta restricción aplica incluso con cláusulas de confidencialidad, dado que riesgos por vulnerabilidades, errores o ataques nunca se eliminan completamente. Cuando sea indispensable usar GenAI sobre información sensible, aplicar anonimización/seudonimización irreversible y procesar en ambientes controlados con máximas medidas de seguridad, cumpliendo la Ley 1581/2012 y el Decreto 1377/2013 (Ley 1581, 2012; Department of Public Expenditure, NDP Delivery and Reform, 2024).

• **Validación y verificación obligatoria:** validar exhaustivamente toda salida antes del uso en decisiones, investigaciones o comunicaciones. Los modelos cometen errores factuales, generan información plausible pero falsa, carecen de comprensión contextual profunda y producen salidas inconsistentes. Toda validación debe incluir la verificación de exactitud factual, evaluación de coherencia lógica, análisis de pertinencia contextual, detección de sesgos/estereotipos/lenguaje discriminatorio, y valoración crítica de razonabilidad según criterio profesional.

Ninguna decisión con efectos jurídicos puede fundamentarse exclusivamente en salidas de GenAI sin validación humana (Dapre, 2021; C.P., 1991, art. 29).

• **Agencia y supervisión humana:** la GenAI complementa —nunca sustituye— juicio humano profesional, experiencia especializada y razonamiento ético. Los usuarios tienen responsabilidad ineludible de evaluar críticamente salidas, ejerciendo escepticismo informado y manteniendo control efectivo. La responsabilidad institucional permanece en funcionarios humanos. Debe implementarse supervisión estructurada para revisar y aprobar contenido antes de difusión, especialmente en áreas críticas que afecten derechos humanos o funciones constitucionales, ejercida por funcionarios con competencia técnica y autoridad suficiente (Department of Public Expenditure, NDP Delivery and Reform, 2024).

• **Mitigación de sesgos y promoción de equidad:** el contenido puede reflejar y amplificar sesgos de datos de entrenamiento, perpetuando discriminaciones históricas y estereotipos. Obligación de revisar salidas para detectar lenguaje, suposiciones o recomendaciones excluyentes, inequitativas o que perpetúen estereotipos relacionados con género, origen étnico, edad, discapacidad, orientación sexual, ubicación geográfica, nivel socioeconómico u otras características protegidas. Cuando se detecten sesgos, ajustar manualmente para garantizar inclusividad y equidad. Aplicar filtros de posprocesamiento cuando sea factible para detectar/modificar automáticamente lenguaje sesgado, operacionalizando el principio 3 de diversidad, no discriminación y equidad (Dapre, 2021).

El uso responsable de la IA generativa requiere establecer un equilibrio entre aprovechar capacidades transformadoras de la actual revolución industrial y salvaguardar rigurosamente los derechos humanos, la seguridad institucional y la confianza pública. Estas orientaciones establecen una base mínima de prácticas aceptables; las diferentes dependencias tienen el deber de desarrollar controles adicionales según sus funciones y riesgos específicos. La actualización continua será necesaria conforme evolucionen capacidades tecnológicas, emerjan nuevos riesgos y se fortalezca el aprendizaje institucional.

Apéndices

Apéndice A. Marco normativo aplicable

Tabla 17. De correspondencia entre principios de gobernanza y marco normativo

Principio de gobernanza	Instrumento normativo	Artículo / Sección	Contenido relevante
Supervisión y control humano	Constitución Política de Colombia	Art. 6	Responsabilidad de los servidores públicos por infringir la Constitución y las leyes, y por omisión o extralimitación en el ejercicio de sus funciones.
	Marco ético para la IA en Colombia (Dapre, 2021)	Principio 3	Control humano de las decisiones propias de un sistema de inteligencia artificial (*human-in-the-loop y human-over-the-loop*).
	Directrices éticas para una IA fiable (UE). (High-Level Expert Group on Artificial Intelligence, 2019)	Requisito 1	Agencia y supervisión humanas.
	Reglamento 2024/1689 (AI Act). (Parlamento Europeo y Consejo de la Unión Europea, 2024)	Art. 14	Requisitos de supervisión humana para sistemas de IA de alto riesgo.
	Principios de IA de la OCDE (2019, 2024)	Principio 1.2	Los sistemas de IA deben incluir salvaguardias que permitan la intervención humana.
Robustez técnica y seguridad	Marco ético para la IA en Colombia (Dapre, 2021)	Principio 4	Seguridad: los sistemas de IA no deben generar afectaciones a la integridad y salud física y mental.
	Reglamento 2024/1689 (AI Act). (Parlamento Europeo y Consejo de la Unión Europea, 2024)	Art. 15	Precisión, robustez y ciberseguridad.
	Directrices éticas para una IA fiable (UE). (High-Level Expert Group on Artificial Intelligence, 2019)	Requisito 2	Robustez técnica y seguridad.
	Política de Seguridad de la Información Resolución 138 del 27 de junio de 2025 (PGN, 2025)	Todo el documento	Marco institucional para la gestión de la seguridad de la información.

Principio de gobernanza	Instrumento normativo	Artículo / Sección	Contenido relevante
Privacidad y gobernanza de datos	Constitución Política de Colombia	Art. 15	Derecho a la intimidad personal y familiar, al buen nombre y al habeas data.
	Ley 1581 de 2012	Todo	Régimen General de Protección de Datos Personales.
	Decreto 1377 de 2013	Todo	Reglamentación parcial de la Ley 1581 de 2012.
	Marco ético para la IA en Colombia (Dapre, 2021)	Principio 2	Privacidad.
	Reglamento 2024/1689 (AI Act). (Parlamento Europeo y Consejo de la Unión Europea, 2024)	Art. 10	Gobernanza y calidad de los datos.
	Política de Gobierno de Datos (PGN, 2024a)	Todo el documento	Marco institucional para la gestión y gobierno de datos.
Transparencia y explicabilidad	Constitución Política de Colombia	Art. 74	Derecho de acceso a los documentos públicos.
	Sentencia T-067 de 2025 (Corte Constitucional)	Considerandos y parte resolutiva	Define la transparencia algorítmica como elemento esencial del derecho de acceso a la información pública.
	Directiva Conjunta 007 de 2025 (PGN, Defensoría del Pueblo)	Todo	Estándares mínimos de transparencia algorítmica para el Estado.
	Ley 1712 de 2014	Arts. 3, 4, 5	Ley de Transparencia y del Derecho de Acceso a la Información Pública.
	Ley 1755 de 2015	Art. 14	Derecho de petición.
	Marco ético para la IA en Colombia (Dapre, 2021)	Principio 1	Transparencia y explicación.
	Reglamento 2024/1689 (AI Act). (Parlamento Europeo y Consejo de la Unión Europea, 2024)	Art. 13	Transparencia y provisión de información a los usuarios.
Equidad, no discriminación y justicia	Constitución Política de Colombia	Art. 13	Derecho a la igualdad y prohibición de la discriminación.
	Marco ético para la IA en Colombia (Dapre, 2021)	Principios 6 y 7	No discriminación - Inclusión.

Principio de gobernanza	Instrumento normativo	Artículo / Sección	Contenido relevante
	Directrices éticas para una IA fiable (UE). (High-Level Expert Group on Artificial Intelligence, 2019)	Requisito 5	Diversidad, no discriminación y equidad.
	Documento CONPES 4144 de 2025 (CONPES [DNP])	Eje 2	IA para un Gobierno y una sociedad más equitativos.
	Código de Procedimiento Penal (Ley 906 de 2004)	Arts. 4, 5 y 7	Igualdad, imparcialidad e inocencia en la administración de justicia.
Bienestar social y ambiental	Constitución Política de Colombia	Art. 2	Fines esenciales del Estado: servir a la comunidad, promover la prosperidad general y garantizar la efectividad de los principios, derechos y deberes.
	Constitución Política de Colombia	Art. 79, 80	Derecho a un ambiente sano y deber del Estado de proteger la diversidad e integridad del ambiente.
	Marco ético para la IA en Colombia (Dapre, 2021)	Principio 9	Beneficio Social.
	Directrices éticas para una IA fiable (UE) (High-Level Expert Group on Artificial Intelligence, 2019)	Requisito 6	Bienestar social y medioambiental.
	Objetivos de Desarrollo Sostenible (ODS)	ODS 16	Paz, justicia e instituciones sólidas.
Rendición de cuentas y responsabilidad	Constitución Política de Colombia	Art. 6	Responsabilidad de los servidores públicos.
	Constitución Política de Colombia	Art. 90	Responsabilidad patrimonial del Estado.
	Ley 1952 de 2019, y sus reformas	Todo	Código General Disciplinario.
	Ley 734 de 2002, y sus reformas	Todo	Código Disciplinario Único (parcialmente vigente).
	Directrices éticas para una IA fiable (UE). (High-Level Expert Group on Artificial Intelligence, 2019)	Principio 7	Rendición de cuentas.
	Principios de IA de la OCDE (2019, 2024)	Principio 1.5	Mecanismos para garantizar la responsabilidad y la rendición de cuentas.
	Constitución Política de Colombia (1991)	Arts. 44 y 45	Derechos fundamentales de los niños («Los derechos de los niños

Principio de gobernanza	Instrumento normativo	Artículo / Sección	Contenido relevante
Prevalencia de los derechos de niños, niñas y adolescentes (NNA)			prevalecen sobre los derechos de los demás») y protección y formación integral de los adolescentes.
	Ley 1098 de 2006	Todo	Código de la Infancia y la Adolescencia, que desarrolla el principio de interés superior del menor, establece derechos fundamentales de NNA y obligaciones de protección del Estado, la familia y la sociedad.
	Ley 1581 de 2012	Arts. 7 y 12	Derechos de los titulares (especial protección a datos de menores) y requisitos especiales para datos de menores (consentimiento de representantes legales).
	Convención sobre los Derechos del Niño (ONU, 1989; Ley 12, 1991)	Arts. 3, 12 y 16	Interés superior del niño (art. 3), derecho a ser escuchado (art. 12) y protección de la vida privada (art. 16).
	Observación general n.° 25 (Comité de los Derechos del Niño de la ONU, 2021)	Todo el documento	Directrices sobre los derechos de los niños en relación con el entorno digital, IA, algoritmos, protección de datos y participación de NNA en entornos tecnológicos.
	Marco ético para la IA en Colombia (Dapre, 2021)	Principio 8	«Prevalencia de los derechos de niños, niñas y adolescentes», establece directrices sobre ética de datos, ética de algoritmos y ética de prácticas aplicables a NNA.
	Reglamento 2024/1689 (AI Act). (Parlamento Europeo y Consejo de la Unión Europea, 2024)	Anexo III, numeral 3	Considera sistemas de IA en ámbito educativo o que afecten a menores como de «alto riesgo», requiriendo evaluaciones de conformidad, supervisión humana y transparencia.
	Principios de IA de la OCDE (2019, 2024)	Principio 1.1	«Crecimiento inclusivo, desarrollo sostenible y bienestar», que incluye consideración especial de impactos sobre grupos vulnerables como NNA.

Fuente. Elaboración de los autores.

Tabla 18. Normativa nacional adicional aplicable

Instrumento	Año	Relevancia para la gobernanza de IA
Decreto 262 de 2000. Modificado por el Decreto Ley 1851 de 2021	2000	Estructura y la organización de la Procuraduría General de la Nación y del Instituto de Estudios del Ministerio Público.
Ley 2422 de 2024	2024	Por medio de la cual se dictan disposiciones para fortalecer el funcionamiento de las personerías en Colombia.
Ley 24 de 1992	1992	Organización y funcionamiento de la Defensoría del Pueblo.
Documento CONPES 3975, Documento CONPES 4144 (CONPES [DNP] 2019, 2025) - Politica Nacional de Transformación Digital y Política Nacional de Inteligencia Artificial.	2019- 2025	Política Nacional para la Transformación Digital e Inteligencia Artificial.
Decreto 1263 de 2022	2022	Reglamentación en materia de transformación digital.
Ley 2502 de 2025	2025	Falsedad personal mediante el uso de IA.
Proyecto de ley 043 de 2025	2025	Marco regulatorio específico para la IA en Colombia (en trámite legislativo).
PETI (PGN, 2024b)	2024	Plan Estratégico de Tecnologías de la Información de la Procuraduría.
Política de Tecnologías de la Información (PGN, 2024c)	2024	Marco de políticas de TI de la Procuraduría.

Fuente. Elaboración de los autores.

Tabla 19. Normativa internacional de referencia

Instrumento	Organización	Año	Relevancia
Reglamento 2024/1689 (AI Act)	Parlamento Europeo y Consejo de la Unión Europea	2024	Primer marco regulatorio integral de IA a nivel mundial, basado en riesgos.
Directrices éticas para una IA fiable (UE)	Comisión Europea (*High-Level Expert Group on Artificial Intelligence*) OCDE	2019	Establece los siete principios fundamentales para una IA fiable.
Principios de IA de la OCDE	UNESCO	2019 (actualizados 2024)	Adoptados por Colombia, promueven una IA innovadora y fiable.
Recomendación sobre la ética de la IA	Unión Europea	2021	Marco global de valores y principios éticos para la IA.
GDPR (Reglamento General de Protección de Datos)	Parlamento Europeo, Consejo de la Unión Europea	2016	Referente internacional en protección de datos personales.

Fuente. Elaboración de los autores.

Apéndice B. Lienzo o canvas de IA responsable Ministerio Público de Colombia

Canvas de IA responsable - Ministerio Público de Colombia

Introducción al Taller de Ideación

Estimados participantes,

En el marco del proyecto «Diseño de una Estrategia Integral para el Uso Ético de la Inteligencia Artificial en el Ministerio Público», desarrollado por el Instituto de Estudios del Ministerio Público, les damos la bienvenida a esta sesión de ideación.

El propósito de este formulario es guiarnos de manera estructurada a través del Canvas de IA responsable, una herramienta diseñada para planificar proyectos de IA desde una perspectiva ética, legal y centrada en el ser humano.

¿Cómo usar este canvas?

Este formulario digitaliza el Canvas para facilitar nuestro taller colaborativo. A medida que avancemos, discutiremos cada sección para:

- Asegurar una visión multifuncional: integrando perspectivas técnicas, legales y misionales desde el inicio.
- Gestionar riesgos de forma proactiva: identificando y planificando la mitigación de posibles amenazas antes de que se materialicen.
- Garantizar el cumplimiento normativo: alineando el diseño de la solución con el marco legal colombiano y los derechos humanos.
- Definir una estrategia clara: desde la definición del problema hasta la comunicación y la asignación de responsabilidades.

Por favor, diligencien cada sección con el mayor detalle posible basándose en la discusión del equipo.

Nota de tratamiento de datos personales (*habeas data*):

De conformidad con la Ley 1581 de 2012 de Protección de Datos Personales, informamos que la información recopilada en este formulario será utilizada exclusivamente para los fines académicos y de planificación del proyecto de investigación «Diseño de una Estrategia Integral para el Uso Ético de la Inteligencia Artificial en el Ministerio Público». Los datos serán tratados con estricta confidencialidad por el equipo investigador del IEMP. Su participación es voluntaria y al completar este formulario, usted autoriza el tratamiento de la información proporcionada para los fines aquí descritos.

Sección 1 - Información general del proyecto de IA generativa para el Ministerio Público

- Pregunta: nombre del proyecto.
  - Tipo: respuesta corta.
  - Descripción: p. ej. sistema de análisis inteligente de casos penales.
- Pregunta: líder del proyecto.
  - Tipo: respuesta corta.
  - Descripción: indique el nombre y cargo del responsable principal.
- Pregunta: fecha de inicio.
  - Tipo: fecha.
- Pregunta: ¿A cuál función misional impacta este proyecto?
  - Tipo: respuesta corta.
- Pregunta: ¿Cuál es el objetivo/s del proyecto?
  - Tipo: respuesta corta.

Sección 2 - Partes interesadas

- Pregunta: ¿Quiénes son las partes interesadas clave afectadas por este proyecto de IA?
  - Tipo: párrafo.
  - Descripción: p. ej. fiscales, investigadores, ciudadanos, víctimas, procesados, defensores públicos.
- Pregunta: ¿Quiénes son los aliados internos?
  - Tipo: párrafo.
  - Descripción: p. ej. empleados, equipos técnicos de la Oficina de Sistemas, directivos de la PGN, analistas de la DAE.
- Pregunta: ¿Quiénes son los aliados externos?
  - Tipo: párrafo.
  - Descripción: p. ej. ciudadanía en general, las ONG de derechos humanos, Rama Judicial, Defensoría del Pueblo, personerías.
- Pregunta: ¿Grupos de impacto especial?
  - Tipo: párrafo.
  - Descripción: identifique si el proyecto afecta de manera particular a grupos vulnerables, minorías étnicas, mujeres, menores, etc.

Sección 3 - Categorización de riesgo del proyecto

- Pregunta: según la normativa colombiana y estándares internacionales, ¿cuáles son los riesgos potenciales y cómo se mitigarán?
  - Tipo: párrafo.
  - Descripción: explique brevemente por qué se ha seleccionado ese nivel de riesgo, considerando los posibles impactos en los ciudadanos y la institución.

Sección 4 - Declaración del problema y justificación de IA

- Pregunta: declaración del problema, ¿qué problema específico estamos resolviendo?
  - Tipo: párrafo.
  - Descripción: p. ej. los fiscales enfrentan un volumen creciente de casos complejos que requieren análisis extenso de evidencias, generando retrasos en los procesos judiciales y afectando el acceso efectivo a la justicia.
- Pregunta: justificación de IA, ¿por qué la inteligencia artificial es la mejor solución para este problema?
  - Tipo: párrafo.
  - Descripción: p. ej. la IA puede procesar grandes volúmenes de información, identificar patrones complejos que pasarían desapercibidos para un humano y asistir en la priorización de casos según criterios objetivos.
- Pregunta: ¿Qué tipo de solución de IA es la más adecuada?
  - Tipo: párrafo.
  - Descripción: p. ej. modelos de clasificación de documentos, procesamiento de lenguaje natural (NLP) para análisis de testimonios, sistemas de recomendación de jurisprudencia.
- Pregunta: alternativas consideradas
  - Tipo: párrafo.
  - Descripción: describa brevemente otras soluciones evaluadas (ej. contratación de más personal, desarrollo de software tradicional) y por qué fueron descartadas.

Sección 5 - Inclusión y diversidad

- Pregunta: ¿Cómo incorporaremos perspectivas diversas en el diseño y validación del sistema?
  - Tipo: párrafo.
  - Descripción: p. ej. mediante la conformación de equipos multidisciplinarios, realización de consultas con grupos de impacto, y asegurando que los datos de entrenamiento sean representativos de la diversidad de la población colombiana.

Sección 6 - Agencia humana y supervisión

- Pregunta: ¿En qué puntos del proceso se integrará la supervisión humana para la toma de decisiones?
  - Tipo: párrafo.
  - Descripción: p. ej. todas las recomendaciones de la IA serán revisadas y validadas por un funcionario público competente antes de tomar cualquier acción procesal. El sistema incluirá puntos de verificación obligatorios en decisiones críticas.
- Pregunta: ¿Podría el sistema afectar derechos humanos? Si es así, ¿se ha realizado una evaluación de impacto?
  - Tipo: párrafo.
  - Descripción: describa el análisis de impacto sobre derechos como la igualdad, el debido proceso, la presunción de inocencia y la privacidad.

Sección 7 - Privacidad y gobernanza de datos

- Pregunta: ¿Qué procedimientos de protección de datos se implementarán?
  - Tipo: párrafo.
  - Descripción: p. ej. seudonimización, encriptación de datos en reposo y en tránsito, controles de acceso basados en roles, auditorías de acceso regulares, políticas de retención y borrado seguro de datos.

Sección 8 - Transparencia, explicabilidad y robustez

- Pregunta: ¿Qué información sobre el sistema será pública para la ciudadanía?
  - Tipo: párrafo.
  - Descripción: describa qué se publicará, como la metodología general, las métricas de rendimiento auditadas, las limitaciones conocidas del sistema y el marco de gobernanza.

- Pregunta: ¿Qué políticas y procedimientos de seguridad técnica se seguirán?
  - Tipo: párrafo.
  - Descripción: p. ej. autenticación multifactor, *logs* de auditoría, monitoreo continuo de seguridad, pruebas de penetración periódicas.
- Pregunta: describa el plan de contingencia ante fallos técnicos.
  - Tipo: párrafo.
  - Descripción: mecanismos de detección temprana de errores, protocolos de respuesta a incidentes y procedimientos de recuperación.

Sección 9 - Diversidad, no discriminación y equidad

- Pregunta: ¿Qué estrategias se usarán para detectar y mitigar sesgos algorítmicos?
  - Tipo: párrafo.
  - Descripción: mencione posibles sesgos (p. ej. por género, origen étnico, estrato socioeconómico) y las medidas para mitigarlos (p. ej. diversificación de datos de entrenamiento, uso de métricas de equidad, revisión por comités diversos).
- Pregunta: ¿Cómo se garantizará que el sistema sea accesible e inclusivo?
  - Tipo: párrafo.
  - Descripción: considere la accesibilidad para personas con discapacidad, diversidad, no discriminación y equidad, las posibles barreras lingüísticas y el impacto de la brecha digital en el acceso al sistema o sus resultados.

Sección 10 - Bienestar social, ambiental y marco normativo

● Pregunta: ¿Cómo podría el sistema contribuir positiva o negativamente a la sociedad colombiana?
• Tipo: párrafo.
• Descripción: p. ej. mejora en el acceso a la justicia, reducción de tiempos procesales vs. riesgo de ampliar brechas digitales.
● Pregunta: ¿Qué marco normativo específico debe considerarse?
• Tipo: casillas de verificación.
• Opciones:
- Constitución Política de Colombia (arts. 15, 20, 29).
- Ley 1712 de 2014 (transparencia).
- Código de Procedimiento Penal.
- Decreto 1377 de 2013 (reglamentación habeas data).
- Ley 1581 de 2012 (protección de datos).
- Otro: (campo abierto).
● Pregunta: ¿Qué entidades estatales son relevantes para este proyecto?
• Tipo: párrafo corto.

Sección 11 - Responsabilidad, comunicación y ciclo de vida

● Pregunta: ¿Quién será responsable de cada etapa del ciclo de vida de la IA?
• Tipo: párrafo.
• Descripción: defina responsables para: (i) diseño y datos; (ii) verificación y validación; (iii) despliegue; (iv) operación y monitoreo, y (v) retiro/actualización.
● Pregunta: ¿Cómo se comunicará el funcionamiento del sistema al usuario final?
• Tipo: párrafo._____

Sección 12 - Aprobaciones

- Pregunta: nombre y cargo del líder del proyecto que aprueba este Canvas.
  - Tipo: respuesta corta.
- Pregunta: nombre y cargo del supervisor legal que aprueba este Canvas.
  - Tipo: respuesta corta.
- Pregunta: nombre y cargo del director/a que aprueba este canvas.
  - Tipo: respuesta corta.

Referencias

Arrieta, A. B., Díaz-Rodríguez, N., Del Ser, J., Bennetot, A., Tabik, S., Barbado, A., García, S., Gil-López, S., Molina, D., Benjamins, R., Chatila, R., y Herrera, F. (2020). Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI. *Information Fusion*, 58, 82-115. https://doi.org/10.1016/j.inffus.2019.12.012

Barocas, S., y Selbst, A. D. (2016). Big Data's Disparate Impact. *California Law Review*, 104(3), 671-732. https://doi.org/10.15779/Z38BG31

Brown, T. B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., Agarwal, S., Herbert-Voss, A., Krueger, G., Henighan, T., Child, R., Ramesh, A., Ziegler, D. M., Wu, J., Winter, C., ... Amodei, D. (2020). Language models are few-shot learners. En *Advances in Neural Information Processing Systems* (Article No.: 159, 1877–1901). Curran Associates, Inc. https://dl.acm.org/doi/abs/10.5555/3495724.3495883

Comité de los Derechos del Niño. (2021). Observación general núm. 25 (2021) relativa a los derechos de los niños en relación con el entorno digital (CRC/C/GC/25). Naciones Unidas. https://docs.un.org/es/crc/c/gc/25

Congreso de la República de Colombia. (2025). Proyecto de Ley 043-2025 del Senado de la República y 324-2025 de la Cámara de Representantes, por medio del cual se regula la inteligencia artificial en Colombia para garantizar su desarrollo ético, responsable, competitivo e innovador. [En trámite legislativo]. https://www.camara.gov.co/wp-content/uploads/2025/12/proyectos-ley/publicaciones/proyecto-35981/PPDSC-PL-043-25S-324-25C-INTELIGENCIA-ARTIFICIAL-SC.pdf

Consejo Nacional de Política Económica y Social [CONPES]. (2019). *Política Nacional para la Transformación Digital e Inteligencia Artificial* (Documento CONPES 3975). Departamento Nacional de Planeación [DNP]. https://colaboracion.dnp.gov.co/CDT/Conpes/Econ%C3%B3micos/3975.pdf

Consejo Nacional de Política Económica y Social [CONPES]. (2025). *Política Nacional de Inteligencia Artificial de Colombia* (Documento CONPES 4144). Departamento Nacional de Planeación [DNP]. https://colaboracion.dnp.gov.co/CDT/Conpes/Econ%C3%B3micos/4144.pdf

Cormen, T. H., Leiserson, C. E., Rivest, R. L., y Stein, C. (2009). *Introduction to Algorithms* (3rd ed.). MIT Press.

Departamento Administrativo de la Presidencia de la República [Dapre] (2021). *Marco Ético para la Inteligencia Artificial en Colombia*. Colombia. https://minciencias.gov.co/sites/default/files/marco-etico-ia-colombia-2021.pdf

Department of Public Expenditure, NDP Delivery and Reform. (2024). *Guidelines for the Responsible Use of Artificial Intelligence in the Public Service*. Government of Ireland. https://assets.gov.ie/static/documents/09fe3ad4/Guidelines_for_the_Responsible_Use_of_AI_in_the_Public_Service_20250918.pdf

Estevez, E., Fillottrani, P., y Linares Lejarraga, S. (2020). *PROMETEA: Transformando la administración de justicia con herramientas de inteligencia artificial*. https://doi.org/10.18235/0002378

Federal Chancellery FCh. (2025). Strategy: Use of AI systems in the Federal Administration. Swiss Confederation. https://www.bk.admin.ch/bk/en/home/digitale-transformation-ikt-lenkung/vorgaben/sb021-strategie-einsatz-von-ki-systemen-in-der-bundesverwaltung.html

Guedes, L., y Oliveira Júnior, M. (2024). Artificial intelligence adoption in public organizations: a case study. *Future Studies Research Journal: Trends and Strategies, 16*(1), e860. https://doi.org/10.24023/FutureJournal/2175-5825/2024.v16i1.860

High-Level Expert Group on Artificial Intelligence. (2019). *Ethics guidelines for trustworthy AI*. European Commission. https://digital-strategy.ec.europa.eu/en/library/ethics-guidelines-trustworthy-ai

LeCun, Y., Bengio, Y., y Hinton, G. (2015). Deep learning. *Nature*, 521(7553), 436-444. https://doi.org/10.1038/nature14539

Neudert, L.-M., Kollanyi, B., y Howard, P. N. (2017). *Junk News and Bots during the German Federal Presidency Election: What Were German Voters Sharing Over Twitter?* Project on Computational Propaganda. https://demtech.oii.ox.ac.uk/research/posts/junk-news-and-bots-during-the-german-federal-presidency-election-what-were-german-voters-sharing-over-twitter/

National Cyber Security Centre. (2023). *Cyber security guidance on generative AI for public sector bodies*. NCSC. https://www.ncsc.gov.ie/pdfs/Cybersecurity_Guidance_on_Generative_AI_for_PSBs.pdf

Organisation for Economic Co-operation and Development. (2019). Recommendation of the Council on Artificial Intelligence (OECD/LEGAL/0449). Modificado el 3/05/2024. OECD. https://legalinstruments.oecd.org/en/instruments/OECD-LEGAL-0449

Organisation for Economic Co-operation and Development. (2024). *Encuesta de la OCDE sobre los determinantes de la confianza en las instituciones públicas 2024: Notas por país - Colombia*. https://www.oecd.org/content/dam/oecd/es/publications/reports/2024/06/oecd-survey-on-drivers-of-trust-in-public-institutions-2024-results-country-notes_33192204/colombia_a71f1c24/e356c624-es.pdf

Parlamento Europeo y Consejo de la Unión Europea. (2024). Reglamento (UE) 2024/1689 por el que se establecen normas armonizadas en materia de inteligencia artificial (Reglamento de Inteligencia Artificial). 13 de junio de 2024. *Diario Oficial de la Unión Europea*, L 1689. https://eur-lex.europa.eu/eli/reg/2024/1689/oj/eng

Procuraduría General de la Nación [PGN]. (2025). Política de seguridad de la información de la PGN. Resolución n.° 138, Por medio de la cual se modifica y/o adicionan los artículos 3°, 4°, 6°, 7°, 8° y 9° de la Resolución 036 de 2009, se derogan el artículo 5° de la resolución 020 de 2021 y el artículo 3 de la resolución 296 de 2021, se asignan funciones y se dictan otras disposiciones. https://apps.procuraduria.gov.co/relatoria/media/file/flas_juridico/4555_RESOLUCION%20138%20DE%2027%20DE%20JUNIO%20DE%202025.pdf

Procuraduría General de la Nación. (2024a, 13 de septiembre). Política de Gobierno de Datos (Código CI-PO-01, Versión 01). https://apps.procuraduria.gov.co/portal/media/file/modulo_calidad/mapa_proceso//3741_CI-PO-01%20POLÍTICA%20GOBIERNO%20DE%20DATOS.pdf

Procuraduría General de la Nación. (2024b, 25 de octubre). Plan Estratégico de Tecnologías de la Información (PETI) 2022-2025: Actualización vigencia 2024 (código TI-PL-01, versión 1) https://www.procuraduria.gov.co/Documents/2024/Noviembre%202024/3758_TI-PL-01%20PLAN%20ESTRATEGICO%20DE%20TECNOLOGIAS%20DE%20LA%20INFORMACI%C3%93N.pdf

Procuraduría General de la Nación. (2024c, 25 de enero). Política de Tecnologías de la Información (código TI-PO-15, versión 1).

Russell, S. J., y Norvig, P. (2020). *Artificial Intelligence: A Modern Approach* (4th ed.). Pearson.

Tejedor, J. M., Medina, C. M., y Martínez, L. E. (2025a). *Diagnóstico para la adopción de la inteligencia artificial generativa en el Ministerio Público*. Procuraduría General de la Nación. [Documento interno de trabajo].

Tejedor, J. M., Medina, C. M., y Martínez, L. E. (2025b). *Portafolio de ideas para la adopción de inteligencia artificial en el Ministerio Público colombiano: Diagnóstico, casos de uso y propuestas de implementación.* Procuraduría General de la Nación. [Documento interno de trabajo].

Transparencia por Colombia. (2024). *Informe sobre desinformación y corrupción en Colombia*. https://transparenciacolombia.org.co/desinformacion-colombia-acceso-informacion-publica/

United Nations Educational, Scientific and Cultural Organization [UNESCO]. (2021). *Recommendation on the ethics of artificial intelligence*. UNESCO. https://unesdoc.unesco.org/ark:/48223/pf0000381137

Willats, R., Pennington, J., Mohan, A., y Vidgen, B. (2025). *Classification is a RAG problem: A case study on hate speech detection*. Contextual AI. https://arxiv.org/abs/2508.06204

World Justice Project. (2024). *Rule of Law Index 2024*.

Xing, Y., Zhai, C., Che, Z., Pan, H., Li, K., Zhang, B., Yao, Z., y Si, X. (2025). A Multimodal Fake News Detection Model Based on Bidirectional Semantic Enhancement and Adversarial Network Under Web3.0. *Electronics, 14*(18), 3652. https://doi.org/10.3390/electronics14183652

Normativas

Constitución Política de Colombia [C.P.]. 7 de julio de 1991.

Ley 12 de 1991. Por medio de la cual se aprueba la Convención sobre los Derechos del Niño adoptada por la Asamblea General de las Naciones Unidas el 20 de noviembre de 1989. 22 de enero de 1991. *Diario Oficial* n.° 39.640.

Ley 24 de 1992. Por la cual se establece la organización y funcionamiento de la Defensoría del Pueblo y se dictan otras disposiciones. 15 de diciembre de 1992. *Diario Oficial* n.° 40.690.

Ley 734 de 2002. Por la cual se expide el Código Disciplinario Único. 5 de febrero de 2002. *Diario Oficial* n.° 44.708.

Ley 906 de 2004. Por la cual se expide el Código de Procedimiento Penal. 31 de agosto de 2004. *Diario Oficial* n.° 45.658.

Ley 1098 de 2006. Por la cual se expide el Código de la Infancia y la Adolescencia. 8 de noviembre de 2006. *Diario Oficial* n.° 46.446.

Ley 1581 de 2012. Por la cual se dictan disposiciones generales para la protección de datos personales. 17 de octubre de 2012. *Diario Oficial* n.° 48.587.

Ley 1712 de 2014. Por medio de la cual se crea la Ley de Transparencia y del Derecho de Acceso a la Información Pública Nacional y se dictan otras disposiciones. 6 de marzo de 2014. *Diario Oficial* n.° 49.559.

Ley 1755 de 2015. Por medio de la cual se regula el derecho fundamental de petición. 30 de junio de 2015. *Diario Oficial* n.° 49.559.

Ley 1952 de 2019. Por medio de la cual se expide el Código General Disciplinario. 28 de enero de 2019. *Diario Oficial* n.° 50.850.

Ley 2502 de 2025. Por medio de la cual se modifica y establece un agravante al artículo 296 de la Ley 599 del 2000, Código Penal colombiano, referente al delito de falsedad personal para la modalidad de suplantación utilizando inteligencia artificial (IA). 28 de julio de 2025. *Diario Oficial* n.° 53.198.

Decreto Ley 262 de 2000. Por el cual se modifican la estructura y la organización de la Procuraduría General de la Nación. 22 de febrero de 2000. *Diario Oficial* n.° 43.904.

Decreto 1377 de 2013 [Presidencia de la República de Colombia]. Por el cual se reglamenta parcialmente la Ley 1581 de 2012. 27 de junio de 2013. *Diario Oficial* n.° 48.834.

Decreto 1263 de 2022. [Ministerio de Tecnologías de la Información y las Comunicaciones]. Por el cual se adiciona el Título 22 a la Parte 2 del Libro 2 del Decreto 1078 de 2015. 22 de julio de 2022. *Diario Oficial* n.° 52.103.

Directiva Conjunta n.° 007 de 2025. [Procuraduría General de la Nación y Defensoría del Pueblo]. (2025). Estándares sobre transparencia algorítmica para los sistemas algorítmicos utilizados por el Estado. 30 de septiembre de 2025.

Jurisprudencia

Corte Constitucional de Colombia. Sentencia T-323/24 de agosto de 2024. M.S. Juan Carlos Cortés. https://www.corteconstitucional.gov.co/relatoria/2025/t-067-25.htm

Corte Constitucional de Colombia. Sentencia T-067. 26 de febrero de 2025. M.P. Natalia Ángel Cabo. https://www.corteconstitucional.gov.co/relatoria/2025/t-067-25.htm

Organisation for Economic Co-operation and Development. (2019). *Recommendation of the Council on Artificial Intelligence* (OECD/LEGAL/0449). Modificado el 3/05/2024. OECD. https://legalinstruments.oecd.org/en/instruments/OECD-LEGAL-0449

PROCURADURÍA
GENERAL DE LA NACIÓN
COLOMBIA

DIÁLOGO PARA CONSTRUIR CONSENSOS

PROCURADURÍA EN LAS REGIONES

Paz Electoral