Gobierno de Colombia

Uso responsable de la IA



Guía Ética para la Implementación, Desarrollo y Uso de Sistemas de Inteligencia Artificial en Entidades Públicas de Colombia

Gobierno de
Colombia

Presidencia de la República

Saúl Kattan Cohen
Alto Consultor para la Transformación Digital

Ingrid Paola Hernández Sierra
Coordinadora Grupo Transformación Digital

Diana Marcela Arias
Asesora del Grupo Transformación Digital

Manuel Humberto Sierra
Asesor Grupo Transformación Digital

Juliana Flórez Durán
Asesora Grupo Transformación Digital

Agradecimientos

Viceministerio de Transformación Digital del Ministerio de Tecnologías de la Información y la Comunicaciones

Viceministerio de Conocimiento, Innovación y Productividad del Ministerio de Ciencia, Tecnología e Innovación

Dirección de Desarrollo Digital del Departamento Nacional de Planeación

Coordinación de Prospectiva Estratégica de la Comisión de Regulación y las Comunicaciones

Y a todas las entidades de gobierno que participaron con sus comentarios para enriquecer este documento

CONTENIDO

1. INTRODUCCIÓN....2

2. REFERENCIAS INTERNACIONALES PARA LA ÉTICA DE LA IA EN EL SECTOR PÚBLICO....4

2.1. Principios de la OCDE sobre la IA....4

2.2. Principios Rectores de las Naciones Unidas para la Gobernanza Internacional de la IA....5

2.3. Recomendación sobre la ética de la inteligencia artificial de la UNESCO....6

2.4. Valores Fundamentales y Principios de Acción del Alan Turing Institute....10

2.5. Aplicación en el sector público colombiano....11

3. PRINCIPIOS Y VALORES ÉTICOS FUNDAMENTALES PARA UNA IA RESPONSABLE EN COLOMBIA....17

3.1. Principios fundamentales....17

3.2. Valores fundamentales....21

3.3. Establecer un ecosistema de principios y valores fundamentales....24

4. MATERIALIZACIÓN DE LOS PRINCIPIOS Y VALORES ÉTICOS PARA UNA IA RESPONSABLE EN COLOMBIA....25

4.1. Habilitadores para una IA responsable....25

4.2. Salvaguardas y buenas prácticas para la implementación de IA....34

5. CONCLUSIONES Y RECOMENDACIONES CLAVE PARA COLOMBIA....41

GLOSARIO....43

SIGLAS....45

BIBLIOGRAFÍA....46

ANEXO TÉCNICO....47
Caso de Uso Ilustrativo: Sistema de Priorización de Cirugías con Inteligencia Artificial (SIPRICI) - Basado en la experiencia del Sistema Nacional de Salud del Reino Unido (NHS)



1. INTRODUCCIÓN

La **inteligencia artificial (IA)** se ha consolidado como una de las fuerzas más transformadoras del siglo XXI, redefiniendo las industrias, las economías y, de manera crucial, la administración pública a nivel global. Su integración con el gobierno digital promete una revolución en la prestación de servicios y en la eficiencia operativa de cualquier gobierno.

Como ya es conocido, los **sistemas de IA** tienen el potencial de automatizar procesos internos, trámites y servicios públicos, mejorar la toma de decisiones y la capacidad de previsión, optimizar la detección de fraudes y elevar la calidad del trabajo de los servidores públicos. Se estima que la IA podría automatizar un porcentaje significativo de transacciones repetitivas en el servicio público, liberando recursos humanos para tareas de mayor valor. (OCDE, 2025)

Sin embargo, la adopción de la IA en el sector público a menudo no ha alcanzado el ritmo observado en el sector privado. Esta disparidad se ha hecho más evidente en momentos de crisis, como la pandemia de COVID-19, que puso de manifiesto la importancia crítica de las tecnologías digitales y los datos para construir resiliencia económica y social y durante la cual, los gobiernos se vieron obligados a acelerar años de avances tecnológicos en cuestión de semanas o meses para mantener la operatividad y asegurar la provisión de servicios esenciales a ciudadanos y empresas.

En la actualidad, las administraciones públicas en todo el mundo, incluida la de Colombia, enfrentan una creciente presión para satisfacer las demandas ciudadanas, al mismo tiempo, que lidian con niveles decrecientes de confianza pública. En este escenario, la adopción estratégica de la IA se vuelve imperativa para aumentar la productividad, mejorar la capacidad de respuesta y fortalecer la rendición de cuentas. (OCDE, 2025)

Esta guía comprende un acercamiento a una visión ética integral, diseñada particularmente para las entidades públicas del orden

nacional en Colombia. Su propósito es orientar la implementación, el desarrollo y el uso de sistemas de IA, asegurando que estas tecnologías sirvan al bien común y se alineen con los valores democráticos del país y con los derechos humanos. Así mismo, se articulen con los marcos nacionales que se han venido desarrollando como el MSPI, la Política de Gobierno Digital y las obligaciones derivadas de la Ley 1581 de 2012 sobre protección de datos personales entre otros.

El documento se fundamenta en principios éticos globales y buenas prácticas internacionales, adaptándolos al contexto y las necesidades específicas colombianas, abordando los elementos y principios éticos fundamentales, los habilitadores clave para una implementación y uso exitosos y las salvaguardas esenciales para identificar, prevenir y mitigar posibles riesgos, al igual que un análisis de los desafíos y lecciones aprendidas, culminando en recomendaciones específicas para el país.

La ausencia de una guía clara y una estrategia proactiva podría hacer que esta visión se quede en fases piloto o enfrente resistencia debido a la falta de confianza. Es así, que, de forma previa a la elaboración de esta guía, se realizó un trabajo de diagnóstico sobre las necesidades de producción o actualización normativa o reglamentaria con el objetivo de asegurar el cumplimiento de los principios éticos presentados en el presente documento, con el apoyo de la UNESCO, a través de la aplicación de su metodología RAM (Readiness Assessment Methodology)[1] (del cual se tratará más adelante)

Este diagnóstico, derivado de la metodología RAM, identificó la necesidad de una producción normativa o reglamentaria urgente. Esto incluye la revisión de la Ley 1581 de 2012 y la creación de lineamientos para la clasificación y supervisión de riesgos en IA a nivel sectorial. Por lo tanto, esta guía funge como un paso indispensable hacia la reglamentación legal vinculante de la IA responsable en Colombia.

Por lo tanto, este documento no es meramente un manual de procedimientos, sino un llamado a la acción estratégica, fundamentado en la necesidad de evitar el rezago y asegurar que la IA se convierta en un motor de valor público en Colombia (OCDE, 2025).

[1] El informe final de la Metodología de Evaluación del Estado de Preparación (RAM) en Colombia para la implementación del marco ético de la inteligencia artificial en el país, puede ser consultado en https://unesdoc.unesco.org/ark:/48223/pf0000396015

2. REFERENCIAS INTERNACIONALES
PARA LA ÉTICA DE LA IA EN EL SECTOR PÚBLICO

La implementación de sistemas de inteligencia artificial en el sector público debe estar enmarcada en un sólido conjunto de principios éticos que guíen su diseño, desarrollo y uso. Estos principios, derivados de consensos internacionales y marcos especializados, buscan asegurar que la IA sirva al bienestar humano y al interés público, desde la identificación, prevención mitigación de los riesgos inherentes a su creciente inserción en la vida diaria. A continuación, se enumeran algunas referencias internacionales relacionadas con principios fundamentales y valores para un diseño, desarrollo y uso ético con participación e inclusión, de la inteligencia artificial.

2.1. PRINCIPIOS DE LA OCDE SOBRE LA IA

Adoptados por 47 países y la Unión Europea, constituyen un marco integral para promover el diseño, desarrollo, despliegue y uso confiable de la IA. (OCDE, 2025). Se estructuran en dos categorías principales: 1) principios basados en valores y 2) recomendaciones para los formuladores de políticas.

Entre los **principios basados en valores** se incluyen el fomento del crecimiento inclusivo, el desarrollo sostenible y el bienestar, destacando el potencial de la IA confiable para contribuir a la prosperidad y los objetivos de desarrollo global. Se enfatiza en el respeto al estado de derecho, los derechos humanos y los valores democráticos, incluyendo la equidad y la privacidad. Esto hace que se requieran salvaguardas adecuadas para una sociedad justa.

La transparencia y la explicabilidad de la IA[2] son cruciales para que las personas comprendan la interacción con los sistemas de IA y puedan impugnar sus resultados. La robustez, seguridad y protección garantizan que los sistemas funcionen de manera fiable a lo largo de su vida útil, con una gestión continua de riesgos. Finalmente, la rendición de cuentas responsabiliza a las organizaciones e individuos por el funcionamiento adecuado de los sistemas de IA, en línea con estos principios basados en valores (OCDE, 2025).

En cuanto a las **recomendaciones para los formuladores de políticas,** la OCDE sugiere invertir en investigación y desarrollo de IA para estimular la innovación confiable. Se promueve el fomento de un ecosistema habilitador de IA inclusivo, que abarque infraestructura digital, tecnologías y mecanismos para el intercambio de datos y conocimientos.

Por tanto, es fundamental dentro de estas recomendaciones, configurar un entorno de gobernanza y políticas interoperable que facilite el despliegue de sistemas de IA confiables. Así mismo, es necesario desarrollar la capacidad humana y preparar la transformación del

[2] La inteligencia artificial explicable es un conjunto de procesos y métodos que permite a los usuarios humanos comprender y confiar en los resultados y los productos creados por los algoritmos de machine learning (https://www.ibm.com/es-es/think/topics/explainable-ai).

mercado laboral, lo que implica que las personas cuenten con las habilidades necesarias para aprovechar la IA y que se apoyen a los trabajadores en una transición laboral justa en el caso que haya desplazamiento laboral originado por la implementación de la IA. Por último, la cooperación internacional es esencial para el intercambio de información, el desarrollo de estándares y la gestión responsable de la IA a nivel global (OCDE, 2025).

2.2. PRINCIPIOS RECTORES DE LAS NACIONES UNIDAS PARA LA GOBERNANZA INTERNACIONAL DE LA IA

Estos principios son propuestos por el Órgano Asesor de Alto Nivel sobre IA y buscan guiar la formación de nuevas instituciones de gobernanza internacional. Estos principios rectores clave establecen que la IA debe ser gobernada de manera inclusiva, en beneficio de todos y en interés público. Se subraya que la gobernanza de la IA debe construirse en consonancia con la gobernanza de datos y la promoción de bienes comunes de datos. Además, se indica que la gobernanza de la IA debe ser universal, en red y arraigada en la colaboración adaptativa de múltiples partes interesadas, teniendo como pilares fundamentales la Carta de las Naciones Unidas, el derecho internacional de los derechos humanos y otros compromisos internacionales acordados, como los Objetivos de Desarrollo Sostenible (ODS) (OCDE, 2025).

Las funciones institucionales propuestas por las Naciones Unidas incluyen la elaboración de normas, el cumplimiento y la rendición de cuentas, así como informes y revisiones por pares. Se promueve la colaboración internacional en datos, capacidad computacional y talento para abordar los ODS.

También se contempla la implementación de regímenes de responsabilidad, la mediación de estándares, la seguridad y los marcos de gestión de riesgos, la interoperabilidad y la alineación con las normas, al igual que la exploración de horizontes y la construcción de consenso científico (OCDE, 2025).

Adicionalmente, el Programa de las Naciones Unidas para el Desarrollo (PNUD), con el apoyo del Ministerio de Ciencia, Tecnología e Innovación, realizó en 2024 la Evaluación del Panorama de la Inteligencia Artificial en Colombia (AILA). Este reporte complementa los marcos éticos internacionales con evidencia empírica local sobre el nivel de preparación del país para desarrollar e implementar la IA de forma ética, inclusiva y responsable. Colombia obtuvo un nivel “diferenciador” (3.4/5), destacándose en responsabilidad e inclusividad, aunque se identifican brechas estructurales en capacidades técnicas, infraestructura de cómputo, equidad en el acceso y gobernanza interinstitucional. Estas evidencias respaldan la necesidad de contar con una guía ética nacional adaptada al contexto colombiano, como la que aquí se propone. (Programa de las Naciones Unidas para el Desarrollo (PNUD), 2024)

2.3. RECOMENDACIÓN SOBRE
LA ÉTICA DE LA INTELIGENCIA ARTIFICIAL DE LA UNESCO

La recomendación de la UNESCO (United Nations Educational, Scientific and Cultural Organization por sus siglas en inglés) es un instrumento normativo global que busca guiar el desarrollo y uso de la IA de manera responsable, centrada en la dignidad y los derechos humanos. Se propone un marco integral, global y multicultural que establece valores y principios interdependientes que deben ser respetados por todos los actores a lo largo del ciclo de vida de los sistemas de IA, desde la investigación y el diseño hasta el despliegue y la terminación. (UNESCO, 2023)

Su aplicación implica la modificación de leyes y regulaciones existentes, así como el desarrollo de nuevas normativas, siempre en consonancia con el derecho internacional y los Objetivos de Desarrollo Sostenible (ODS) de la ONU. Estas modificaciones en Colombia deberán realizarse en coordinación con entidades como MinCiencias, MINTIC, la Superintendencia de Industria y Comercio, la Ministerio de Educación Nacional, el Archivo General de la Nación, entre otras, garantizando consistencia con el derecho internacional y con los lineamientos del Sistema de Gobierno Digital y demás normatividad asociada.

Valores Fundamentales

La Recomendación de la UNESCO (UNESCO, 2023) articula cuatro valores fundamentales base de todas las acciones relacionadas con la IA:

- **Respeto, protección y promoción de los derechos humanos, las libertades fundamentales y la dignidad humana:** Se centra en que los sistemas de IA deben mejorar la calidad de vida humana sin objetivar a las personas ni socavar su dignidad o derechos. Los gobiernos y todos los actores de la IA deben respetar los instrumentos de derechos humanos en su interacción con los sistemas de IA, asegurando que las nuevas tecnologías promuevan, defiendan y ejerzan los derechos, en lugar de violarlos.

- **Prosperidad del medio ambiente y los ecosistemas:** Llama a que todos los actores de la IA respeten las leyes y normas ambientales internacionales y nacionales, reduciendo el impacto ambiental de los sistemas de IA, incluyendo su huella de carbono y consumo de energía, y previniendo la explotación insostenible de recursos naturales.

- **Garantizar la diversidad y la inclusión:** Este valor exige el respeto, la protección y la promoción de la diversidad y la inclusión a lo largo de todo el ciclo de vida de la IA. Esto implica fomentar la participación de todos los individuos y grupos, sin importar su origen, género, edad, idioma, religión o discapacidad. Se deben realizar esfuerzos para mitigar la falta de infraestructura, educación y habilidades tecnológicas en comunidades desfavorecidas, especialmente en países de ingresos bajos y medios, y evitar la explotación de estas situaciones.

- **Vivir en sociedades pacíficas, justas e interconectadas:** Insta a los actores de la IA a promover sociedades pacíficas y justas, basadas en un futuro interconectado que beneficie a todos y sea compatible con los derechos humanos. Este valor destaca el potencial de la IA para contribuir a la interconexión de todos los seres vivos entre sí y con el entorno natural, promoviendo la paz, la inclusión, la justicia, la equidad y la solidaridad, y evitando la segregación o la amenaza a la coexistencia.

Principios

Además de los valores antes citados, la (UNESCO, 2023) establece una serie de principios concretos para guiar la implementación y uso de la IA:

- **Proporcionalidad e inocuidad:** El alcance de los procesos de IA no debe exceder aquello que sea necesario para lograr sus objetivos legítimos y debe ser apropiados al contexto en que se desenvuelva. Si existe riesgo de daño a humanos, derechos humanos, comunidades o el medio ambiente, se deben implementar evaluaciones de riesgo y medidas preventivas. Las decisiones finales con impacto irreversible o de vida o muerte deben ser tomadas por un ser humano. La IA no debe usarse para calificación social o vigilancia masiva.

- **Seguridad y protección:** Se deben evitar y eliminar los daños no deseados (riesgos de seguridad) y las vulnerabilidades a ataques (riesgos de protección) a lo largo del ciclo de vida de la IA para garantizar la seguridad de humanos, el medio ambiente y los ecosistemas.

- **Equidad y no discriminación:** Los actores de la IA deben promover la justicia social, salvaguardar la equidad y combatir la discriminación, asegurando que los beneficios de la IA sean accesibles para todos, incluyendo grupos vulnerables. Se deben minimizar y evitar aplicaciones y resultados discriminatorios o sesgados, y garantizar recursos efectivos contra la discriminación.

- **Sostenibilidad:** La IA debe contribuir al desarrollo sostenible, evaluando continuamente sus efectos humanos, sociales, culturales, económicos y ambientales en relación con los ODS.

- **Derecho a la intimidad y protección de datos:** Los datos para sistemas de IA deben ser recolectados, usados, compartidos, archivados y eliminados de acuerdo con el derecho internacional y los principios de la Recomendación, con marcos de protección de datos y gobernanza robustos. Se debe aplicar el principio de privacidad desde el diseño.

- **Supervisión y decisión humana:** La responsabilidad ética y legal de las decisiones basadas en IA siempre debe ser atribuible a individuos o entidades legales. La supervisión humana es crucial, y las decisiones de vida o muerte no deben cederse a los sistemas de IA.

- **Transparencia y explicabilidad:** Son requisitos fundamentales para el respeto de los derechos humanos y los principios éticos. La transparencia permite la rendición

de cuentas y la capacidad de impugnar decisiones. La explicabilidad implica hacer compresibles para todos, los resultados y el funcionamiento de los sistemas de IA, de manera proporcional a su impacto y contexto.

- **Responsabilidad y rendición de cuentas:** Los actores de la IA deben respetar y promover los derechos humanos y la protección del medio ambiente, asumiendo su responsabilidad ética y legal. Se deben desarrollar mecanismos de supervisión, evaluación de impacto, auditoría y debida diligencia para asegurar la rendición de cuentas y la trazabilidad de los sistemas de IA.

- **Sensibilización y educación:** Se debe promover la conciencia y comprensión pública sobre la IA y el valor de los datos a través de educación accesible, participación cívica, la promoción de habilidades digitales y capacitación en ética de la IA, para que la sociedad en general pueda tomar decisiones informadas y estén protegidos de influencias indebidas.

- **Gobernanza y colaboración adaptativas y de múltiples partes interesadas:** La gobernanza de la IA debe ser inclusiva y adaptativa, a partir de la participación de diversos actores (gobiernos, sociedad civil, academia, sector privado). El uso de datos debe respetar el derecho internacional y la soberanía nacional. Se deben adoptar estándares abiertos y asegurar la interoperabilidad para facilitar la colaboración.

Ámbitos de Acción Política

La Recomendación de la (UNESCO, 2023) detalla once ámbitos de acción política que traducen estos valores y principios en medidas concretas. Estos ámbitos abarcan desde la evaluación del impacto ético y la gobernanza, hasta políticas de datos, desarrollo y cooperación internacional, medio ambiente, género, cultura, educación e investigación, comunicación e información, economía y trabajo, y salud y bienestar social. Cada ámbito propone acciones específicas para asegurar que la IA se desarrolle y utilice de manera ética y beneficiosa para la sociedad, abordando desafíos como la mitigación de riesgos, la promoción de la diversidad y la inclusión, y la garantía de la transparencia y la rendición de cuentas.

Para Colombia, la Recomendación de la UNESCO ofrece una hoja de ruta detallada para integrar la ética en la IA pública. Al alinear sus políticas con estos valores y principios, el país puede asegurar que sus iniciativas de IA no solo sean innovadoras, sino también socialmente responsables, equitativas y respetuosas de los derechos humanos, contribuyendo a un desarrollo sostenible e inclusivo.

Vale la pena destacar que, en el marco del compromiso de Colombia con el desarrollo ético de la inteligencia artificial, la UNESCO, con el apoyo del Ministerio de Ciencia, Tecnología e Innovación, lideró el proceso de implementación de los principios éticos de la IA en el país. Este esfuerzo se materializó a través de la aplicación de la metodología RAM (Readiness Assessment Methodology), a través de la cual se realizó la identificación de necesidades de producción o actualización normativa relacionada con IA, con el objetivo de asegurar el cumplimiento de los principios éticos sobre uso, desarrollo e implementación de sistemas basados en esta tecnología por parte de las entidades públicas.

El documento resultante fue construido a partir de un riguroso proceso que incluyó tanto la revisión de información pública emitida por diferentes entidades del gobierno colombiano, como entrevistas personales con representantes de sectores como: ciencia, tecnología, trabajo, educación, estadística y planeación, quienes compartieron sus avances, normativas existentes y capacidades institucionales en materia de IA. Esta evaluación permitió establecer el nivel de preparación del país para adoptar la IA de forma ética y derivó en una serie de recomendaciones orientadas a fortalecer la gobernanza nacional en este campo.

Algunas de las recomendaciones formuladas en este documento, se encuentran en el ámbito de la gobernanza e institucionalidad:

- Creación de espacios de diálogo sectorial y multiactor
- Fortalecimiento de capacidades de cooperación internacional
- Actualización de las políticas de datos abiertos
- Fortalecimiento de la infraestructura y gobernanza, e interoperabilidad de datos públicos
- Revisión de la Ley 1581 de 2012
- Articulación sector privado, público y academia (triple hélice)
- Creación de KPI para la apropiación ética de IA
- Promoción de sostenibilidad de la demanda de sistemas de IA
- Promoción de mecanismos exploratorios de regulación (sandbox)
- Fortalecimiento de la verificación y regulación de contenido sintético
- Promoción de mecanismos de compra pública de tecnología (sistemas de IA)
- Consolidación de la Estrategia Nacional de Seguridad Digital y creación de la Agencia de Seguridad Digital
- Creación de lineamientos para la clasificación y supervisión de riesgos en IA a nivel sectorial
- Y a nivel de creación de capacidades:
- Aumento de inversión en Investigación, Desarrollo e innovación (I+D+i) Fortalecimiento de la oferta pública y privada de capacitación y certificación en IA
- Fortalecimiento de las capacidades de las entidades territoriales
- Fortalecimiento de la capacitación de funcionarios y servidores públicos
- Creación de un repositorio centralizado de formación e investigación sobre IA
- Actualización de los currículos educativos para incluir capacidades técnicas y ética en IA
- Creación del Comité de Expertos de Alto Nivel de IA

Estas recomendaciones son de suma importancia para seguir avanzando en el establecimiento ético de la IA en el país y que permitirá asegurar un entorno mucho más favorable para el desarrollo de esta tecnología.

2.4. VALORES FUNDAMENTALES Y PRINCIPIOS DE ACCIÓN DEL ALAN TURING INSTITUTE

El Alan Turing Institute, a través de su programa de Ética y Gobernanza de la IA en la Práctica, ha desarrollado un conjunto de valores fundamentales y principios de acción que complementan los marcos globales, ofreciendo una orientación más operativa para la implementación de la IA en el sector público.

Valores SUM

Los **Valores SUM (Support, Underwrite, Motivate)** buscan establecer un vocabulario moral accesible para que los equipos puedan explorar y discutir las implicaciones éticas de sus proyectos de IA, proporcionando un marco para evaluar su permisibilidad ética (The Alan Turing Institute, 2019). Estos valores son:

- **Respetar:** Implica asegurar la capacidad de las personas para tomar decisiones libres e informadas sobre sus vidas, salvaguardar su autonomía y su derecho a ser escuchadas, y apoyar su desarrollo pleno y la búsqueda de sus pasiones.

- **Conectar:** Busca fomentar el diálogo interpersonal, la cohesión social, la diversidad, la participación y la inclusión. La IA debe utilizarse para fortalecer los lazos de solidaridad y confianza, promoviendo que todas las voces sean escuchadas y consideradas seriamente.

- **Cuidar:** Se centra en promover el bienestar de todas las partes interesadas, minimizando los riesgos de mal uso y abuso de la IA. Prioriza la seguridad y la integridad física y mental de las personas al concebir y desplegar aplicaciones de IA.

- **Proteger:** Aboga por priorizar los valores sociales, la justicia y el interés público. Esto incluye tratar a todos los individuos por igual, usar las tecnologías digitales para proteger el trato justo y equitativo bajo la ley, y emplear la IA para empoderar y avanzar los intereses de la mayor cantidad de individuos posible, considerando los impactos a largo plazo y en el medio ambiente.

Principios FAST

Los **Principios FAST Track (Fairness, Accountability, Sustainability, Transparency)** ofrecen una orientación práctica para el diseño y uso responsable de sistemas de IA. Estos principios buscan "llenar el vacío" entre la capacidad de las máquinas para realizar tareas que requieren inteligencia y su inherente falta de responsabilidad moral (The Alan Turing Institute, 2024).

- **Equidad (Fairness):** Los diseñadores e implementadores de sistemas de IA tienen la obligación de ser equitativos y de no causar daño a ninguna persona a través de sesgos o discriminación. Esto requiere una atención constante a la mitigación de sesgos en todas las fases del proyecto.

- **Rendición de Cuentas (Accountability):** Este principio exige que los seres humanos sean responsables de su papel en todas las etapas del flujo de trabajo de la IA, desde el diseño hasta el despliegue. Los resultados de este trabajo deben ser trazables de principio a fin, asegurando una cadena continua de responsabilidad humana.

- **Sostenibilidad (Sustainability):** Este principio se refiere a la producción de innovación en IA que sea segura y ética en sus resultados y en sus impactos más amplios. Busca garantizar que la IA contribuya al bienestar a largo plazo de las personas y el planeta.

- **Transparencia (Transparency):** Implica que los procesos de diseño e implementación de la IA deben ser justificables en su totalidad. Además, los resultados influenciados por algoritmos deben ser interpretables y comprensibles para las partes afectadas, permitiendo un escrutinio público.

El marco FAST puede utilizarse como instrumento complementario para evaluar equidad, responsabilidad y sostenibilidad en IA pública.

2.5. APLICACIÓN EN EL SECTOR PÚBLICO COLOMBIANO

A partir de la información anterior, es fundamental la interacción entre principios y valores generales para la adopción efectiva de la IA en Colombia. Los principios de alto nivel desarrollados por organizaciones como la OCDE, las Naciones Unidas y UNESCO establecen la visión estratégica y los valores fundamentales a nivel macro. Por otro lado, por ejemplo, los principios FAST Track y los valores SUM del Alan Turing Institute proporcionan una lente más granular y accionable, traduciendo esos valores abstractos en directrices concretas para la implementación diaria.

Esta sinergia es vital porque asegura que la ética de la IA no se quede en una mera declaración teórica, sino que se traduzca en procedimientos operativos estandarizados, programas de capacitación de personal y herramientas de evaluación que reflejen estos valores. La integración de estos diferentes niveles de principios garantiza una aproximación holística, desde la visión estratégica hasta la ejecución táctica. En entidades del sector público, esta aproximación holística implica garantizar que, en el uso de herramientas de analítica predictiva para la formulación de política pública, se mantenga la responsabilidad humana continua, se eviten sesgos algorítmicos y se fortalezcan las capacidades de capacitación, formación de los funcionarios y contratistas para mantener el control final sobre las decisiones asistidas por IA.

Esto se puede lograr, entre otros, mediante la implementación de Evaluaciones de Impacto Algorítmico (AIA) o evaluaciones de impacto en derechos fundamentales (HUDERIA) antes del despliegue, y a través de la conformación de equipos interdisciplinarios (con expertos en tecnología, ética, derecho y derechos humanos) para el diseño, desarrollo, y auditoría de los sistemas.

Para las entidades públicas colombianas, alinear su marco ético con los estándares internacionales más reconocidos no solo valida su enfoque, sino que también facilita la

cooperación y el intercambio de conocimientos con otras naciones. Esto promueve que la IA en Colombia sea no solo innovadora, sino también profundamente ética y socialmente responsable. En la Tabla 1 se presenta un resumen de principios y valores éticos globales de la IA y un ejercicio de cómo éstos pueden ser aplicados en el sector público.

Tabla 1. Principios y valores éticos globales de la IA y aplicación general para el sector público colombiano.

Fuente	Principio/Valor	Descripción Breve	Aplicación general para el Sector Público Colombiano
OCDE	Crecimiento Inclusivo, Desarrollo Sostenible y Bienestar	La IA debe beneficiar a todos los individuos, la sociedad y el planeta, contribuyendo a los ODS.	Los proyectos de IA deben diseñarse para mejorar los servicios públicos y la calidad de vida de todos los ciudadanos, priorizando la equidad y el impacto social positivo.
	Respeto al Estado de Derecho, Derechos Humanos y Valores Democráticos	La IA debe respetar los derechos humanos, la diversidad, la equidad y la privacidad, con salvaguardas adecuadas.	Garantizar que los sistemas de IA no infrinjan derechos fundamentales, protejan la privacidad de los datos personales y promuevan la no discriminación.
	Transparencia y Explicabilidad	Las personas deben entender cuándo interactúan con la IA y poder impugnar sus resultados.	Implementar mecanismos para explicar cómo funcionan las decisiones de IA, especialmente en servicios críticos, y permitir la revisión y apelación.
	Robustez, Seguridad y Protección	Los sistemas de IA deben funcionar de manera fiable y segura a lo largo de su vida útil, gestionando riesgos.	Asegurar la fiabilidad, ciberseguridad y resiliencia de los sistemas de IA, con monitoreo continuo de riesgos.
	Rendición de Cuentas	Organizaciones e individuos son responsables del funcionamiento adecuado de la IA.	Establecer líneas claras de responsabilidad humana para el diseño, desarrollo, implementación y uso de sistemas de IA.

Fuente	Principio/Valor	Descripción Breve	Aplicación general para el Sector Público Colombiano
Naciones Unidas	Gobernanza Inclusiva y en Beneficio de Todos	La IA debe ser gobernada por y para el beneficio de toda la humanidad, incluyendo las poblaciones vulnerables o de trato especial. (Niños, niñas y adolescentes)	Fomentar la participación de diversos grupos de interés, incluyendo comunidades vulnerables, en la definición de políticas de IA.
	Interés Público	La IA debe ser gobernada en interés público.	Desarrollar marcos robustos de gobernanza de datos que faciliten el acceso, uso y compartición segura de datos para la IA, protegiendo la privacidad.
	Gobernanza de Datos y Bienes Comunes de Datos	La gobernanza de la IA debe estar ligada a la gobernanza de datos y la promoción de datos como bien común.	Desarrollar marcos robustos de gobernanza de datos que faciliten el acceso, uso y compartición segura de datos para la IA, protegiendo la privacidad.
	Universalidad, Red y Colaboración Multi-Actor	La gobernanza de la IA debe ser global, conectada y basada en la colaboración de múltiples partes interesadas.	Promover la colaboración entre el gobierno, la academia, el sector privado y la sociedad civil en el desarrollo e implementación de la IA.
	Anclaje en la Carta de la ONU y Derechos Humanos	La gobernanza de la IA debe basarse en el derecho internacional de los derechos humanos y los ODS.	Asegurar que todas las políticas y usos de la IA respeten y promuevan los derechos humanos y contribuyan a los ODS.

Fuente	Principio/Valor	Descripción Breve	Aplicación general para el Sector Público Colombiano
UNESCO (Valores)	Respeto, Protección y Promoción de los Derechos Humanos, Libertades Fundamentales y Dignidad Humana	La IA debe mejorar la calidad de vida humana sin objetivar a las personas ni violar sus derechos.	Asegurar que todo sistema de IA implementado por entidades públicas respete la Constitución colombiana y los tratados internacionales de derechos humanos, con garantías de protección.
	Prosperidad del Medio Ambiente y los Ecosistemas	La IA debe contribuir a la protección y restauración del medio ambiente, reduciendo su impacto ecológico.	Priorizar el desarrollo de IA con baja huella de carbono y consumo responsable de los recursos energéticos e hídricos, y utilizarla para monitorear y mitigar problemas ambientales en Colombia (ej. deforestación, gestión de recursos hídricos).
	Garantizar la Diversidad y la Inclusión	La IA debe promover la participación de todos los grupos, sin discriminación, y mitigar las brechas digitales.	Diseñar sistemas de IA que sean accesibles para todas las poblaciones colombianas, incluyendo comunidades étnicas y rurales, y asegurar que la apropiación de conocimiento y datos de entrenamiento reflejen la diversidad cultural y lingüística del país.
	Vivir en Sociedades Pacíficas, Justas e Interconectadas	La IA debe fomentar la cohesión social, la justicia y la solidaridad, evitando la segregación o la confrontación.	Utilizar la IA para fortalecer la participación ciudadana, mejorar la transparencia en la administración pública y de justicia. Promover la resolución pacífica de conflictos, contribuyendo a la construcción de paz.
	Proporcionalidad e Inocuidad	Los sistemas de IA no deben ir más allá de lo necesario para un fin legítimo y no deben causar daño. Decisiones de vida o muerte siempre deben ser humanas.	Implementar evaluaciones de riesgo rigurosas para cada proyecto de IA, prohibiendo usos como la calificación social o la vigilancia masiva que atenten contra la dignidad humana u otro valor esencial para la sociedad
UNESCO (Principios)	Seguridad y Protección	Prevenir daños no deseados y vulnerabilidades a ataques a lo largo del ciclo de vida de la IA.	Establecer protocolos de ciberseguridad robustos y marcos de acceso a datos que garanticen la protección de la información sensible, almacenada y manejada por sistemas de IA en el gobierno.
	Equidad y No Discriminación	Combatir la discriminación y asegurar que los beneficios de la IA sean accesibles para todos los grupos.	Desarrollar políticas que aseguren el acceso equitativo a los servicios públicos y del estado impulsados por IA y que los algoritmos sean auditados para detectar y corregir sesgos que puedan afectar a poblaciones vulnerables.
	Sostenibilidad	Evaluar el impacto de la IA en las dimensiones humanas, sociales, culturales, económicas y ambientales para contribuir a los ODS.	Integrar la evaluación de impacto en los ODS en todos los proyectos de IA del sector público, buscando soluciones que promuevan el desarrollo sostenible y la resiliencia.

Fuente	Principio/Valor	Descripción Breve	Aplicación general para el Sector Público Colombiano
UNESCO (Principios)	Derecho a la Intimidad y Protección de Datos	Respetar y promover la privacidad y la protección de datos personales a lo largo del ciclo de vida de la IA.	Fortalecer los marcos legales y operativos para la protección de datos personales, exigiendo evaluaciones de impacto en la privacidad y el diseño de sistemas con privacidad por defecto.
	Supervisión y Decisión Humanas	La responsabilidad ética y legal de la IA debe ser siempre atribuible a humanos; la IA no reemplaza la decisión final humana.	Asegurar que los servidores públicos mantengan la supervisión y el control final sobre las decisiones asistidas por IA, especialmente en áreas críticas como la justicia o la asignación de beneficios.
	Transparencia y Explicabilidad	Los sistemas de IA deben ser comprensibles y sus decisiones explicables, de manera proporcional al impacto.	Implementar mecanismos que permitan a los ciudadanos entender cómo funcionan los sistemas de IA que los afectan y por qué se toman ciertas decisiones, con posibilidad de solicitar explicaciones.
	Responsabilidad y Rendición de Cuentas	Los actores de la IA deben asumir su responsabilidad ética y legal, con mecanismos de supervisión y auditoría.	Establecer un marco de rendición de cuentas claro para cada etapa del ciclo de vida de la IA, con auditorías regulares y protección para denunciantes de irregularidades.
	Sensibilización y Educación	Promover la comprensión pública de la IA y la ética a través de la educación y la participación cívica.	Desarrollar programas de capacitación para servidores públicos y ciudadanos sobre la ética de la IA, sus oportunidades y riesgos, fomentando la alfabetización digital y el pensamiento crítico.
	Gobernanza y Colaboración Adaptativas y de Múltiples Partes Interesadas	La gobernanza de la IA debe ser inclusiva, adaptativa y fomentar la colaboración entre diversos actores.	Fomentar la creación de espacios de diálogo y colaboración entre el gobierno, la academia, el sector privado y la sociedad civil para co-crear políticas y soluciones de IA.
Alan Turing Institute (Valores SUM)	Respetar	Dignidad individual, autonomía y capacidad de decisión informada.	Diseñar sistemas de IA que empoderen a los ciudadanos y respeten su libertad de elección.
	Conectar	Diálogo interpersonal, cohesión social, diversidad, participación e inclusión.	Usar la IA para mejorar la comunicación entre el gobierno y los ciudadanos, fomentando la participación y la diversidad de opiniones.

Fuente	Principio/Valor	Descripción Breve	Aplicación general para el Sector Público Colombiano
Alan Turing Institute (Valores SUM)	Cuidar	Bienestar de las partes interesadas, minimización de riesgos, seguridad e integridad.	Implementar la IA con un enfoque en la prevención de daños y la protección de la salud y el bienestar de las personas. Con énfasis en niños, niñas y adolescentes.
	Proteger	Valores sociales, justicia, interés público, equidad y empoderamiento.	Asegurar que la IA contribuya a la justicia social, la equidad y el empoderamiento de los más vulnerables.
Alan Turing Institute (Principios FAST Track)	Rendición de Cuentas (Accountability)	Responsabilidad humana continua y trazabilidad de resultados en todo el ciclo de vida de la IA.	Establecer quién es responsable en cada etapa del proyecto de IA y documentar las decisiones para permitir auditorías.
	Transparencia (Transparency)	Procesos justificables y resultados interpretables y comprensibles para las partes afectadas.	Hacer públicos los procesos de diseño e implementación de la IA y explicar sus resultados de manera clara y accesible.
	Equidad (Fairness)	Mitigación de sesgos y no discriminación en el diseño y los resultados de la IA.	Implementar medidas activas para identificar y corregir sesgos en los datos y algoritmos, evitando resultados discriminatorios.
	Sostenibilidad (Sustainability)	Innovación en IA segura, ética y con impactos positivos a largo plazo.	Evaluar los impactos a largo plazo de la IA en la sociedad y el medio ambiente, buscando soluciones que generen valor duradero.

Tabla 1. Principios y valores éticos globales de la IA y aplicación general para el sector público colombiano.

3. PRINCIPIOS Y VALORES ÉTICOS
FUNDAMENTALES PARA UNA **IA** RESPONSABLE EN COLOMBIA

Para una implementación, desarrollo y uso de sistemas de Inteligencia Artificial en Entidades Públicas del país que asegure que la IA sirva al bienestar de las personas y al interés público, es necesario establecer un conjunto de principios y valores que sirvan como marco orientador que garantice que los beneficios de la IA lleguen, de manera equitativa, inclusiva y ética en todo el territorio nacional. Cabe anotar que la intención de esta propuesta no es limitar la adopción de valores y principios relacionados con la implementación, desarrollo y uso de sistemas de Inteligencia Artificial, lo que se busca es generar unos "mínimos" sobre los cuales las entidades públicas puedan hacer la construcción de sus propios principios y valores, adecuados a su misionalidad y objetivo. Así que exhortamos a todas las entidades públicas que, con base en esta guía, expandan y enriquezcan sus propios principios y valores base con la intención y visión del servicio público.



En ese sentido, se presenta a continuación una propuesta de principios y valores que se consideran fundamentales para alcanzar una IA ética y responsable, a partir de una síntesis del conjunto de principios y valores expuestos en la **Tabla 1,** adaptándolos al contexto colombiano:

3.1. PRINCIPIOS FUNDAMENTALES

1. Centralidad Humana y Bien Público

Los sistemas de IA en el sector público deben ser diseñados, desarrollados y desplegados con el objetivo primordial de promover el bienestar de las personas, proteger los derechos humanos y fortalecer el interés general. Su propósito fundamental es aumentar las capacidades humanas y no dejar la toma de decisiones críticas exclusivamente en manos de la IA. Este enfoque garantiza que el despliegue de las tecnologías en el estado colombiano se realice bajo principios de soberanía digital, equidad, inclusión y rendición de cuentas, asegurando que la innovación tecnológica sea una herramienta al servicio de las personas, mejorando su calidad de vida, fortaleciendo los valores democráticos.

La aplicación de este principio conlleva la priorización de las necesidades ciudadanas, la garantía de un acceso equitativo a los servicios y el fomento de sociedades inclusivas. La finalidad última es que la IA impulse un gobierno más eficiente, justo y sensible a las necesidades de su población.

2. Transparencia, Explicabilidad y Rendición de Cuentas

Los sistemas de IA en el sector público deben operar bajo principios de transparencia, garantizando que su propósito, las fuentes de datos utilizadas, los mecanismos de procesamiento y toma de decisiones sean comprensibles para todos los actores involucrados. Sus resultados deben ser explicables, es decir, debe permitir a los usuarios humanos y a las personas afectadas comprender el razonamiento detrás de las decisiones impulsadas por la IA, incorporando, por ejemplo, registros técnicos obligatorios como control de versiones del modelo, bitácoras de entrenamiento, metadatos, fuentes de datos y logs de auditoría. Además, deben establecerse líneas claras de rendición de cuentas para cada paso del proceso desde el diseño, despliegue, desarrollo, implementación hasta los resultados de la IA.

Este principio es crucial para fomentar la confianza pública, permitir la supervisión democrática y facilitar mecanismos de reparación, de apelación y revisión para la identificación de oportunidades de mejora. Adicionalmente, se debe asegurar la auditabilidad (Auditability) y la responsabilidad continua (Answerability) a lo largo del ciclo de vida de la IA, con el fin de fortalecer la justicia ciudadana y la supervisión democrática de los sistemas implementados para su permanente actualización. La transparencia y la explicabilidad son vitales para construir la confianza pública, mientras que los mecanismos de rendición de cuentas deben garantizar la supervisión humana y la responsabilidad por los resultados impulsados por la IA.

3. Equidad, Igualdad y No Discriminación

Los sistemas de IA en el sector público deben ser diseñados e implementados para garantizar un trato justo y equitativo para todos los ciudadanos, trabajando activamente para prevenir y mitigar el sesgo algorítmico que podría conducir a la discriminación o exacerbar las desigualdades sociales existentes.

La aplicación de este principio implica asegurar la igualdad de acceso a los servicios públicos, prevenir impactos desproporcionados en grupos vulnerables y defender los principios de justicia. Esto requiere medidas proactivas para identificar y mitigar el sesgo en los datos, y los algoritmos en el despliegue de sistemas de IA.

4. Privacidad, Gobernanza de Datos y Seguridad

Los sistemas de IA en el sector público deben garantizar el respeto y la protección efectiva de la privacidad de los ciudadanos en consonancia con los derechos fundamentales y principios de soberanía digital. Esto exige marcos de gobernanza de datos robustos que regulen de manera rigurosa la recolección, almacenamiento, procesamiento y uso de datos conforme a estrictos estándares éticos y legales, incluyendo la minimización de

[3] Se refiere a la recopilación y procesamiento de la cantidad mínima de datos personales necesarios para un propósito específico, y solo durante el tiempo estrictamente necesario.

datos[3], la anonimización y su uso seguro y responsable. La implementación de medidas de ciberseguridad debe ser prioritaria con el fin de prevenir accesos no autorizados, vulneraciones de datos y riesgos asociados a la integridad de los sistemas. Así mismo, se requiere establecer mecanismos claros, accesibles y verificables para la obtención del consentimiento informado por parte de los titulares de los datos, garantizando que las personas comprendan cómo, por qué y para qué se utiliza su información en entornos de Inteligencia Artificial Pública.

La observancia de este principio es vital para mantener la confianza ciudadana y prevenir el uso indebido de información sensible con especial atención en los casos de niños, niñas y adolescentes.

5. Robustez, Fiabilidad y Seguridad

Los sistemas de IA implementados en el sector público deben cumplir con altos estándares de robustez técnica, fiabilidad operativa y seguridad en su funcionamiento. Esto significa que deben funcionar de manera consistente y precisa en diferentes situaciones y entornos diversos, resistir errores y posibles ataques y minimizar el riesgo de producir resultados inesperados o dañinos.

La aplicación rigurosa de este principio fortalece la integridad de los servicios ofrecidos al público, previene interrupciones que pueden afectar derechos fundamentales y contribuye a la sostenibilidad y eficiencia de las operaciones institucionales. Para ello se requiere un enfoque integral que incluya pruebas exhaustivas, validación técnica continua y mecanismos de monitoreo permanente, orientados a anticipar fallos, mitigar vulnerabilidades y garantizar la seguridad del sistema en todo su ciclo de vida.

6. Sostenibilidad Ambiental y Bienestar Social

El diseño, desarrollo y despliegue de la IA en el sector público deben incorporar una evaluación integral de su huella ambiental (incluyendo consumo energético, hídrico y de otros recursos) así como sus impactos sociales a largo plazo como la transformación del mercado laboral, el riesgo de exclusión digital.

Es imperativo que estos procesos se orienten a minimizar las consecuencias negativas y potenciar los beneficios sociales, económicos y ambientales en coherencia con los Objetivos de Desarrollo Sostenible (ODS).

Este principio implica la promoción activa de prácticas de IA ecológicas y ambientalmente sostenibles, que reduzcan su impacto ecológico a lo largo de todo el ciclo de vida tecnológico. Así mismo, exige preparar a la fuerza laboral para los cambios estructurales que la IA genera en los entornos productivos, mediante estrategias de capacitación, reconversión y fortalecimiento de habilidades digitales para lograr un acceso equitativo a la tecnología.

Para concluir, en la **Tabla 2,** se resumen los principios, sus implicaciones y riesgos para el sector público colombiano.

Tabla 2 Principios fundamentales para la IA en el sector público colombiano.

Principio	Definición Central (Contexto del Sector Público)	Implicaciones Clave para los Ciudadanos/Bien Público	Riesgos Asociados si se Viola
1. Centralidad Humana y Bien Público	La IA debe servir al bienestar humano, defender los derechos y aumentar las capacidades, no reemplazarlas en decisiones críticas.	Prioriza las necesidades ciudadanas, asegura acceso equitativo, fomenta sociedades inclusivas, salvaguarda derechos fundamentales.	Deshumanización de servicios, erosión de derechos, desatención de grupos vulnerables, pérdida de legitimidad democrática.
2. Transparencia, Explicabilidad y Rendición de Cuentas	Los sistemas de IA deben ser comprensibles en su propósito, datos y decisiones; las responsabilidades deben ser claras.	Fomenta la confianza pública, permite la supervisión, facilita mecanismos de reparación, asegura control democrático.	Desconfianza, arbitrariedad, imposibilidad de impugnación, falta de control democrático.
3. Equidad, Igualdad y No Discriminación	La IA debe garantizar un trato justo para todos, previniendo y mitigando activamente el sesgo algorítmico.	Asegura igualdad de acceso a servicios, previene impacto dispar en grupos vulnerables, defiende principios de justicia.	Discriminación, perpetuación de desigualdades, injusticia social, pérdida de confianza en la imparcialidad del gobierno.
4. Privacidad, Gobernanza de Datos y Seguridad	La IA debe proteger la privacidad ciudadana con marcos robustos de datos, minimización, anonimización y ciberseguridad.	Mantiene la confianza ciudadana, cumple regulaciones, previene uso indebido de información sensible, salvaguarda seguridad nacional.	Violaciones de privacidad, uso indebido de datos, robo de identidad, vulnerabilidades de seguridad que afectan a la Nación.
5. Robustez, Fiabilidad y Seguridad	Los sistemas de IA deben ser técnicamente sólidos, fiables y seguros, funcionando consistentemente y minimizando riesgos.	Asegura la integridad de los servicios públicos, previene fallos sistémicos, mantiene eficiencia operativa, construye confianza en soluciones de IA.	Fallos del sistema, decisiones erróneas, interrupción en servicios públicos y del estado, pérdida de confianza en la eficacia del gobierno.
6. Sostenibilidad Ambiental y Bienestar Social	El desarrollo de la IA debe considerar su huella ambiental y sus impactos sociales a largo plazo, buscando minimizar externalidades negativas.	Promueve prácticas de IA "verdes", prepara la fuerza laboral para cambios, asegura acceso equitativo a la tecnología, fomenta una sociedad resiliente.	Mayor consumo energético, mayor consumo de fuentes hídricas, obsolescencia laboral, ampliación de la brecha digital, impactos sociales no anticipados. (La salud mental, desarrollo cognitivo, acceso a la información, acceso a la educación, entre otros)

3.2. VALORES FUNDAMENTALES

La formulación de principios éticos, por muy bien intencionada y detallada que sea, no garantiza por si sola una implementación ética de la IA. Persiste una brecha crítica entre las aspiraciones éticas de alto nivel y las complejidades operativas que caracterizan el despliegue de sistemas de IA en entornos complejos del sector público.

Materializar los valores éticos en el desarrollo y uso de la IA exige mucho más que declaraciones de intención: requiere la construcción de un ecosistema integral compuesto por herramientas técnicas, procesos institucionales y capacidades humanas articuladas. Sin una operación robusta y coherente, incluso los principios más bien formulados pueden ser vacíos en sus contenidos. La ausencia de mecanismos efectivos de implementación conduce directamente a fallos éticos, vulneración de derechos y a una erosión progresiva de la confianza pública.

3.2.1. Marcos institucionales y Mecanismos de Supervisión

Para que los principios éticos sean efectivos, requieren marcos institucionales sólidos, políticas claras y una supervisión regulatoria que permita su implementación y cumplimiento.

- **Directrices Éticas y Códigos de Conducta:** Es fundamental desarrollar directrices específicas orientadas a los funcionarios públicos que participan en el diseño, implementación y supervisión de proyectos de IA. Estos códigos deben ir más allá de los comportamientos generales, traduciendo los valores éticos en declaraciones exactas de comportamiento, toma de decisiones y responsabilidad institucional en el día a día. El propósito de estas directrices es ofrecer una guía práctica que facilite la aplicación de los principios éticos en las actividades cotidianas, promoviendo una cultura de integridad, transparencia y rendición de cuentas en el uso de las tecnologías emergentes dentro del sector público.

- **Comités de Ética de la IA y Juntas de Revisión:** Crear instancias internas dedicadas a la revisión ética de los proyectos de IA, especialmente aquellos de alto riesgo, constituye un componente esencial de una gobernanza responsable. Estas instancias pueden anticipar y evaluar las implicaciones éticas antes del despliegue, incorporando una capa de escrutinio técnico, jurídico y social que fortalece la legitimidad institucional y la alineación con los valores públicos.

- **Evaluaciones de Impacto (Ético, de Privacidad, de Derechos Humanos):** Diseñar e implementar evaluaciones exhaustivas antes del despliegue de sistemas de IA es vital para adelantar una gestión proactiva del riesgo. Estas evaluaciones deben considerar no solo los impactos técnicos, sino también las dimensiones sociales, económicas y jurídicas, con especial énfasis en la protección de los derechos fundamentales de los ciudadanos. Para sistemas de IA de alto riesgo o con implicaciones en derechos fundamentales, la Evaluación de Impacto Algorítmico (AIA) debe ser un requisito obligatorio y trazable, asegurando la rendición de cuentas anticipatoria (ex-ante).

- **Sandboxes Regulatorios y Programas Piloto:** Crear entornos controlados para el desarrollo y validación de soluciones innovadoras de IA, bajo orientación ética clara, constituye una herramienta clara para fomentar la innovación responsable. Estos entor-

nos permiten la experimentación en condiciones seguras facilitando el aprendizaje en un marco ético seguro. Al operar dentro de un marco de gobernanza que prioriza la protección de derechos y la transparencia, se facilita la identificación temprana de desafíos éticos y la iteración de soluciones antes de su despliegue a gran escala reduciendo riesgos y fortaleciendo la confianza institucional en el uso de las tecnologías emergentes.

3.2.2. Participación de las Partes Interesadas y Participación Pública

Involucrar activamente a los ciudadanos, las organizaciones de la sociedad civil, academia y expertos en cada etapa del ciclo de vida de los sistemas de IA es fundamental para garantizar una gobernanza inclusiva y legitima. Esta participación permite incorporar diversas perspectivas, identificar necesidades reales y construir confianza en el marco de los valores democráticos y derechos fundamentales.

- **Consultas Públicas y Foros Deliberativos y Cocreación:** La implementación de mecanismos de participación ciudadana en el diseño y evaluación de las estrategias de IA y aplicaciones específicas, a través de consultas públicas y foros deliberativos, es fundamental para garantizar la pertinencia social en los proyectos de IA respondiendo a necesidades sociales y siendo aceptadas por la sociedad.

- **Estrategias de Comunicación Claras:** Explicar las iniciativas de IA al público de manera clara, accesible y transparente, es esencial para construir y mantener la confianza ciudadana. Este esfuerzo debe ir más allá de la divulgación técnica, promoviendo una comprensión genuina de los objetivos, beneficios, riesgos y límites de la IA por parte de todos los sectores de la sociedad. Al adoptar lenguajes inclusivos de alfabetización digital, las instituciones pueden fortalecer la apropiación social de la tecnología.

3.2.3. Fortalecimiento de Capacidades y Alfabetización Ética para Servidores Públicos

- **Programas de Alfabetización en IA:** Ofrecer una comprensión básica de la IA a todos los servidores públicos es fundamental para desmitificar la tecnología, reducir barreras cognitivas y fomentar una cultura institucional de adaptación e innovación.

- **Capacitación Especializada en Ética de la IA:** Para quienes participen en el desarrollo, adquisición y despliegue de soluciones basadas en IA, la capacitación especializada en ética resulta indispensable. Esta formación no solo les permite identificar y abordar dilemas éticos complejo, sino que también fortalece su capacidad para incorporar principios como la equidad, la transparencia, la rendición de cuentas y el respeto por los derechos humanos en cada etapa del ciclo tecnológico.

- **Equipos Interdisciplinarios:** Impulsar la colaboración entre equipos interdisciplinarios que integren expertos técnicos, los expertos en ética, los expertos juristas y otros profesionales, es vital para abordar la complejidad de la IA desde múltiples perspectivas.

3.2.4. Monitoreo Continuo, Evaluación de Impacto y Mejora Iterativa

Dado que los sistemas de IA evolucionan, y sus impactos pueden cambiar con el tiempo, es indispensable establecer mecanismos de monitoreo y la evaluación continuos que aseguren una alineación sostenida con los principios éticos. El sector público debe adoptar un enfoque ágil y adaptativo, basado en evidencia que permita revisar, aprender y perfeccionar las políticas relacionadas con la IA, adoptando una gestión ética continua.

- **Monitoreo del Rendimiento:** El seguimiento del rendimiento, la precisión y el sesgo de los sistemas de IA a lo largo del tiempo es fundamental para detectar desviaciones, riesgos emergentes y efectos no previstos. Este monitoreo continuo permite anticipar impactos, fortalecer la confiabilidad de los sistemas y garantizar su alineación con estándares éticos, normativos y de calidad técnica.

- **Evaluaciones de Impacto Post-Despliegue:** La realización periódica de evaluaciones sobre los impactos sociales y en los derechos humanos de los sistemas de IA desplegados es esencial para comprender de manera profunda y contextual sus consecuencias reales. Este ejercicio permite identificar efectos no previstos, mitigar riesgos emergentes y fortalecer la rendición de cuentas institucional. Integrar estas evaluaciones en los ciclos de vida de los sistemas contribuye a una gobernanza ética, transparente y centrada en las personas.

- **Mecanismos de Retroalimentación y Reparación:** Es fundamental establecer canales claros para que los ciudadanos proporcionen retroalimentación, planteen preocupaciones y soliciten reparación frente a decisiones impulsadas por la IA. Estos mecanismos fortalecen la justicia y promueven la rendición de cuentas institucional y garantizan el respeto a los derechos fundamentales.

- **Gobernanza Adaptativa:** En un entorno marcado por la rápida evolución de la IA y la aparición constante de nuevos desafíos técnicos, éticos y sociales, resulta imprescindible que las instituciones públicas adopten una gobernanza flexible y proactiva. Esto implica estar preparados para revisar y actualizar de manera oportuna las políticas, normativas y marcos regulatorios, garantizando su pertinencia y eficacia.

3.2.5. El Rol Indispensable de la Experiencia y el Juicio Humano

Aunque la IA automatiza tareas y optimiza procesos, el juicio humano sigue siendo insustituible, especialmente en el sector público, donde las decisiones a menudo implican dilemas éticos complejos, interpretaciones normativas matizadas y circunstancias individuales que los algoritmos no pueden captar plenamente. Se debe preservar la responsabilidad última de los seres humanos, así como su capacidad de juicio discrecional. Esto implica la necesidad de estrategias claras que involucren de manera responsable a los humanos en las etapas del ciclo de la IA, al igual que establecer mecanismos robustos para la revisión y anulación humana de las decisiones automatizadas de la IA, particularmente en aplicaciones de alto riesgo o con implicaciones sensibles para los derechos fundamentales.



3.3. ESTABLECER UN ECOSISTEMA DE PRINCIPIOS Y VALORES FUNDAMENTALES

Los principios y valores éticos que orientan el desarrollo y uso de la IA no deben abordarse como componentes aislados, sino como parte de un ecosistema profundamente interrelacionado-. La verdadera rendición de cuentas, por ejemplo, resulta inviable, sin transparencia y explicabilidad; por su parte, la equidad depende de una gobernanza de datos robusta y de la privacidad para evitar sesgos en la información de entrada y garantizar la representatividad de los datos utilizados. Una falla en cualquiera de estos principios puede tener un efecto cascada, que comprometen la legitimidad y eficacia del conjunto. Por ello se requiere un enfoque holístico e integrado, más allá de una simple lista de verificación, y garantice la coherencia, la integridad y resiliencia del marco ético en su totalidad.

Aunque muchos enfoques sobre la ética de la IA se centran en la mitigación técnica de riesgos como el sesgo o la seguridad de los datos, la reiterada invocación "bien público", los "derechos humanos" y los "valores democráticos" evidencia que la mera conformidad técnica por si sola resulta insuficiente. La IA en el ámbito del sector público debe ser proactivamente diseñada para incorporar y reflejar estos valores de manera estructural, en lugar de ser "ajustada" posteriormente con salvaguardas éticas. Esto representa una evolución de una mentalidad de cumplimiento reactivo a una filosofía de diseño proactiva, donde las consideraciones normativas, sociales, éticas y democráticas se integran de manera transversal en todo el ciclo de vida de los sistemas de IA.

Estos valores y principios ofrecen una base ética para el uso responsable de la inteligencia artificial en el sector público colombiano. Más que un marco cerrado, representa un punto de partida orientador en línea con la invitación a que cada entidad pública desarrolle sus propios principios y valores, adaptados a su misionalidad, contexto institucional y objetivos estratégicos.

La adopción de principios y valores éticos en el uso de la IA por parte del sector público del país no solo permitirá aprovechar su potencial transformador, sino también fortalecer la confianza pública, proteger los derechos fundamentales de los ciudadanos y garantizar que el avance tecnológico se traduzca en mayor bienestar social, equidad y sostenibilidad.

Así mismo se requiere que las entidades públicas establezcan medidas e indicadores claves de desempeño que evalúen de forma continua la transparencia, equidad y responsabilidad de la IA en la práctica, asegurando la alineación con los Objetivos de Desarrollo Sostenible (ODS)

4. MATERIALIZACIÓN DE LOS PRINCIPIOS Y VALORES ÉTICOS PARA UNA IA RESPONSABLE EN COLOMBIA

Para lograr un desarrollo, implementación y uso responsable y ético de la IA, es necesario contar con habilitadores, salvaguardas y buenas prácticas que permitan materializar los principios y valores fundamentales expuestos previamente. A continuación, se presentan los aspectos más relevantes de cada uno, como guía orientadora para su adopción e integración institucional.

4.1. HABILITADORES PARA UNA IA RESPONSABLE

La implementación exitosa y ética de la inteligencia artificial en el sector público nacional requiere más que la simple adopción de principios. Requiere una arquitectura de gobernanza robusta, acompañada de capacidades técnicas, institucionales y humanas clave. Estos elementos son interdependientes y conforman un ecosistema que cuando se fortalece de manera integral, puede maximizar el valor público de la IA, mitigar riesgos sistemáticos y consolidar una base sólida de confianza pública en el uso de las tecnologías emergentes.

Estrategias y coordinación de la IA a nivel nacional

Para asegurar un despliegue coherente y efectivo de la IA, las entidades públicas deben adoptar un enfoque de "gobierno integral"[4] (whole-of-government). Esto implica la formulación e implementación de estrategias y directrices nacionales articuladas que orienten, prioricen y coordinen el uso y desarrollo de la IA en función de los valores democráticos y los objetivos de política pública y las necesidades institucionales. Este enfoque promueve la coherencia intersectorial, evita duplicidades y asegura que la adopción tecnológica responda a principios éticos, criterios de equidad y metas de desarrollo sostenible.

La definición de roles y responsabilidades claras es crucial para facilitar el desarrollo, uso y escalamiento coherente de la IA. Esto incluye entre otros aspectos, la designación de figuras estratégicas como el Chief AI Officer (CAIO) y equipos interdisciplinarios permanentes compuestos por expertos en temas como seguridad digital, analítica de datos, derecho, ética, archivística y gestión documental para la asignación de responsabilidades específicas a las instituciones públicas. Esta arquitectura de responsabilidades contribuye a consolidar una estructura institucional sólida, capaz de implementar estrategias nacionales de IA de manera coordinada, ética y alineada con los objetivos de política pública. (OCDE, 2025).

Además, reforzar los mecanismos de coordinación y colaboración, tanto a nivel interministerial como entre diferentes niveles de gobierno, es vital para asegurar un enfoque holístico en la implementación de la IA. Esta articulación previene la fragmentación y evita la duplicación

[4] Este enfoque implica una colaboración entre los distintos organismos públicos que va más allá de sus respectivos ámbitos de competencia con vistas a ofrecer a la ciudadanía una respuesta conjunta desde un único organismo.

de esfuerzos, una mejor asignación de recursos, asegurando que las iniciativas de IA se desarrollen de manera coherente, complementaria y alineada con los objetivos nacionales. Asimismo, es fundamental la creación de espacios y dedicación de tiempo para la experimentación con la IA. Combinados con procesos de aprendizaje iterativo, estos entornos controlados son esenciales para desarrollar y fortalecer las capacidades en IA, identificar nuevas posibilidades y enfoques, y detectar a tiempo posibles riesgos técnicos. Estos entornos controlados permiten explorar el potencial de la IA de manera segura y eficiente.

Datos como Fundamento: Calidad, Representatividad y Protección

Los datos constituyen la base estructural sobre el cual se diseñan, construyen y entrenan los sistemas de IA. Su calidad, representatividad y trazabilidad, junto con la garantía de protección de los derechos asociados son habilitadores críticos para el desarrollo de una IA confiable, legitima y alineada con el interés público.

Por tanto, el desarrollo de marcos, directrices y mecanismos para una gobernanza de datos sólida es esencial para aprovechar y maximizar los beneficios de la IA, al tiempo que se protegen la privacidad, la propiedad intelectual y la seguridad en las iniciativas de IA. Estos esfuerzos a menudo implican una articulación multisectorial, la colaboración entre organismos reguladores, actores de la industria y la sociedad civil, con el fin de garantizar enfoques legítimos, inclusivos y alineados al interés público.

Colombia ha avanzado significativamente en el fortalecimiento de su ecosistema de datos mediante el Decreto 1389 de 2022, que establece los lineamientos generales para la gobernanza en la infraestructura de datos y se crea el Modelo de gobernanza de la infraestructura de datos. A esto se suman el Plan Nacional de Infraestructura de Datos (PNID) y la Estrategia Sectorial de Datos definida en el Plan Nacional de Desarrollo 2022 – 2026 "Colombia Potencia Mundial de la Vida". Estos instrumentos constituyen pilares fundamentales para seguir preparando al país frente a desafíos y oportunidades que plantean las tecnologías emergentes como la Inteligencia Artificial, en donde, como ya se mencionó, su desarrollo depende críticamente de la disponibilidad, calidad y gobernanza de los datos.

La exploración e implementación de tecnologías de mejora de la privacidad (PETs, por sus siglas en ingles), como la anonimización de datos sensibles utilizados en el entrenamiento de sistemas de IA, se consolida como una práctica en expansión dentro de los marcos de gobernanza responsable. Estas herramientas permiten mitigar riesgos asociados al uso de datos personales, fortalecer la confianza en los procesos algorítmicos y facilitar el cumplimiento normativo de entornos de innovación.

Para asegurar la representatividad en los datos es vital para que los sistemas de IA produzcan resultados precisos, relevantes y socialmente justos. La utilización de datos no representativos, incluyendo aquellos que reflejan desigualdades históricas o que son incompletos, pueden introducir sesgos en los algoritmos, distorsionar toma de decisiones automatizadas y limitar el desarrollo de servicios inclusivos. Una gobernanza de datos orientada a la equidad permite mitigar estos riesgos y avanzar hacia una IA que responda de manera legítima y efectiva a la diversidad de contextos sociales.

La habilitación de mecanismos efectivos y confiables para el acceso y la compartición de datos representa un desafío fundamental, en el desarrollo responsable de sistemas de IA. Persisten brechas significativas en la disponibilidad de datos que sean de alta calidad, pertinente y fácilmente accesible, lo que limita su utilidad para el entrenamiento, validación y monitoreo de modelos. Superar esta brecha, requiere un esfuerzo articulado entre sectores, orientado a fortalecer infraestructuras de datos, promover esquemas de recopilación colaborativa y acuerdos de apertura. (OCDE, 2025)

Los datos abiertos gubernamentales (OGD) son otra iniciativa clave para fortalecer la transparencia, la innovación pública y el desarrollo de sistemas IA confiables, aún más teniendo en cuenta que en "promedio solo el 46% de los conjuntos de datos gubernamentales de alto valor están disponibles como datos abiertos en la OCDE" (OCDE, 2025). Aumentar el valor de los OGD para los sistemas de IA implica la estandarización de la estructura y los formatos de los datos, así como un mayor uso de interfaces de programación de aplicaciones (APIs) para la compartición automatizada de datos. (OCDE, 2025).

Colombia ha consolidado avances significativos en su estrategia de datos abiertos, logrando que varias entidades de gobierno ya tengan publicada esta información y la actualicen de acuerdo con las necesidades de cada una, al tenor de la implementación del PNID y su Estrategia Sectorial de Datos.

Por su parte, el aprovechamiento de datos del sector privado representa una oportunidad clave para potenciar el desarrollo de aplicaciones de IA más precisas, eficientes y contextualizadas. La combinación de datos públicos (que ofrecen información demográfica y de servicios públicos) con datos del sector privado (como patrones de movilidad, comportamiento del consumidor o tendencias financieras) permite enriquecer modelos de IA, ampliar su capacidad de respuesta y generar soluciones más adaptadas a las realidades sociales y económicas.

Finalmente, la implementación consistente y operacional de la visión y estrategia de datos para la IA debe estar respaldada por una capacidad adecuada en toda la administración pública, junto con directrices y marcos legales claros. A su vez resulta prioritario impulsar la alfabetización en datos y el desarrollo de competencias en IA entre servidores públicos, como condición habilitante para una transformación digital inclusiva, sostenible y centrada en el ciudadano.

La parte operativa de la gobernanza de datos implica procesos, mecanismos y herramientas para la implementación a nivel organizacional y de equipo, asegurando que las prácticas de gestión de datos se integren en todo el ciclo de valor de los datos de la IA.

Infraestructura Digital Pública y Modelos Fundacionales de IA

La infraestructura digital pública (IDP) es un habilitador fundamental para el desarrollo y uso de la IA en el gobierno, proporcionando los cimientos tecnológicos necesarios. Entonces, el poder computacional y la infraestructura de datos son cruciales para el desarrollo y uso efectivo de la IA en el gobierno.

Se puede elegir entre soluciones en la nube, en las instalaciones (on-premises) o un enfoque híbrido, considerando la sensibilidad de los datos, los requisitos regulatorios, las limitaciones

presupuestarias y los objetivos a largo plazo. Las soluciones on-premises ofrecen mayor control y seguridad para aplicaciones altamente sensibles, mientras que las soluciones en la nube proporcionan escalabilidad, eficiencia de costos y acceso a tecnologías de IA de vanguardia. Por su parte, un enfoque híbrido puede ofrecer un equilibrio.

También es fundamental considerar la huella ambiental de la IA, ya que la demanda global de capacidad de centros de datos listos para IA se triplicará para 2030, lo que subraya la necesidad de prácticas sostenibles y la gestión del impacto ambiental (OCDE, 2025). Al respecto, existe una tendencia creciente hacia modelos de IA más pequeños y especializados (SLMs) que consumen menos recursos, requieren menos datos y son menos costosos (OCDE, 2025). Algunos países, como Corea, están construyendo centros de datos compartidos y nubes gubernamentales para modernizar la tecnología, asegurar el cumplimiento y lograr eficiencias de costos. Por su parte, la Infraestructura Nacional de Datos (IND) de Brasil es una iniciativa estratégica que establece políticas, estándares y mecanismos de gobernanza para organizar y compartir datos del sector público de manera segura y eficiente, promoviendo los principios FAIR (Encontrables, Accesibles, Interoperables y Reutilizables).

De otra parte, se pueden desarrollar modelos propios fundacionales de IA[5], o construir sobre los existentes, adaptándolos al contexto nacional y público. El "ajuste fino" (fine-tuning) de un modelo fundacional en un conjunto de datos más específico, puede reducir significativamente los costos financieros y de tiempo. Aunque existen modelos propietarios desarrollados por empresas privadas, también se pueden aprovechar modelos de código abierto "pre-entrenados", donde la arquitectura del modelo y los pesos son públicamente accesibles para su modificación y uso (OCDE, 2025). La adaptación de estos modelos al contexto nacional y del sector público puede reducir drásticamente los costos de adopción para los equipos.

Sin embargo, el desarrollo y uso de modelos fundacionales conllevan riesgos, como la generación de resultados sesgados o inexactos, lo cual llama la atención con el creciente interés en invertir en modelos fundacionales nacionales o regionales para mejorar la soberanía tecnológica y reflejar la diversidad lingüística y cultural. (OCDE, 2025)

Finalmente, las herramientas comunes de IA, a menudo construidas sobre modelos fundacionales, actúan como una forma de IDP que habilita y mejora otros servicios en el gobierno. Estas herramientas proporcionan una capa de servicio compartida que puede automatizar tareas rutinarias, mejorar la interacción con el usuario y optimizar la prestación de servicios. Los chatbots, por ejemplo, pueden gestionar un gran volumen de consultas ciudadanas, proporcionando respuestas instantáneas y liberando recursos humanos.

Para ser consideradas IDP, estas herramientas de IA deben satisfacer una necesidad básica común y ser utilizables en una amplia gama de organizaciones del sector público.

[5] Los modelos fundacionales son una forma de inteligencia artificial generativa (IA generativa). Ellos generan resultados a partir de una o más entradas (indicaciones) en forma de instrucciones en lenguaje humano. (https://aws.amazon.com/es/what-is/foundation-models/#:~:text=Los%20modelos%20fundacionales%20son%20una,de%20instrucciones%20en%20lenguaje%20humano.)

Desarrollo de Capacidades y Talento en el Servicio Público

La brecha de habilidades es uno de los desafíos más significativos para la adopción de IA confiable en el gobierno. Para abordarlo, es crucial fomentar las capacidades y el talento en el servicio público. En ese marco, evaluar las necesidades de los diferentes grupos de usuarios es fundamental para asegurar la adopción y el uso efectivo de la IA. Estos grupos varían desde usuarios generales de sistemas de IA hasta líderes institucionales, profesionales de datos y digitales, y roles más especializados en IA, cada uno con requisitos de alfabetización y habilidades específicas.

Por ejemplo, los servidores públicos no especializados necesitan capacitación en alfabetización general sobre datos y tecnologías de IA, su uso óptimo y consideraciones éticas y legales. Por su parte los funcionarios líderes son esenciales para impulsar la adopción de la IA, y requieren de una visión estratégica y una comprensión ejecutiva de las capacidades, impactos y gestión de riesgos de la IA. Otros perfiles de servidores más especializados pueden liderar el diseño e implementación de servicios, necesitando un alto grado de alfabetización en IA para un despliegue efectivo, mientras que los servidores públicos especialistas en IA, aunque constituyen una parte más pequeña de la fuerza laboral, son críticos para el desarrollo, despliegue y gestión de sistemas de IA, requiriendo esfuerzos de atracción, retención, promoción de incentivos y desarrollo específicos.

Entonces, resulta necesario diseñar rutas formativas diferenciadas que respondan a los distintos niveles de especialización requeridos en el sector público, lo que implica desarrollar programas de capacitación adaptados tanto para quienes interactúan cotidianamente con herramientas de IA, como para quienes lideran la toma de decisiones estratégicas sobre su adopción e implementación. En complemento, la actualización constante y el acceso a materiales educativos relevantes son componentes esenciales para cerrar la brecha de conocimiento y fomentar una cultura organizacional orientada a la innovación y la ética digital.

Adicionalmente, la creación de incentivos que promuevan la participación en iniciativas de formación y la certificación de competencias digitales puede elevar la motivación y profesionalización del personal. Es clave también instaurar mecanismos de evaluación y retroalimentación que permitan ajustar los contenidos formativos conforme evolucionan las tecnologías y las necesidades institucionales.

La preparación de los servidores públicos para la IA implica una combinación de habilidades digitales fundamentales y una alfabetización específica en datos e IA. Esto incluye comprender el potencial de la transformación digital, las necesidades de los usuarios, la colaboración abierta y el uso confiable de datos y tecnología. La alfabetización en IA es necesaria para que las personas evalúen críticamente las tecnologías de IA, se comuniquen y colaboren eficazmente con la IA, y la utilicen como herramienta en diversos contextos, incluyendo la comprensión de los sistemas de IA, el manejo de datos y la ética.

El desarrollo de habilidades y talento en IA debe abordarse mediante prácticas de desarrollo interno (marcos de competencias, aprendizaje formal e informal) y estrategias de reclutamiento externo (compensación competitiva, trayectorias profesionales claras y flexibilidad laboral) para cerrar las brechas de habilidades críticas. Las asociaciones con la industria y la academia son fundamentales para obtener el apoyo y la experiencia necesarios cuando la capacidad interna es limitada.

La facilitación de conexiones y el intercambio de conocimientos entre los servidores públicos a través de comunidades de práctica y redes es vital para la colaboración, el aprendizaje y la identificación de problemas comunes. Estos espacios también sirven como canales para la retroalimentación de los usuarios sobre los sistemas internos de IA, ayudando a superar los desafíos iniciales de adopción y a replicar proyectos exitosos de manera más efectiva. Finalmente, la formación de equipos multidisciplinarios es esencial para asegurar que las iniciativas de IA se beneficien de diversas perspectivas y experiencias. La sensibilidad y complejidad de la IA requieren la colaboración de expertos en tecnología, ética, derecho y políticas públicas para establecer un enfoque estratégico e inclusivo en el uso de la IA en toda la administración pública.

Este enfoque integral en el desarrollo de capacidades debe ir acompañado de liderazgos comprometidos y de estructuras organizacionales que favorezcan la colaboración interdisciplinaria, permitiendo que la experimentación y el aprendizaje se consoliden como valores centrales en la gestión pública contemporánea.

Inversión y Financiamiento Estratégico

La inversión financiera es un habilitador crítico para el avance de la IA en el gobierno, permitiendo la experimentación y el escalamiento de iniciativas. Esto hace que el fortalecimiento de la planificación estratégica para una inversión coherente sea fundamental, requiriendo de la coordinación entre las autoridades presupuestarias, de gobierno digital y de contratación pública, para identificar las necesidades de IA y alinearlas con los recursos disponibles y las posibles adquisiciones del sector privado. (OCDE, 2025) (Federal Ministry of the Interior, 2025).

También se pueden utilizar herramientas de gestión ya existentes, como mecanismos de evaluación de propuestas de valor y de riesgos de inversión, con tal de reforzar la coherencia en las decisiones de inversión en IA, asegurando el cumplimiento de las regulaciones y estándares de políticas (OCDE, 2025).

Por otro lado, el financiamiento de la IA en el sector público requiere recursos específicos para fomentar su adopción y evitar esfuerzos fragmentados. (OCDE, 2025).

Por último, los mecanismos de monitoreo para una inversión coherente son esenciales para asegurar que el desarrollo y despliegue de las inversiones en IA se realicen dentro del presupuesto y el cronograma, entregando los resultados esperados. También se pueden utilizar herramientas de monitoreo para supervisar la gestión y el desarrollo de sistemas de IA en toda la administración, incluyendo el desarrollo de indicadores clave de rendimiento y enfoques estructurados para gestionar los desarrollos continuos de los sistemas de IA a través de la gestión de carteras de TI.

Procesos de contratación pública ágiles y colaboración con actores no gubernamentales

Los procesos de contratación pública deben ser ágiles y estratégicos para facilitar la adquisición de soluciones de IA y fomentar la colaboración con actores no gubernamentales. Por tanto, requieren de una preparación y planificación cuidadosa como punto de partida

para lograr que sean flexibles, eficientes y que fomenten una amplia participación. Para ello, es fundamental el establecimiento de equipos multidisciplinarios, la evaluación de los datos y enfoques de gobernanza existentes, y la identificación de riesgos en todo el ciclo de vida de la IA con sus respectivas estrategias de mitigación.

La adopción de métodos de contratación ágiles e innovadores también es crucial para acelerar la incorporación de nuevas tecnologías y el desarrollo y uso confiable de la IA. Esto puede incluir concursos tecnológicos, demostraciones, procesos de diálogo competitivo y desafíos.

También se pueden aprovechar los acuerdos marco de políticas para establecer reglas, prioridades y directrices generales de contratación que pueden ser específicas para la IA o aplicarse a proveedores clave. Estos acuerdos son importantes para la interacción con el sector privado y pueden generar marcos predefinidos que permitan la contratación de IA bajo principios rectores más amplios.

Igualmente es necesario considerar la contratación como habilitador para el bien público y la IA confiable. Esto implica que los países consideren cómo se preparan para el mercado de IA y cómo la contratación pública puede ser una herramienta estratégica para moldear dicho mercado y asegurar que los sistemas de IA se alineen con los estándares gubernamentales. La contratación pública juega un papel crítico en el establecimiento de requisitos para los sistemas de IA que reflejen los valores públicos, garantizando la rendición de cuentas, la seguridad y la equidad en la adopción de la IA.

Otro enfoque puede ser el Impulsar startups GovTech, lo que implica la colaboración entre el gobierno y startups, innovadores, "intraemprendedores" gubernamentales y la academia, para el desarrollo de soluciones digitales innovadoras para el gobierno. Este enfoque complementa las capacidades gubernamentales existentes para procesos y servicios ágiles, centrados en el usuario. Esto incluye aprovechar las colaboraciones GovTech para experimentar y desarrollar sistemas de IA que aborden desafíos gubernamentales y sociales. España (España Digital 2026, 2025).

Por otro lado, la expansión del potencial de la IA a través de alianzas es fundamental, ya que existen beneficios significativos en las asociaciones intersectoriales activas, donde cada sector tiene un rol y contribuciones concretas. Estas alianzas facilitan la colaboración entre entidades públicas y especialistas en IA de otros sectores, incluyendo empresas privadas, instituciones académicas y fundaciones, fomentando el desarrollo e implementación de soluciones de vanguardia. Las asociaciones público-privadas (APP) son un tipo común de acuerdo en este sentido. (OCDE, 2025).

Para finalizar, vale la pena mencionar que, dentro de procesos de contratación pública, la interrelación de estos habilitadores tiene un impacto directo en la confianza pública. Ninguno de ellos funciona por separado, es decir que, una inversión sin talento o estrategia es ineficaz y la calidad de los datos afecta directamente el desempeño y equidad de la IA. El desarrollo exitoso de la IA requiere fortalecer todos los componentes, si falta alguno el valor generado será limitado.

En la **Tabla 3** se presenta un resumen de los habilitadores, al igual que acciones clave para el gobierno nacional.

Tabla 3. Habilitadores para una IA confiable y acciones clave para el gobierno nacional.

Habilitador	Aspecto	Acciones Clave
Estrategia	Estrategias de Gobierno Integral (Whole-of-Government)	Formular e implementar estrategias de IA que alineen su uso con los objetivos de gobierno y los ODS.
	Definición de Roles y Responsabilidades	Designar y empoderar a líderes de IA (ej. CAIOs) y definir responsabilidades claras para el desarrollo y uso de IA en cada entidad.
	Coordinación de Esfuerzos de IA	Establecer mecanismos de coordinación interinstitucional y entre niveles de gobierno para evitar duplicidades y fragmentación.
	Espacios de Experimentación	Crear entornos seguros para la prueba y el aprendizaje iterativo de soluciones de IA.
Datos	Privacidad, Seguridad y Derechos de Propiedad Intelectual	Desarrollar e implementar marcos de gobernanza de datos que protejan la privacidad, la seguridad y la propiedad intelectual.
	Representatividad de los Datos	Invertir en la recopilación y limpieza de datos representativos, incluyendo recursos lingüísticos y culturales, para mitigar sesgos.
	Acceso y Compartición Efectiva de Datos	Mejorar la disponibilidad de datos de alta calidad mediante la estandarización y el uso de APIs para la compartición automatizada.
	Aprovechamiento de Datos del Sector Privado	Establecer marcos para la colaboración segura y ética en el uso de datos del sector privado.
	Reforma de la Gobernanza de Datos	Implementar estrategias de datos integrales y definir roles de liderazgo para la gestión de datos en entornos IA.
	Implementación Consistente y Operacional	Asegurar la capacidad para una implementación consistente de la gobernanza de datos, con directrices claras y cumplimiento normativo.
Infraestructura	Poder Computacional e Infraestructura de Datos	Evaluar y adoptar soluciones de infraestructura (nube, on-premises, híbrida) que soporten el desarrollo y uso de IA, considerando el impacto ambiental.
	Modelos Fundacionales de IA	Explorar el desarrollo o ajuste fino de modelos fundacionales nacionales o regionales para la soberanía tecnológica y la diversidad lingüística.
	Herramientas Comunes de IA	Implementar herramientas de IA compartidas (ej. chatbots) que sirvan como capa de servicio común para la automatización y mejora de la interacción.

Habilitador	Aspecto	Acciones Clave
Talento	Evaluación de Necesidades por Grupo de Usuarios	Realizar evaluaciones de necesidades para adaptar los programas de capacitación en IA a los diferentes perfiles de servidores públicos.
	Preparación de Servidores Públicos	Desarrollar programas de alfabetización digital y en IA que combinen habilidades fundamentales con conocimientos específicos.
	Desarrollo de Habilidades y Talento en IA	Implementar estrategias de desarrollo interno (formación) y reclutamiento externo (compensación, flexibilidad) para atraer y retener talento.
	Facilitación de Conexiones e Intercambio de Conocimientos	Fomentar comunidades de práctica y redes para la colaboración y el intercambio de experiencias entre servidores públicos.
	Equipos Multidisciplinarios	Crear equipos con expertos en tecnología, ética, derecho y políticas públicas para un enfoque integral de la IA.
Inversión	Fortalecimiento de la Planificación Estratégica	Coordinar la planificación de inversiones en IA entre autoridades presupuestarias, de gobierno digital y de contratación.
	Financiamiento de la IA en el Gobierno	Asignar recursos financieros específicos y dirigidos para la experimentación y el escalamiento de proyectos de IA.
	Mecanismos de Monitoreo de Inversiones	Implementar herramientas de monitoreo (KPIs, gestión de cartera de TI) para supervisar el progreso y el retorno de la inversión de las iniciativas de IA.
Contratación y Alianzas	Preparación y Planificación Cuidadosa	Establecer equipos multidisciplinarios y realizar análisis de riesgos en la fase de preparación de la contratación de IA.
	Métodos de Contratación Ágiles e Innovadores	Utilizar concursos, demostraciones y diálogos competitivos para adquirir soluciones de IA de manera eficiente.
	Acuerdos Marco de Políticas	Establecer directrices de contratación que aseguren la alineación de las soluciones de IA con los valores públicos y estándares gubernamentales.
	Aprovechamiento de Startups GovTech	Colaborar con startups y el mundo académico en soluciones digitales innovadoras a través de la co-creación y experimentación.
	Expansión del Potencial de la IA a través de Alianzas	Fomentar asociaciones público-privadas y con la academia para el desarrollo e implementación de soluciones de IA de vanguardia.

4.2. SALVAGUARDAS Y BUENAS PRÁCTICAS PARA LA IMPLEMENTACIÓN DE IA

La implementación de sistemas de inteligencia artificial en el sector público, si bien promete eficiencia y mejora de servicios, también introduce complejidades y riesgos que deben ser gestionados proactivamente. Estas complejidades y riesgos hacen que varios procesos internos o lógica de toma de decisiones de la IA sean opacos e incomprensibles para los humanos, lo que se denomina como la problemática de la Caja Negra.

Las salvaguardas y las buenas prácticas son esenciales para resolver dicha problemática y garantizar que **la IA se utilice de manera ética,** responsable y que fomente la confianza pública.

Respecto a la transparencia y explicabilidad para fomentar la confianza pública

La transparencia en la IA implica hacer que los algoritmos sean abiertos, comprensibles y accesibles al escrutinio público, resolviendo así el problema de la caja negra. Esto incluye la divulgación de los procesos mediante los cuales se utilizan y las decisiones a las que contribuyen, a través de instrumentos de transparencia proactiva, que son fundamentales para que el sector público comparta información sobre sus sistemas de IA sin necesidad de solicitudes específicas.

Actualmente a través de la Sentencia T-067 de 2025, la Corte Constitucional Colombiana establece un marco conceptual jurídico que sustenta este concepto indicando que la finalidad de la transparencia algorítmica es "que el público en general pueda comprender cómo los sistemas de toma de decisiones automatizadas (SDA) procesan los datos que capturan y cómo toman decisiones que afectan la vida de las personas." (Corte Constitucional República de Colombia, 2025). La Procuraduría General de la Nación y la Defensoría del Pueblo, con el apoyo técnico de Agencia Nacional Digital desarrollaron lo dispuesto en esta Sentencia, mediante la Directiva 007 de 2025.

Respecto a los instrumentos de transparencia proactiva se cuentan los registros públicos de sistemas de IA, repositorios centralizados y de fácil búsqueda que consolidan información sobre los sistemas de IA en uso, detallando su propósito, el sector al que aplican y las jurisdicciones afectadas.

Otro instrumento es la publicación del código fuente y la documentación de los algoritmos públicos. Esta es considerada una buena práctica, especialmente para audiencias técnicas, ya que permite examinar y verificar el funcionamiento de los sistemas, promoviendo la rendición de cuentas y la confianza. En su defecto, la publicación de documentación exhaustiva es crucial, tal como lo exige Francia con la ley de República Digital, dónde los organismos gubernamentales deben poner a disposición del público, en un formato abierto y fácilmente reutilizable, las reglas que definen el algoritmo utilizado y con el cual se da la toma de decisiones (Kiteworks, 2025).

Otras formas de transparencia proactiva incluyen las publicaciones impulsadas por el usuario, donde algunos gobiernos divulgan proactivamente información solicitada

frecuentemente sobre algoritmos, y las respuestas automatizadas por interacción, que proporcionan información sobre sistemas automatizados durante las interacciones del usuario con servicios gubernamentales.

Por otro lado, los instrumentos de transparencia reactiva permiten al gobierno responder a solicitudes específicas de información sobre un algoritmo o su uso, generalmente bajo leyes de Acceso a la Información (ATI). Sin embargo, estos regímenes pueden tener limitaciones si no están diseñados específicamente para la transparencia algorítmica, como problemas de gestión de registros o excepciones relacionadas con la propiedad intelectual.

Respecto a la explicabilidad de la IA, ésta se refiere a la capacidad de hacer inteligible la lógica detrás de las decisiones o comportamientos de un sistema de IA, como si hubieran sido producidos por una persona que razona y utiliza evidencia. Esto significa que, más allá de la mera descripción técnica, se debe poder comunicar el "porqué" de una decisión de IA en un lenguaje comprensible para los no especialistas.

Finalmente, la implementación centrada en el ser humano es clave para lograr la transparencia y explicabilidad. Los protocolos de implementación deben considerar las necesidades, competencias y capacidades de las personas afectadas por los sistemas de IA. Esto asegura que las explicaciones sean relevantes y accesibles, fomentando una comprensión genuina y, por ende, la confianza.

Respecto a la rendición de cuentas y supervisión a lo largo del ciclo de vida de la IA
La rendición de cuentas es un principio fundamental que exige que los seres humanos sean responsables de su papel en todo el flujo de trabajo del proyecto de IA, y que los resultados sean trazables de principio a fin.

Una referencia interesante es la presentada por el Alan Turing Institute (The Alan Turing Institute, 2024) quien desglosa la rendición de cuentas en varios conceptos interrelacionados: La responsabilidad (Answerability), que implica establecer una cadena continua de responsabilidad humana a lo largo de todo el flujo de trabajo del proyecto de IA, sin lagunas. Las autoridades humanas competentes deben ofrecer explicaciones y justificaciones claras, comprensibles y coherentes sobre la lógica de los resultados del sistema de IA y los procesos detrás de su producción y uso.

La auditabilidad (Auditability), la cual se refiere a la capacidad de demostrar la responsabilidad en las prácticas de diseño, desarrollo y despliegue, así como la capacidad de justificar los resultados. Esto se logra mediante la trazabilidad de todas las etapas del ciclo de vida de la IA, con documentación accesible y fácilmente comprensible.

La rendición de cuentas anticipatoria (Anticipatory Accountability), que se centra en asegurar la rendición de cuentas durante las etapas de diseño y desarrollo del proyecto (ex-ante), priorizando la prevención de daños. Por ejemplo, documentar las decisiones de selección de características en la fase de preprocesamiento y la justificación de esas elecciones. Por otro lado, la rendición de cuentas remedial (Remedial Accountability), aborda la rendición de cuentas después del despliegue del proyecto (ex-post), remediando problemas y proporcionando justificaciones para el impacto del sistema. Un ejemplo sería recurrir a registros de medidas de mitigación de sesgos para explicar una decisión de contratación desfavorable generada por IA.

La gestión de riesgos para sistemas de IA de alto riesgo es crucial para identificar y mitigar los riesgos inherentes. Esto implica establecer guías sobre niveles de riesgo aceptables para diferentes usos y contextos, y es necesario tanto antes como después del despliegue de los sistemas de IA. Por ejemplo, el Marco de Gestión de Riesgos de IA del NIST de EE. UU. es una herramienta que ayuda a las organizaciones a identificar riesgos únicos y proponer acciones de mitigación.

Las evaluaciones de impacto, especialmente las Evaluaciones de Impacto Algorítmico (AIA), ayudan a las organizaciones públicas a anticipar y evaluar cómo un algoritmo puede funcionar en un contexto específico. Se realizan en las primeras etapas de desarrollo (ex-ante) y pueden repetirse después del despliegue (ex-post). El objetivo principal de las AIA ex-ante es evaluar los impactos potenciales de un sistema algorítmico en las economías y sociedades, proporcionando un mecanismo de rendición de cuentas. La "Directiva sobre la Toma de Decisiones Automatizadas" de Canadá, por ejemplo, requiere una AIA que considera varios factores y asigna una puntuación de riesgo que prescribe ciertas acciones. El Consejo de Europa también ha desarrollado la Evaluación de Impacto en Derechos Humanos, Democracia y Estado de Derecho (HUDERIA), que ofrece una metodología para la evaluación y mitigación de riesgos.

Las auditorías algorítmicas implican el monitoreo continuo del comportamiento del sistema después del despliegue para identificar riesgos esperados o inesperados y asegurar una implementación responsable. Estas auditorías pueden ser técnicas (examinando entradas/salidas), de cumplimiento (verificando el cumplimiento de requisitos regulatorios), inspecciones regulatorias (monitoreando el comportamiento del sistema a lo largo del tiempo) o sociotécnicas (evaluando el impacto del sistema en procesos y contextos sociales más amplios). Las auditorías sirven para evaluar el rendimiento, asegurar el cumplimiento, detectar sesgos, mejorar la transparencia y explicabilidad, y responsabilizar a las organizaciones.

Para abordar de manera efectiva la problemática de la 'Caja Negra', las entidades pueden desarrollar capacidades internas y externas para realizar no solo auditorías técnicas y de cumplimiento, sino fundamentalmente auditorías sociotécnicas. Estas auditorías son esenciales para evaluar cómo el sistema impacta los procesos sociales y los contextos más amplios, superando las limitaciones de la mera transparencia técnica.

Las evaluaciones de capacidades de los sistemas de IA son similares a las evaluaciones de impacto, pero se centran específicamente en la probabilidad de resultados dañinos debido a las capacidades de un sistema de IA. Suelen realizarse en las primeras etapas de desarrollo, pero también pueden hacerse después del despliegue.

El empoderamiento de los órganos de supervisión y asesoramiento es fundamental. Instituciones como las Entidades Fiscalizadoras Superiores (EFS) están ampliando sus actividades de auditoría para escrutar algoritmos de IA en cuanto a precisión, seguridad, eficacia, transparencia y equidad. Los órganos asesores, como los consejos de IA, proporcionan orientación experta y recomendaciones sobre cuestiones emergentes de IA. En Colombia, el Consejo Superior de la Judicatura emitió el Acuerdo PCSJA24-12243, estableciendo directrices para el uso respetuoso, responsable, seguro y ético de la IA en la Rama Judicial, enfatizando la necesidad de programas de capacitación y guías para su uso.

En la **Tabla 4** se presenta un resumen mecanismos de transparencia y rendición de cuentas en la IA pública.

Tabla 4. Mecanismos de transparencia y rendición de cuentas en la IA.

Mecanismo	Aspecto	Función Principal	Implicación para Colombia
Transparencia Proactiva	Registros Públicos de Sistemas de IA	Centralizar y hacer pública la información sobre los sistemas de IA en uso, sus propósitos y alcances.	Mantener actualizado el registro de sistemas de IA utilizados por entidades nacionales.
	Publicación de Código Fuente y Documentación	Permitir el escrutinio técnico y la verificación del funcionamiento de los algoritmos.	Publicar documentación exhaustiva (o código fuente cuando sea posible) de los algoritmos de IA de alto impacto, siguiendo el ejemplo de PretorIA.
	Publicaciones Proactivas Impulsadas por el Usuario	Divulgar información frecuente sobre algoritmos sin necesidad de solicitudes.	Identificar y publicar proactivamente preguntas frecuentes y explicaciones sobre los sistemas de IA más consultados por los ciudadanos.
	Respuestas Automatizadas por Interacción	Proporcionar información sobre sistemas automatizados durante las interacciones del usuario.	Asegurar que los chatbots y asistentes virtuales informen a los usuarios cuando están interactuando con IA y cómo se procesa su información.
Transparencia Reactiva	Respuesta a Solicitudes Específicas	Atender solicitudes de información sobre algoritmos o su uso, bajo leyes de Acceso a la Información.	Fortalecer la capacidad de las entidades para responder a solicitudes ciudadanas sobre el funcionamiento y los impactos de la IA.
Evaluaciones de Impacto	Evaluación de Impacto Algorítmico (AIA)	Anticipar y evaluar cómo un algoritmo puede funcionar en un contexto específico y sus impactos.	Realizar AIAs obligatorias para sistemas de IA de alto riesgo o alto impacto, como los utilizados en justicia o detección de fraude.
	Evaluación de Capacidades de los Sistemas de IA	Evaluar la probabilidad de resultados dañinos debido a las capacidades de un sistema de IA.	Integrar análisis de riesgo de capacidades en las fases tempranas de desarrollo de IA.
Auditorías	Auditorías Algorítmicas	Monitorear continuamente el comportamiento del sistema después del despliegue para identificar riesgos.	Desarrollar capacidades internas y externas para realizar auditorías técnicas, de cumplimiento y sociotécnicas de los sistemas de IA.
Órganos de Supervisión	Órganos de Supervisión y Asesoramiento	Escudriñar algoritmos de IA en cuanto a precisión, seguridad, eficacia, transparencia y equidad.	Fortalecer los órganos de control para auditar el uso de IA, y establecer un consejo asesor de IA a nivel nacional.



Respecto a la mitigación de sesgos y promoción de la equidad (fairness)

La equidad es un pilar central en la ética de la IA, y su promoción requiere un esfuerzo continuo para mitigar los sesgos en todas las etapas del ciclo de vida del sistema. La equidad debe desarrollarse a partir del principio de no daño discriminatorio, que se refiere al umbral mínimo de equidad. Esto exige que los diseñadores y usuarios de sistemas de IA que prioricen la mitigación de sesgos y la exclusión de influencias discriminatorias en los resultados e implementaciones de sus modelos, significando que las decisiones y comportamientos de los modelos de IA no deben generar impactos discriminatorios o inequitativos en individuos y comunidades afectadas.

La equidad como pilar central en la ética de la IA cuenta con diferentes componentes (The Alan Turing Institute, 2019) que describiremos a continuación:

La equidad en los datos (Data Fairness) es un componente necesario de la equidad algorítmica. Si los resultados de un proyecto de IA se generan a partir de conjuntos de datos sesgados, comprometidos o distorsionados, las partes interesadas no estarán adecuadamente protegidas del daño discriminatorio. Esto implica asegurar que los datos de entrenamiento sean representativos, suficientes, íntegros en su origen, oportunos y relevantes. La creación de un "Dataset Factsheet" es una buena práctica para documentar la procedencia, el procesamiento y las decisiones sobre los datos, garantizando la calidad y la mitigación de sesgos.

La equidad en el diseño (Design Fairness) se refiere a las precauciones que deben tomarse en todo el flujo de trabajo del proyecto de IA para evitar que el sesgo tenga una influencia discriminatoria. Esto abarca desde la formulación del problema (donde las elecciones sobre cómo clasificar y estructurar las entradas pueden introducir sesgos), hasta el preprocesamiento de datos, la determinación de características y la evaluación de estructuras analíticas. Es crucial que equipos multidisciplinarios colaboren para identificar y mitigar sesgos en cada una de estas etapas.

La equidad en los resultados (Outcome Fairness) es el componente que implica definir y medir la equidad de los impactos y resultados del sistema de IA. Existen diversas definiciones formalizables de equidad como: i) la paridad demográfica/estadística (donde cada grupo recibe beneficios en proporciones iguales); ii) la paridad de tasa de verdaderos positivos (donde la precisión del modelo es equivalente entre subgrupos de población); iii) la paridad de tasa de falsos positivos (evitando clasificaciones erróneas desproporcionadas) y, iv) la paridad de valor predictivo positivo. También se consideran la equidad individual (tratar a individuos similares de manera similar) y la equidad contractual (donde una decisión sobre un individuo sería la misma si este perteneciera a un grupo diferente en un mundo alternativo cercano). La preparación de una "Declaración de Posición sobre Equidad" (Fairness Position Statement - FPS) es una buena práctica para explicitar y justificar los criterios de equidad empleados en el sistema de IA.

Finalmente, la equidad en la implementación (Implementation Fairness) se refiere a la preparación y capacitación de los usuarios para el despliegue responsable y sin sesgos del sistema de IA. Los sistemas de soporte a la decisión automatizados presentan riesgos de sesgo y mala aplicación en el punto de entrega. Es crucial capacitar a los implementadores para que comprendan las limitaciones de la IA, eviten el sesgo de automatización (la tendencia a confiar excesivamente en los resultados de la IA sin escrutinio) y fomenten el juicio humano.

A través de la equidad, se da respuesta al desafío de la "caja negra" en los sistemas de IA, particularmente aquellos basados en aprendizaje profundo, el cual dificulta la comprensión de cómo producen un resultado dado, lo que erosiona la transparencia, la rendición de cuentas y la detección de sesgos. Si las decisiones asistidas por IA en el gobierno no pueden explicarse de manera inteligible a los ciudadanos, se socava la confianza pública y la legitimidad de las acciones gubernamentales.

En contextos de alto impacto, como la justicia o la asignación de beneficios, esto puede generar resistencia ciudadana y la percepción de injusticia. La solución a esta problemática no reside únicamente en la transparencia técnica (la publicación del código), sino en la "explicabilidad" en términos humanos y la "auditabilidad" sociotécnica. Las auditorías algorítmicas deben ir más allá de los aspectos técnicos para evaluar cómo el sistema impacta los procesos sociales y los contextos más amplios.

La T**abla 5** presenta un resumen de los tipos de equidad en la IA y cuáles son sus implicaciones para el sector público.

Tabla 5. Tipos de equidad en la IA y sus implicaciones para el sector público.

Tipo de Equidad	Definición Breve	Implicación para el Diseño/ Implementación de IA en el Gobierno
Paridad Demográfica/ Estadística	Un resultado es justo si cada grupo en el conjunto seleccionado recibe beneficios en proporciones iguales o similares, sin correlación entre un atributo sensible y el resultado.	Prevenir el impacto dispar en grupos protegidos o vulnerables al diseñar sistemas de asignación de beneficios o servicios.
Paridad de Tasa de Verdaderos Positivos	Un resultado es justo si las tasas de "verdaderos positivos" de una predicción o clasificación algorítmica son iguales entre grupos.	Asegurar que la precisión del modelo sea equivalente entre subgrupos de población, por ejemplo, en sistemas de detección de enfermedades o elegibilidad.
Paridad de Tasa de Falsos Positivos	Un resultado es justo si no maltrata desproporcionadamente a personas de un grupo social al clasificarlas erróneamente a una tasa más alta que a miembros de otro grupo.	Evitar que ciertos grupos sean injustamente penalizados o etiquetados erróneamente (ej., en sistemas de detección de fraude o riesgo criminal).
Valor Predictivo Positivo (VPP) Paridad	Un resultado es justo si las tasas de VPP (fracción de casos positivos correctamente predichos de todos los positivos predichos) son iguales entre grupos.	Asegurar que la probabilidad de que los miembros de diferentes grupos realmente tengan la cualidad predicha sea la misma.
Equidad Individual	Un resultado es justo si trata a individuos con calificaciones relevantes similares de manera similar.	Diseñar sistemas que apliquen reglas de manera consistente a individuos con características pertinentes idénticas, independientemente de otros atributos.
Equidad Contrafactual	Un resultado es justo si una decisión automatizada sobre un individuo sería la misma si ese individuo perteneciera a un grupo diferente en un mundo alternativo cercano.	Permitir a los afectados entender qué factores, si hubieran cambiado, habrían influido en un resultado diferente, facilitando recursos y apelaciones.

5. CONCLUSIONES Y RECOMENDACIONES CLAVE PARA COLOMBIA

Conclusiones Generales

La inteligencia artificial (IA) representa una herramienta estratégica clave para modernizar y fortalecer el sector público en Colombia, mejorando la productividad, la atención ciudadana y la confianza institucional. Su implementación debe abordarse con una visión integral que combine principios éticos, marcos de gobernanza adaptados al contexto nacional e inversión en capacidades clave como datos, infraestructura, talento y financiamiento. En esta visión debe prevalecer la protección de los derechos humanos y la dignidad sobre la eficiencia y el crecimiento económico. Un objetivo principal debe ser que las entidades públicas diseñen sus sistemas de IA para maximizar las capacidades humanas y con visión de optimización y mejora de procesos, evitando la instrumentalización o el control invasivo.

El éxito de una IA responsable en el gobierno dependerá de la colaboración multisectorial entre Estado, sector privado, academia y sociedad civil, así como del fomento de iniciativas GovTech y asociaciones público-privadas. Además, es indispensable adoptar una cultura de aprendizaje continuo, con mecanismos sólidos de monitoreo y evaluación que permitan ajustar las estrategias conforme evolucionen la tecnología y el entorno. Finalmente, se destaca la necesidad de mantener la adaptabilidad institucional, priorizando siempre los derechos humanos como eje central en el desarrollo y uso de sistemas de IA.

Recomendaciones

Basado en el análisis de los principios éticos, los habilitadores y las salvaguardas, se formulan las siguientes recomendaciones para las entidades públicas nacionales:

1. Invertir Estratégicamente en Habilitadores Clave:

- **Datos:** Inversión en la creación y mantenimiento de una infraestructura de datos públicos de alta calidad, representativa y accesible, garantizando la privacidad y la seguridad. Esto incluye la estandarización de datos y el fomento de la compartición segura entre entidades.

- **Talento:** Implementar programas de capacitación masiva en IA y alfabetización digital para todos los niveles de la administración pública, desde usuarios generales hasta especialistas. Crear incentivos para atraer y retener talento especializado en IA en el sector público, explorando modelos de colaboración con la academia y el sector privado.

- **Infraestructura:** Evaluar y adoptar soluciones de poder computacional (nube, híbrida) que soporten las necesidades de IA, considerando la eficiencia y la sostenibilidad ambiental.

2. Implementar Sistemas Integrales de Auditoría y Evaluación:

- Desarrollar capacidades internas y externas para realizar Evaluaciones de Impacto Algorítmico (AIA) y auditorías sociotécnicas de manera regular y obligatoria para sistemas de IA de alto riesgo. Tanto las Evaluaciones como las Auditorías deben incluir mecanismos robustos para la verificación de la integridad de la información y la mitigación proactiva de riesgos derivados de la manipulación con contenido sintético o deepfakes.

- Asegurar la rendición de cuentas mediante la identificación temprana de riesgos, la verificación del cumplimiento de los principios éticos y la evaluación del impacto real en los ciudadanos. Establecer mecanismos de recurso y apelación claros para las decisiones asistidas por IA.

3. Fomentar la Transparencia Proactiva y la Explicabilidad:

- Establecer y mantener un registro público y accesible de todos los sistemas de IA utilizados por las entidades nacionales, detallando su propósito, los datos que utilizan, los mecanismos de operación y las evaluaciones de impacto.

- Priorizar el diseño de sistemas de IA que sean interpretables y capaces de explicar sus decisiones en un lenguaje claro y comprensible para los no especialistas, especialmente en áreas sensibles como la justicia.

- Para el caso colombiano, el registro público debe ser obligatorio y, en cumplimiento de la Sentencia T-067 de 2025, los lineamientos de transparencia algorítmica deben incluir la evaluación proactiva y sistemática de la publicación del código fuente (total o parcial/condicionada) para sistemas de alto riesgo, garantizando la anonimizarían de datos y la separación de credenciales de acceso a bases de datos sensibles.

4. Aprovechar y Escalar las Experiencias Exitosas:

- Documentar sistemáticamente las lecciones aprendidas de iniciativas nacionales y territoriales. Estas experiencias deben servir de base para escalar y replicar soluciones exitosas en otros contextos y funciones gubernamentales, adaptándolas a las necesidades específicas de cada área.

- Fomentar el intercambio de conocimientos y mejores prácticas entre las entidades para maximizar el impacto de las inversiones en IA, a través de estructuras de colaboración permanente, con el fin de realizar el escalamiento de soluciones de IA probadas a diversas entidades nacionales y territoriales.

La adopción de estas recomendaciones permitirá a Colombia no solo aprovechar los beneficios transformadores de la IA, sino también asegurar que su despliegue en el sector público se realice de manera ética, equitativa y con una sólida base de confianza ciudadana, contribuyendo así a un gobierno más eficiente, transparente y centrado en las personas.

GLOSARIO

NOTA: Este glosario busca generar términos comunes para el entendimiento de este documento, más no establecer definiciones formales para los conceptos aquí consignado.

Algoritmo: Conjunto finito de instrucciones o reglas bien definidas, ordenadas y finitas que permiten realizar una actividad, resolver un problema, mediante pasos sucesivos que no generen dudas a quien deba realizar dicha actividad. (Data Scientist, 2025), (Blockchain España, 2025)

Alfabetización en IA: Conjunto de conocimientos y habilidades necesarios para comprender, evaluar críticamente, utilizar y colaborar eficazmente con sistemas de inteligencia artificial.

Auditoría Algorítmica: Proceso de estudio para la evaluación de un sistema de IA y su proceso de desarrollo, incluyendo el diseño y los datos utilizados para entrenar el sistema. Esto incluye la evaluación de los impactos en materia de precisión, justicia algorítmica, sesgos, discriminación, privacidad, cumplimiento, equidad, seguridad y transparencia. Las auditorías pueden ser realizadas por entidades externas designadas por la organización, o por reguladores, investigadores u otras partes que realizan una auditoría de un sistema por iniciativa propia. (BID, 2022) (Foro de Cooperación en Regulación Digital, 2022)

Caja Negra: Término que describe un sistema de IA cuyos procesos internos o lógica de toma de decisiones son opacos e incomprensibles para los humanos.

Datos Sesgados: Datos que no son representativos de la población o fenómeno que se pretende modelar, lo que puede llevar a resultados discriminatorios o inexactos en los sistemas de IA.

Despliegue de IA: Fase final del ciclo de vida de un sistema de IA, donde el modelo entrenado se integra en un entorno operativo para su uso en aplicaciones del mundo real.

Explicabilidad de la IA (XAI): Capacidad de un sistema de IA a través de un conjunto de procesos y métodos para explicar sus decisiones o comportamientos de modo que permita a los usuarios humanos comprender y confiar en los resultados y los productos creados. (IBM, 2025)

GovTech: Colaboración entre el gobierno y startups, innovadores, o la academia para desarrollar soluciones digitales innovadoras en el sector público.

Huella Ambiental de la IA: Impacto ambiental derivado del consumo de energía y recursos (ej. agua para refrigeración) de los sistemas de IA y sus infraestructuras (ej. centros de datos).

Impacto Dispar: Ocurre cuando una decisión o sistema, aunque aparentemente neutral, produce un efecto negativo desproporcionado en un grupo protegido o vulnerable.

Infraestructura Digital Pública (IDP): Sistemas digitales compartidos, seguros e interoperables que pueden apoyar la prestación inclusiva de servicios públicos y privados en toda la sociedad.

Inteligencia Artificial (IA): Tecnología que permite a las computadoras y máquinas simular el aprendizaje humano, la comprensión, la resolución de problemas, la toma de decisiones, la creatividad y la autonomía. (IBM, 2025 b)

Modelos Fundacionales (Foundation Models): Forma de inteligencia artificial generativa (IA generativa). Ellos generan resultados a partir de una o más entradas (indicaciones) en forma de instrucciones en lenguaje humano. (AWS, 2025)Principios FAIR: Acrónimo de Findable (Encontrables), Accessible (Accesibles), Interoperable (Interoperables) y Reusable (Reutilizables), y corresponden a un conjunto directrices para mejorar la facilidad de búsqueda, la accesibilidad, la interoperabilidad y la reutilización de los recursos digitales. (GO FAIR, 2025)

Sandbox de Datos: Entorno seguro y controlado que permite la experimentación con datos (incluidos los personales) para desarrollar y probar sistemas de IA, garantizando la privacidad y el cumplimiento normativo.

Sesgo de Automatización: Tendencia de las personas a confiar excesivamente en los resultados o recomendaciones de sistemas automatizados, incluso sin un escrutinio crítico.
Sistema de IA de Alto Riesgo: Sistema de IA que, debido a su propósito o contexto de uso, tiene el potencial de causar un daño significativo a la salud, la seguridad, los derechos fundamentales o el estado de derecho.

Tecnologías de Mejora de la Privacidad (PETs): Soluciones digitales que permiten recopilar, procesar, analizar y compartir información protegiendo la confidencialidad y privacidad de los datos.

SIGLAS

AIA: Evaluación de Impacto Algorítmico
(Algorithmic Impact Assessment)

APP: Asociación Público-Privada

APIs: Interfaces de Programación de Aplicaciones
(Application Programming Interfaces)

CAIO: Chief AI Officer
(Director de IA)

EFS: Entidades Fiscalizadoras Superiores

FAST: Fairness, Accountability, Sustainability, Transparency
(Equidad, Rendición de Cuentas, Sostenibilidad, Transparencia)

FPS: Declaración de Posición sobre Equidad
(Fairness Position Statement)

IA: Inteligencia Artificial

IDP: Infraestructura Digital Pública

IND: Infraestructura Nacional de Datos

KPIs: Indicadores Clave de Rendimiento
(Key Performance Indicators)

OCDE: Organización para la Cooperación y el Desarrollo Económicos

ODS: Objetivos de Desarrollo Sostenible

OGD: Datos Abiertos Gubernamentales
(Open Government Data)

PETs: Tecnologías de Mejora de la Privacidad
(Privacy-Enhancing Technologies)

ROI: Retorno de la Inversión
(Return on Investment)

SLMs: Modelos de IA más Pequeños y/o Especializados
(Smaller and/or more Specialized AI Models)

SUM: Support, Underwrite, Motivate
(Apoyar, Garantizar, Motivar)

BIBLIOGRAFÍA

España Digital 2026. (29 de julio de 2025). GobTechLab (Laboratorio ciudadano de innovación tecnológica en la Administración). Obtenido de https://espanadigital.gob.es/lineas-de-actuacion/gobtechlab-laboratorio-ciudadano-de-innovacion-tecnologica-en-la-administracion

Federal Ministry of the Interior. (27 de enero de 2025). Artificial intelligence in the federal administration: AI Opportunity Market and transparency database now open. Obtenido de https://www.bmi.bund.de/SharedDocs/pressemitteilungen/EN/2025/01/maki-pm.html?nn=9384552

Kiteworks. (29 de julio de 2025). ¿Qué es la Ley de la República Digital de Francia? Obtenido de https://www.kiteworks.com/es/glosario-riesgo-cumplimiento/que-es-la-ley-de-la-republica-digital-de-francia/#:~:text=La%20Ley%20de%20la%20Rep%C3%BAblica%20Digital%20de%20Francia%20tambi%C3%A9n%20establece,consumidores%20en%20esta%20era%20digital.

Organisation for Economic Co-operation and Development (OCDE). (11 de Abril de 2025). Governing with Artificial Intelligence. Obtenido de https://www.oecd.org/en/publications/governing-with-artificial-intelligence_26324bc2-en.html

Organización de las Naciones Unidas para la Educación, la Ciencia y la Cultura (UNESCO). (30 de agosto de 2023). Recomendación sobre la ética de la inteligencia artificial. Obtenido de https://www.unesco.org/es/articles/recomendacion-sobre-la-etica-de-la-inteligencia-artificial

The Alan Turing Institute. (2019). Understanding artificial intelligence ethics and safety: A guide for the responsible design and implementation of AI systems in the public sector. Obtenido de https://www.turing.ac.uk/sites/default/files/2019-08/understanding_artificial_intelligence_ethics_and_safety.pdf

The Alan Turing Institute. (2024). AI Accountability in Practice. Obtenido de https://aiethics.turing.ac.uk/modules/accountability/

U.S. General Services Administration. (8 de febrero de 2024). Technology Modernization Fund seeking proposals for Artificial Intelligence projects. Obtenido de https://www.gsa.gov/about-us/newsroom/news-releases/technology-modernization-fund-seeking-proposals-fo-02082024

United Nations (UN). (septiembre de 2024).Governing AI for Humanity.Obtenido de https://www.un.org/sites/un2.un.org/files/governing_ai_for_humanity_final_report_en.pdf

ANEXO TÉCNICO

CASO DE USO ILUSTRATIVO:
SISTEMA DE PRIORIZACIÓN DE CIRUGÍAS CON INTELIGENCIA ARTIFICIAL (SIPRICI) - BASADO EN LA EXPERIENCIA DEL SISTEMA NACIONAL DE SALUD DEL REINO UNIDO (NHS)

NOTA IMPORTANTE

El presente caso de uso es de carácter ilustrativo y no corresponde a un ejercicio ejecutado en Colombia. Se basa en la experiencia real del Sistema Nacional de Salud del Reino Unido (NHS) en la región de Cheshire y Merseyside, donde se implementó exitosamente un sistema de priorización de cirugías con Inteligencia Artificial.

Este anexo tiene como propósito demostrar cómo los principios y valores éticos de la IA se pueden operacionalizar en la práctica, ofreciendo un modelo de referencia que las entidades públicas colombianas podrían adaptar a su contexto institucional, normativo y cultural.

INTRODUCCIÓN

1.1. ¿Por qué este caso de uso?

Cuando hablamos de implementar Inteligencia Artificial en el sector público, es natural preguntarse: ¿cómo se traducen los principios éticos en acciones concretas? Este caso de uso responde precisamente a esa pregunta, mostrando paso a paso cómo un sistema de IA puede diseñarse, desarrollarse y desplegarse manteniendo siempre a las personas en el centro de cada decisión.

El Sistema de Priorización de Cirugías (que llamaremos SIPRICI) representa un ejemplo emblemático de cómo la tecnología puede convertirse en aliada de la equidad y la justicia social. En lugar de reemplazar el criterio médico, esta herramienta lo potencia, ayudando a los profesionales de la salud a tomar decisiones más informadas sobre quién debe operarse primero cuando los recursos son limitados.

1.2. El desafío: cuando la demanda supera la capacidad

Imagine un hospital público con cientos de pacientes esperando una cirugía. Algunos llevan meses en lista de espera; otros acaban de ser diagnosticados. Algunos tienen condiciones que empeoran rápidamente; otros pueden esperar sin mayor riesgo. ¿Cómo decidir quién pasa primero? Tradicionalmente, esta decisión ha dependido del criterio individual de cada médico, generando inconsistencias y, en ocasiones, percepciones de inequidad.

Este escenario no es hipotético. El NHS de Reino Unido enfrentó esta realidad con especial intensidad tras la pandemia de COVID-19, cuando las listas de espera quirúrgicas alcanzaron niveles históricos. La respuesta no fue automatizar la decisión humana, sino crear una herramienta que proporcionara información objetiva y estandarizada para apoyar el juicio clínico.

2. FICHA TÉCNICA DEL CASO

La siguiente ficha presenta las características generales del caso de uso, facilitando su comprensión y eventual adaptación al contexto colombiano.

Categoría	Descripción
Tipo de Caso	Caso Completo End-to-End (abarca todo el ciclo de vida del sistema de IA)
Sector de Aplicación	Salud Pública - Gestión de listas de espera quirúrgicas
Nivel de Madurez Requerido	Intermedio a Avanzado
Tipo de Inteligencia Artificial	Sistema de Soporte a Decisiones basado en Machine Learning Supervisado y Deep Learning
Principios Éticos Prioritarios	Responsabilidad, Equidad, Transparencia, Rendición de Cuentas, Centralidad Humana
Referencia Internacional	NHS Cheshire and Merseyside (Reino Unido) - Sistema C2-Ai PTL
Entidades Aplicables (Colombia)	Hospitales públicos de II-III nivel, Secretarías de Salud, IPS del régimen subsidiado
Alineación con CONPES 4144	Eje 6: Uso y Adopción de IA en entidades públicas para mejorar servicios
Tiempo Estimado de Implementación	12-18 meses (piloto a despliegue completo)

Tabla 1. Ficha técnica del caso de uso SIPRICI – Elaboración Propia

3. PRINCIPIOS ÉTICOS EN ACCIÓN

Este caso de uso demuestra cómo los principios éticos establecidos en la Guía se materializan en decisiones concretas de diseño e implementación. No se trata de cumplir una lista de verificación, sino de integrar estos valores como parte esencial del sistema desde su concepción.

Principio Ético	Aplicación en SIPRICI	Resultado Esperado
Centralidad Humana y responsabilidad	El sistema recomienda, pero el médico decide. Siempre existe la posibilidad de sobrescribir la recomendación con justificación.	Preservación del juicio clínico y la autonomía profesional. El paciente sigue siendo persona, no número.
Transparencia	Cada puntuación incluye explicación de los factores que la determinaron, usando técnicas como SHAP values para explicabilidad.	pacientes pueden entender por qué su cirugía fue programada en determinada fecha.
Equidad y No Discriminación	Métricas de fairness integradas: se monitorea que el sistema no discrimine por estrato socioeconómico, etnias, género o ubicación geográfica	Reducción de brechas de acceso entre diferentes grupos poblacionales y étnicos.
Rendición de Cuentas	Trazabilidad completa: Cada decisión queda registrada con su justificación, permitiendo el desarrollo de Auditorías programadas.	Posibilidad de revisar y explicar cualquier decisión con mecanismos de apelación claros.
Robustez y Seguridad	Validación con panel de expertos, monitoreo de data drift, alertas ante comportamientos anómalos del sistema.	Sistema confiable que funciona consistentemente y detecta sus propias limitaciones.

Tabla 2. Operacionalización de principios éticos en el caso SIPRICI – Elaboración propia contextualizando a Colombia

4. ARQUITECTURA TÉCNICA DEL SISTEMA

4.1. ¿Cómo funciona el sistema?

SIPRICI no es una "caja negra" que toma decisiones incomprensibles. Es un sistema híbrido que combina tres elementos complementarios:

1. **Reglas clínicas basadas en evidencia:** criterios médicos establecidos y validados por la comunidad científica que definen umbrales de urgencia y riesgo. Para este caso se debe entrenar la Inteligencia Artificial con el apoyo de comités médicos y generando casuísticas relacionadas con casos reales presentes en los contextos de salud.

2. **Modelo de Machine Learning supervisado:** algoritmos entrenados con datos históricos que identifican patrones de deterioro y predicen riesgos individualizados, ya que, en el caso de la salud, las enfermedades y los diagnósticos pueden servir como etiquetas que sirven para el procesamiento y los entrenamientos de los modelos.

3. **Capa de optimización:** considera la disponibilidad real de recursos (quirófanos, especialistas, equipos) para proponer fechas factibles. Esto significa tener sistemas interoperables con bases de datos interconectadas que permiten tener un inventario de recursos disponibles y monitoreables en tiempo real.

4.2. Variables consideradas

El sistema considera múltiples dimensiones del paciente y su contexto, reconociendo que la salud es un fenómeno integral que no puede reducirse a un solo indicador. La siguiente tabla presenta las principales categorías de variables utilizadas:

Dimensión	Variables Clave	Fuente de Datos
Clínica	Tipo de patología, severidad, tiempo de evolución, comorbilidades, deterioro funcional (escalas validadas), riesgo anestésico (ASA), etc.	Historia clínica electrónica del paciente
Socioeconómica	Régimen de afiliación, zona de residencia (urbano/rural), barreras de acceso identificadas, estratificación, nivel de estudios, edad, etnia, género	Registros administrativos, Bases de datos de las EPS, SISBEN
Temporal	Tiempo en lista de espera, número de reprogramaciones previas, tiempo desde diagnóstico	Sistema de gestión hospitalaria
Contexto Institucional	Disponibilidad de quirófano, especialista requerido, insumos y equipos necesarios	Sistema de recursos hospitalarios

Tabla 3. Dimensiones y variables del modelo de priorización – Elaboración Propia

4.3. El resultado: Score de Prioridad

El sistema SIPRICI genera un Score de Prioridad (0-100) para cada paciente. Este puntaje no es una decisión automática, sino una herramienta de información que considera ponderaciones de acuerdo a la interpretabilidad del caso:

- **Urgencia clínica (peso aproximado 40%):** qué tan rápido puede deteriorarse el paciente si no se interviene.
- **Equidad de acceso (peso aproximado 30%):** factores que históricamente han generado desigualdad en el acceso.
- **Eficiencia operativa (peso aproximado 20%):** optimización del uso de recursos disponibles.
- **Factores de riesgo (peso aproximado 10%):** consideraciones adicionales de seguridad del paciente.

Nota: *Estas ponderaciones son ilustrativas y deben ser calibradas por equipos clínicos locales según el contexto específico de cada institución y las prioridades de política pública de salud del país.*

5. CICLO DE VIDA:
IMPLEMENTACIÓN END-TO-END

La implementación responsable de un sistema de IA requiere un enfoque sistemático que abarque desde la planificación inicial hasta el monitoreo continuo. Las siguientes fases reflejan la experiencia del NHS y pueden adaptarse al contexto colombiano.

FASE 1: Diseño y Planificación (Meses 1-6)

El éxito de cualquier proyecto de IA comienza mucho antes de desarrollar los sistemas o soluciones tecnológicas. Esta fase establece los cimientos éticos, técnicos e institucionales del sistema, lo que comúnmente se reconoce como la ideación y planificación. Por tal razón, esta guía les permitirá reconocer los siguientes elementos fundamentales.

Actividades clave:

- **Conformación de equipos multidisciplinares:** incluyendo personal clínico, científicos de datos, especialistas en ética, representantes jurídicos, administradores, y, crucialmente, representantes de pacientes.
- **Análisis de Impacto Algorítmico (AIA):** evaluación anticipada de riesgos éticos potenciales, impacto en derechos fundamentales e identificación de grupos vulnerables que requieren salvaguardas especiales.
- **Definición de requisitos y restricciones:** especificaciones técnicas (precisión mínima, latencia máxima), requisitos éticos (métricas de fairness y otros principios éticos a cumplir) y restricciones regulatorias aplicables según las leyes y normas vigentes.
- **Diseño de gobernanza:** establecimiento de comités de supervisión, procedimientos de apelación, comités o juntas para toma de decisiones cruciales y protocolos de auditoría.

FASE 2: Desarrollo y Validación (Meses 6-12)

Esta fase es la encargada de transformar los requisitos y la planeación estratégica en un sistema funcional y robusto, con especial énfasis en garantizar el procesamiento de los datos, la detección de patrones y la mitigación de riesgos y sesgos operacionales.
Gestión del sesgo en datos históricos:

Un desafío crítico es que los datos históricos pueden contener sesgos sistémicos. Si históricamente ciertos grupos poblacionales recibieron atención más rápida, sea porque cuentan con planes de salud preferenciales u otro tipo de servicios que modifiquen el comportamiento, el modelo podría aprender y perpetuar estas inequidades. La mitigación requiere:

- Auditoría de datos para detectar disparidades sistemáticas.
- Re-ponderación de muestras para equilibrar grupos subrepresentados.
- Validación con expertos clínicos de patrones sospechosos.
- Definición de outcomes deseados (equidad clínica) en lugar de replicar outcomes históricos.

Pruebas de equidad:

Dimensión	Métrica	Umbral	Acción Correctiva
Género	Equal Opportunity Difference	< 0.05	Recalibración por género
Estrato socioeconómico	Disparate Impact Ratio	0.80 - 1.25	Re-ponderación de features
Área geográfica	Predictive Parity	< 0.10	Ajuste por disponibilidad
Grupo étnico	Calibration by Group	$R^2 > 0.90$	Sub-modelos o constraints

Tabla 3. Dimensiones y variables del modelo de priorización – Elaboración Propia

FASE 3: Implementación (Meses 12-15)

Para cualquier solución tecnológica y transformación digital soportada con IA, se recomienda hacer un despliegue gradual con pilotos o pruebas de concepto, ya que permiten identificar y corregir problemas antes de afectar a toda la población de pacientes. Además, se convierten en soluciones escalables que facilitan la iteración constante y la gestión de cambios.

Estrategia de despliegue:

- **Piloto controlado:** inicio en 1-2 servicios quirúrgicos de alto volumen, con monitoreo intensivo, por ejemplo: tratamientos cardiovasculares y renales.
- **Expansión gradual:** incorporación progresiva de especialidades, adaptando criterios clínicos específicos. Esto permite hacer una integración progresiva y un análisis de cada una de las especialidades junto al proceso de la construcción de bases de conocimiento específicas y la validación de los comités médicos para cada una.
- **Capacitación diferenciada:** programas adaptados para personal médico, administrativo y liderazgo. La transferencia de conocimiento es crucial desde la estrategia de implementación puesto que todos los actores involucrados deben entender el funcionamiento de las soluciones IA, esto a su vez garantiza la apropiación, uso y transparencia.

Transparencia hacia pacientes:

Un elemento distintivo del caso NHS fue el compromiso con la comunicación proactiva hacia los pacientes, incluyendo notificación personalizada al ingresar a lista de espera, portal de consulta del estado, y procedimientos claros de apelación. Es fundamental encontrar y diseñar canales de atención y comunicación a la ciudadanía.

Las entidades gubernamentales deben aprovechar los distintos medios de comunicación y canales de atención para asegurar la participación ciudadana, la difusión de la información, la responsabilidad colectiva y los ecosistemas de innovación abierta.

FASE 4: Monitoreo y Mejora Continua (Mes 15 en adelante)

Este tipo de sistemas de información apoyados en Inteligencia Artificial no pueden ser estáticos. Requieren constantes cambios, vigilancia permanente y capacidad de evolución basada en evidencia. Por esa razón, se sugiere para cualquiera de los casos de implementación encontrar métricas como las siguientes, las cuales estén relacionadas directamente a los principios éticos sobre el uso de la IA y la centralidad en las necesidades humanas. Este es un ejemplo de ello:

Indicador	Métrica	Meta	Alerta
Equidad	Demographic Parity Difference entre estratos	< 0.05	> 0.10
Eficiencia	Tasa de reprogramación de cirugías	< 10%	> 15%
Calidad Clínica	Net Promoter Score pacientes	< 5%	+20% vs baseline
Satisfacción	Calibration by Group	< 70	< 50
Confianza	Tasa de apelaciones	< 2%	> 5%

Tabla 5. Indicadores clave de desempeño para monitoreo continuo

6. RESULTADOS DOCUMENTADOS:
LA EXPERIENCIA DEL NHS

Para el ejemplo de la solución en Reino Unido sobre la implementación del sistema C2-Ai en la región de Cheshire y Merseyside del NHS ha generado evidencia concreta de los beneficios de un enfoque ético y centrado en el paciente. Los siguientes resultados fueron documentados tras la implementación en tres hospitales (St. Helens and Knowsley, Warrington and Halton, y Liverpool University Hospitals) con una cohorte de 125,000 pacientes:

RESULTADOS CLAVE - NHS CHESHIRE AND MERSEYSIDE

- **Reducción del 8%+** en admisiones de emergencia desde la lista de espera.
- **125 días-cama ahorrados** por cada 1,000 pacientes en lista, reduciendo daños evitables.
- **27.1% de reducción** en pacientes esperando más de 52 semanas, en solo 6 semanas.
- **250,000 pacientes evaluados semanalmente** de forma automatizada en la región.
- **Expansión a 9 hospitales** y replicación en docenas de trusts del NHS en todo Reino Unido.

Fuentes: NHS Cheshire and Merseyside (2023); Health Innovation Network Case Study; Digital Health News Europe; PMC Implementation Report (PMCID: PMC9872449).

6.1. Factores de éxito identificados

El análisis de la experiencia del NHS revela factores críticos que determinaron el éxito de la implementación:

- **Liderazgo clínico comprometido:** los médicos fueron co-creadores del sistema, no usuarios pasivos de una herramienta impuesta. Esto sugiere una integración de los diferentes actores de la sociedad.

- **Control humano preservado:** el sistema siempre opera como soporte a la decisión, nunca como sustituto del juicio clínico. La toma de decisiones son responsabilidad del humano, al igual que la supervisión de las mismas.

- **Integración con infraestructura existente:** baja barrera de implementación al conectarse con sistemas ya en uso. Aprovechamiento de los recursos e interoperabilidad.

- **Enfoque en prevención:** identificar riesgos antes de que se materialicen, no solo gestionar crisis. Es fundamental garantizar la mitigación de posibles riesgos de ocurrencia y la materialización.

7. MARCO DE GOBERNANZA

La gobernanza efectiva es el pilar que sostiene la implementación ética de cualquier sistema de IA. El siguiente marco propone una estructura de responsabilidades que garantiza supervisión, rendición de cuentas y mejora continua.

Instancia	Composición	Responsabilidades
Comité de Gobernanza IA-Salud (Estratégico)	Director Médico, Director Administrativo, Representante de Pacientes, Bioeticista, Jurídico	Aprobar políticas de uso, revisar auditorías, autorizar cambios mayores, gestionar controversias
Comité Técnico de IA (Operacional)	Jefe de Analítica, Data Scientists, Arquitectos de Software, Ciberseguridad	Mantenimiento técnico, implementar mejoras, gestionar incidentes, reportar performance
Comité Clínico de Revisión (Supervisión)	Cirujanos de especialidades, Anestesiólogos, Enfermería, Bioética	Validar coherencia clínica, revisar casos discordantes, actualizar criterios, resolver apelaciones
Punto Focal de Transparencia (Comunicación)	Profesional en comunicaciones con conocimiento de IA y salud	Responder consultas ciudadanas, preparar reportes públicos, gestionar información

Tabla 6. Estructura de gobernanza propuesta para SIPRICI

8. INSTRUMENTOS OPERACIONALES

La operacionalización efectiva de los principios éticos requiere instrumentos concretos que guíen el trabajo cotidiano de los equipos. Si bien el desarrollo detallado de estos instrumentos corresponde a cada entidad según su contexto específico, es fundamental reconocer su importancia y función dentro del ciclo de vida del sistema.

8.1. Instrumentos recomendados

Instrumento	Función	Momento de Uso
Evaluación de Impacto Algorítmico (AIA)	Identificar y documentar sesgos y riesgos éticos potenciales antes del despliegue	Fase de Diseño y planeación
Dataset Factsheet	Documentar procedencia, procesamiento y decisiones sobre datos de entrenamiento	Fase de Desarrollo
Fairness Position Statement	Explicitar y justificar criterios de equidad empleados en el sistema	Fase de Validación

Instrumento	Función	Momento de Uso
Protocolo de Consentimiento Informado	Asegurar que pacientes comprendan cómo se usa su información a través del correcto tratamiento de los datos	Fase de Implementación
Checklist de Auditoría	Guiar revisiones periódicas de performance, equidad y cumplimiento	Fase de Monitoreo
Formulario de Apelación	Canalizar solicitudes de revisión de pacientes de forma estandarizada	Operación continua

Tabla 7. Instrumentos operacionales recomendados

9. SALVAGUARDAS Y GESTIÓN DE RIESGOS

Ningún sistema o solución de IA está exento de los riesgos y problemas de seguridad. La diferencia entre una implementación responsable y una problemática radica en la capacidad de anticipar, detectar y mitigar estos riesgos de manera sistemática. Para el caso de uso se generó la siguiente matriz de riesgos sobre el uso e implementación de Inteligencia Artificial.

Riesgo	Prob	Impacto	Medidas de Mitigación
Generación de predicciones y asignaciones erróneas	Media	Alto	Implementar entrenamientos constantes y garantizar métricas de precisión, F1 Scores, Recall de los modelos de Machine Learning
Sesgo algorítmico contra grupos vulnerables	Media	Alto	Auditorías trimestrales de equidad, métricas obligatorias, representación de poblaciones vulnerables en comité de ética
Dependencia excesiva del sistema (automation bias)	Alta	Alto	Capacitación continua en limitaciones, auditorías de decisiones sobrescritas, requerir siempre justificación médica
Violación de privacidad de datos médicos	Baja	Muy Alto	Cumplimiento Ley 1581/2012, anonimización de datos, auditorías de seguridad semestrales, cifrado end-to-end
Resistencia del personal médico	Media	Medio	Involucrar médicos desde diseño, enfatizar rol de soporte (no reemplazo), mostrar evidencia de mejora
Data drift (cambios en distribución de datos)	Media	Medio	Monitoreo continuo de distribución de variables, reentrenamiento semestral, alertas automáticas

Tabla 8. Matriz de riesgos y medidas de mitigación

10. CONSIDERACIONES PARA LA ADAPTACIÓN AL CONTEXTO COLOMBIANO

Si bien este caso de uso se basa en la experiencia del Reino Unido, su adaptación al contexto colombiano es no solo posible sino deseable. Las siguientes consideraciones pueden guiar este proceso de contextualización:

10.1. Alineación normativa

- **Ley 1581 de 2012:** todo tratamiento de datos personales de salud debe cumplir con los principios de finalidad, libertad, veracidad, transparencia, acceso y circulación restringida, seguridad y confidencialidad.
- **Ley Estatutaria de Salud (Ley 1751 de 2015):** el derecho a la salud es fundamental y las decisiones deben garantizar acceso equitativo.
- **CONPES 4144:** las implementaciones deben alinearse con la Política Nacional de Inteligencia Artificial.
- **Directiva Conjunta 007:** aplicar los estándares de transparencia algorítmica establecidos para el sector público.

10.2. Fuentes de datos disponibles

- **RIPS (Registros Individuales de Prestación de Servicios):** información de atención en salud.
- **BDUA (Base de Datos Única de Afiliados):** régimen de afiliación y características de aseguramiento.
- **SISBEN:** información socioeconómica para identificar determinantes sociales de salud.
- **Historias clínicas electrónicas de las IPS:** información clínica detallada.

10.3. Desafíos específicos del contexto colombiano

- **Fragmentación del sistema de salud:** la existencia de múltiples aseguradores (EPS) y prestadores (IPS) requiere mecanismos de interoperabilidad.
- **Brechas territoriales:** las diferencias entre regiones urbanas y rurales deben reflejarse en el modelo para no perpetuar inequidades.
- **Diversidad poblacional:** Colombia tiene grupos étnicos con necesidades específicas (pueblos indígenas, comunidades afrocolombianas) que deben considerarse.
- **Calidad de datos:** la completitud y consistencia de los registros de salud varía entre instituciones, requiriendo estrategias de validación y limpieza.

11. Lecciones aprendidas y recomendaciones

La experiencia internacional ofrece aprendizajes valiosos para cualquier entidad que considere implementar un sistema similar:

11.1. Lo que funciona

1. **Empezar pequeño y escalar gradualmente:** un piloto bien ejecutado genera confianza y evidencia para la expansión.
2. **Involucrar a la ciudadanía desde el diseño:** los médicos y pacientes deben ser co-creadores, no receptores pasivos.
3. **Transparencia radical:** publicar metodología, métricas y auditorías genera confianza y permite escrutinio constructivo.
4. **Humildad técnica:** reconocer las limitaciones del sistema evita expectativas desmedidas y facilita la aceptación.
5. **Medir el impacto real:** métricas claras de equidad y eficiencia permiten demostrar valor y detectar problemas.

11.2. Lo que debe evitarse

1. **Automatizar sin supervisión:** la IA es una herramienta, no un oráculo. La decisión final siempre debe ser humana.
2. **Ignorar la calidad de los datos:** entrenar con datos sin calidad, genera resultados no deseables y riesgosos. La auditoría de datos es prerrequisito, no opcional.
3. **Subestimar la gestión del cambio:** la tecnología es la parte fácil; cambiar prácticas institucionales requiere tiempo y dedicación.
4. **Desplegar sin mecanismos de apelación:** los pacientes deben tener voz; ignorarla erosiona la legitimidad del sistema.

12. CONCLUSIÓN

Este caso de uso demuestra que es posible implementar sistemas de Inteligencia Artificial en el sector público de manera ética, transparente y centrada en las personas. La experiencia del NHS prueba que, cuando se diseñan correctamente, estas herramientas pueden:

- Mejorar la equidad: al estandarizar criterios y visibilizar sesgos que antes permanecían ocultos.
- Aumentar la eficiencia: al optimizar el uso de recursos escasos sin comprometer la calidad.
- Fortalecer la confianza: al hacer transparentes los criterios de decisión y habilitar mecanismos de rendición de cuentas.
  Potenciar el juicio humano: al proporcionar información objetiva que complementa, pero nunca sustituye, la experiencia clínica.

El camino hacia una IA responsable en el sector público colombiano no requiere reinventar la rueda. Experiencias como la aquí documentada ofrecen aprendizajes valiosos y modelos replicables. Lo que sí requiere es compromiso institucional, inversión en capacidades y, sobre todo, la convicción de que la tecnología debe estar siempre al servicio de las personas, especialmente de aquellas en situación de mayor vulnerabilidad.

REFERENCIAS

C2-Ai. (2022). Hospitals Prioritise Patients Most in Need of Surgery and Increase Capacity with AI. Digital Health News Europe.

Health Innovation Network. (2023). AI to prioritise patients waiting for elective surgery. Case Study.

NHS Cheshire and Merseyside. (2023). Region-wide AI deal to help tackle waiting lists across nine NHS trusts.

NHS England. (2021). NHS's £160 million 'accelerator sites' to tackle waiting lists.

PMC. (2023). Accuracy of a tool to prioritise patients awaiting elective surgery: an implementation report. PMCID: PMC9872449.

The Alan Turing Institute. (2019). Understanding artificial intelligence ethics and safety: A guide for the responsible design and implementation of AI systems in the public sector.

UNESCO. (2023). Recomendación sobre la ética de la inteligencia artificial.

Gobierno de Colombia. (2024). Directiva Conjunta 007 sobre Transparencia Algorítmica. Procuraduría General de la Nación y Defensoría del Pueblo.

Gobierno de
Colombia

