Informe 2

Conceptos fundamentales y uso responsable de la INTELIGENCIA ARTIFICIAL EN EL SECTOR PÚBLICO

CAF BANCO DE DESARROLLO DE AMÉRICA LATINA

/02

Conceptos fundamentales y uso responsable de la

INTELIGENCIA ARTIFICIAL EN EL SECTOR PÚBLICO

CAF BANCO DE DESARROLLO DE AMÉRICA LATINA

Título: **Conceptos fundamentales y uso responsable de la Inteligencia Artificial en el sector público. Informe 2**

Editor: **CAF**

Gerencia de Infraestructura Física y Transformación Digital

Gerente de Infraestructura Física y Transformación Digital, Antonio Silveira.

Elaborado bajo la dirección de Carlos Santiso, anteriormente responsable de la Dirección de Innovación Digital del Estado y Claudia Flores, directora (E) de Transformación Digital, la supervisión de María Isabel Mejía Jaramillo, ejecutiva senior de la Dirección de Transformación Digital y la coordinación editorial de Nathalie Gerbasi, directora (E) de Capacitación.

Este informe estuvo a cargo de María Isabel Vélez, Cristina Gómez Santamaría y Mariutsi Alexandra Osorio Sanabria, junto con Tibisay Sánchez Quintero (para el Capítulo 2), del Centro para la Cuarta Revolución Industrial de Colombia (C4IR.CO), afiliado al Foro Económico Mundial.

Martha Cecilia Rodríguez fue la responsable de la edición de contenidos y corrección editorial.

Las ideas y planteamientos contenidos en la presente edición son de exclusiva responsabilidad de sus autores y no comprometen la posición oficial de CAF.

Diseño gráfico: Good, Comunicación para el Desarrollo Sostenible

Fotografía de portada: iStockphoto

Esta y otras publicaciones sobre el uso e impacto de la inteligencia artificial en el sector público se encuentran en: scioteca.caf.com

INFORME 2

Conceptos fundamentales y uso responsable de la

INTELIGENCIA ARTIFICIAL EN EL SECTOR PÚBLICO

CAF BANCO DE DESARROLLO DE AMÉRICA LATINA
Telefónica
Microsoft
C4IR.CO

Prólogo

Las tecnologías emergentes y en particular la inteligencia artificial (IA) tienen alto potencial disruptivo para resetear las administraciones públicas en la era digital, mejorando la definición de las políticas públicas, la entrega de los servicios a los ciudadanos y la eficiencia interna de las administraciones. El sector público puede potenciar su capacidad para lograr impactos sociales, económicos y ambientales para el bienestar de los ciudadanos, siempre que la IA se implemente en una forma ética y estratégica.

Para lograrlo es preciso comprender los conceptos fundamentales y tener un conocimiento técnico de esta herramienta, saber para qué sirve, cuáles son sus riesgos potenciales, ventajas e inconvenientes y cómo aprovecharla en beneficio de todos.

Es preciso reconocer también que se generan nuevas preocupaciones. Los sistemas de IA deben ser justos, eficientes y eficaces, lo que plantea desafíos en cuatro aspectos críticos de su diseño y operación: el uso efectivo de los datos y la tecnología, las capacidades humanas, la cultura de lo público, y la legitimidad y confianza.

La serie de estudios sobre el uso e impacto de la IA en el sector público en América Latina, que incluye este segundo estudio, tiene precisamente el objetivo de informar este necesario debate, porque las decisiones que tomamos hoy están definiendo nuestro futuro digital. CAF - banco de desarrollo de América Latina, a través de su Dirección de Transformación Digital, promueve la modernización digital para impulsar gobiernos más ágiles, abiertos e innovadores, que se apoyen en las nuevas tecnologías y la inteligencia de datos y fomenten mejoras en la eficiencia de las administraciones y en la calidad de los servicios a los ciudadanos.

En septiembre 2021 CAF lanzó el **Reporte regional "ExperiencIA: datos e IA en el sector público" que aborda el uso estratégico y responsable de esta tecnología en la administración pública,** con el fin de aportar reflexiones y experiencias que permitan a los gobiernos de América Latina responder a los retos que afrontan en un periodo, sin lugar a duda de grandes incertidumbres y, a la vez, decisivo para su desarrollo sostenible futuro. Esta serie de estudios profundiza con mayor detalle algunas de las temáticas clave abordadas en el reporte.

En particular este informe, realizado por un equipo del Centro para la Cuarta Revolución Industrial de Colombia (C4IR.CO), afiliado al Foro Económico Mundial, se enfoca en los conceptos fundamentales y el uso responsable de la Inteligencia Artificial en el sector público, como una primera contribución al debate.

Este conjunto de publicaciones es parte de una agenda más amplia de apoyo de CAF a la apropiación ética de la IA en el sector público liderada por María Isabel Mejía, ejecutiva senior de la Dirección de Transformación Digital, a través de un abanico de instrumentos que incluyen la generación de conocimiento accionable y la asesoraría técnica a gobiernos.

Antonio Silveira
Gerente de Infraestructura Física y Transformación Digital

Reconocimientos

La publicación de este reporte es responsabilidad de la Gerencia de Infraestructura Física y Transformación Digital de CAF, banco de desarrollo de América Latina, a cargo de Antonio Silveira. El documento ha sido elaborado bajo la dirección de Carlos Santiso, anteriormente responsable de la Dirección de Innovación Digital del Estado y Claudia Flores, directora (E) de Transformación Digital, la supervisión de María Isabel Mejía, ejecutiva senior de la Dirección de Transformación Digital y la coordinación editorial de Nathalie Gerbasi, directora (E) de Capacitación.

CAF agradece a las autoras, María Isabel Vélez, Cristina Gómez Santamaría y Mariutsi Alexandra Osorio Sanabria, junto con Tibisay Sánchez Quintero (para el Capítulo 2), del C4IR.CO, y a Martha Rodríguez, por el apoyo editorial.

CAF agradece también a Telefónica, Microsoft y el Centro para la Cuarta Revolución Industrial de Colombia, sus socios estratégicos en esta agenda.

Índice

CUADROS

RECUADROS

FIGURAS

Introducción

La inteligencia artificial (IA) es una tecnología de propósito general que está extendiéndose rápidamente, que se perfecciona a medida que avanza y que soporta y complementa otras tecnologías, dando lugar a un amplio espectro de aplicaciones e innovaciones. Se reconoce en ella un gran potencial para alterar las dinámicas sociales y económicas actuales, resolver o, por el contrario, agudizar muchos de los grandes desafíos ambientales, sociales y económicos de nuestro tiempo.

La difusión, desarrollo y uso creciente de la IA han sido posibles, principalmente, gracias a los enormes volúmenes de datos disponibles, las técnicas analíticas y una capacidad computacional suficiente para procesarlos. De manera particular se destaca la madurez alcanzada por una de sus subcategorías, el denominado aprendizaje de máquina o aprendizaje automático (*machine learning*). Este tipo de IA se apoya en los recursos de computación en la nube y el auge de la economía digital, que incluye nuevas plataformas y mercados para productos basados en datos, así como incentivos económicos para el desarrollo de nuevos productos y mercados (Stone *et al*., 2016).

Si bien los análisis y consideraciones sobre el potencial de esta tecnología se han centrado principalmente en el sector privado, el reconocimiento de las entidades públicas como usuarias de la IA ha venido en aumento. Este incremento ha estado impulsado no solo por la complejidad creciente de las demandas sociales o por el trabajo de investigadores y de organismos internacionales, sino también por los mismos Gobiernos que, como se muestra en el presente informe, han venido incorporando la IA para fortalecer su desempeño en diferentes frentes.

Explorar los beneficios que puede obtener el sector público en este sentido implica la comprensión de ciertos aspectos técnicos. Ese conocimiento permite a los funcionarios, por ejemplo, elegir los métodos o modelos que mejor se ajustan a situaciones específicas o incluso orientar la adquisición de soluciones ya disponibles en el mercado. Con ese propósito, en el presente documento se presentan bases conceptuales relacionadas con la IA y los datos, siendo estos últimos un insumo crítico para la mayoría de los sistemas de IA, particularmente los basados en aprendizaje automático.

En lo que se refiere al uso de la IA, se destaca la contribución potencial al logro de beneficios sociales y económicos con avances, entre otros aspectos, en la prestación de servicios por parte de los Gobiernos. Esta tecnología ofrece la posibilidad de que esos servicios sean más eficientes, equitativos y personalizados. Sin embargo, si bien no cabe duda de las oportunidades y potencialidades que brinda, su desarrollo e implementación también entrañan múltiples desafíos para la sociedad, comenzando por el riesgo de discriminación de grupos e individuos, el uso indebido de los datos o la vulneración del derecho a la privacidad.

Para un uso responsable y un óptimo aprovechamiento de esta tecnología es necesario identificar esos desafíos, comprenderlos en profundidad e identificar formas de mitigar los riesgos que conlleva su explotación. También es preciso reflexionar y abrir diálogos sobre lo que significa e implica un uso responsable y confiable de la IA en un marco ético, en el que participen el sector público —como responsable de reglamentar e incentivar su utilización y, a la vez, como usuario— y la sociedad civil —puesto que sus miembros son beneficiarios y pueden verse afectados por su utilización—, además de involucrar a los expertos en la materia y la academia.

/1

CONCEPTOS FUNDAMENTALES

sobre inteligencia artificial en el **sector público**

Con el fin de ofrecer los conceptos fundamentales para la comprensión de la IA en general y en particular en el sector público, se aborda primero la definición de datos en el ámbito de la IA y las oportunidades que estos plantean para la administración pública, para explorar luego los tipos de datos y uno de sus aspectos críticos: la calidad. Seguidamente, y antes de abordar el concepto de IA, se consideran la gestión de los datos y su cadena de valor, cuya aplicación y comprensión determina su aprovechamiento y uso estratégico, particularmente mediante esta tecnología.

Una vez aclarados esos conceptos fundamentales, se propone una definición de la IA junto con una explicación de las principales clasificaciones utilizadas (general, específica, simbólica y no simbólica), de gran utilidad para ubicar en el contexto de la tecnología términos usados cada vez con más frecuencia, como redes neuronales o aprendizaje profundo. Finalmente, se lleva el foco al sector público, donde se exploran las principales oportunidades que ofrece la IA y se llama la atención sobre los posibles riesgos que conlleva y que requieren especial atención para lograr su uso ético y responsable.

¿QUÉ SON **LOS DATOS?**

Los datos, del latín *datum*, que significa lo que se da o sucede, son cifras o hechos sencillos, discretos y objetivos que representan eventos que ocurren, se estructuran, capturan, cuantifican y transfieren con facilidad (Davenport y Prusak, 2000). Estos hechos, de acuerdo con el contexto, son presentados de tal forma que puedan ser comprendidos, interpretados y comunicados por un ser humano o procesados por una máquina para servir de antecedente en la obtención de una conclusión (Guillén *et al.*, 2015; ODLIS, 2020). Aunque los datos describen solo una parte de lo que sucede, sin ningún juicio o interpretación, son un recurso de gran relevancia que, cuando se combinan con tecnologías digitales, generan oportunidades para promover cambios sociales, políticos y económicos (Nugroho *et al.*, 2015).

Actualmente, el volumen de recolección y uso de datos aumenta de forma acelerada como resultado de la proliferación de dispositivos, sensores y servicios utilizados por personas, organizaciones y Gobiernos. A raíz de este aumento, cada vez más organizaciones reconocen en ellos un activo estratégico que puede utilizarse de forma eficaz para tomar decisiones más efectivas, operar de manera más eficiente, priorizar objetivos, crear productos o evaluar riesgos y procesos (Zafar *et al.*, 2017).

En el caso específico del sector público, se ha reconocido el potencial de los datos, tanto de fuentes internas como externas, para tomar decisiones basadas en la evidencia, impulsar su eficiencia y prestar mejores y nuevos servicios, entre otras cosas (Khtira *et al.*, 2017). Según la OCDE, las oportunidades que generan los datos pueden dividirse en tres áreas: gobernanza anticipatoria, políticas y servicios, y gestión del desempeño (en el Cuadro 1.1 se ofrecen ejemplos para cada una).

Aunque los datos describen solo una parte de lo que sucede, sin ningún juicio o interpretación, son un recurso de gran relevancia que, cuando se combinan con tecnologías digitales, generan oportunidades para promover cambios sociales, políticos y económicos (Nugroho *et al.*, 2015).

Cuadro 1.1
Oportunidades del uso estratégico de los datos

Área:	Oportunidades:	Ejemplos:
Gobernanza anticipatoria	› Pronosticar e identificar proactivamente desarrollos y necesidades futuras a nivel social, económico o ambiental para anticiparse, intervenir e impactar eventos. › Prever situaciones para prepararse ante múltiples resultados probables y definir mejores políticas públicas de acuerdo con los escenarios analizados.	› Sistemas de alerta temprana para activar planes de contingencia para la gestión de riesgos y la prevención de desastres. › Predicción de poblaciones en riesgo de enfermedades. › Análisis de sentimientos en las redes sociales. › Toma de decisiones en tiempo real. › Análisis de riesgos de desplazamiento de mano de obra. › Apoyo en las decisiones de financiación de la investigación y el desarrollo (I+D).
Políticas y servicios	› Apoyar el rediseño de políticas existentes. › Mejorar las soluciones de predicción de políticas innovadoras. › Evaluar el impacto de las políticas públicas para mejorar la respuesta a las necesidades de los ciudadanos. › Fomentar el compromiso con los ciudadanos como cocreadores de valor.	› Combinar datos producidos y recopilados por diferentes entidades públicas para desarrollar políticas integrales. › Mejorar la prestación de servicios transfronterizos y migratorios a partir de acuerdos de cooperación internacional para el intercambio de datos. › Intercambio automatizado de datos sobre registro de empresas. › Analizar datos de iniciativas de tercerización masiva (*crowdsourcing*) en las cuales participen los ciudadanos para cocrear políticas.
Gestión del desempeño	› Apoyar el uso eficiente de los recursos públicos. › Incrementar los recursos públicos. › Aumentar la calidad y evaluación de la gestión pública. › Fomentar la mejora continua.	› Compartir datos de seguridad alimentaria para priorizar la inspección de establecimientos de comida y asegurar la protección de los ciudadanos. › Compartir datos para la detección, análisis y prevención del fraude y la corrupción en el sector público. › Uso del análisis de datos para reducir la carga de los auditores que evalúan el costo de las obras públicas. › Uso de datos para comprender y comparar el desempeño de los sistemas estratégicos de recursos humanos y la gestión para maximizar el impacto del capital humano.

Fuente: Basado en van Ooijen *et al.* (2019) y OCDE (2019b).

Cabe mencionar además que, a partir de las políticas de acceso a la información pública y la adopción del enfoque de gobierno abierto, se ha impulsado la publicación de datos públicos como una estrategia para mejorar la rendición de cuentas, aumentar la participación ciudadana y facilitar su uso por parte de diferentes actores (ciudadanos, organizaciones y entidades públicas) con el fin de generar ideas, conocimiento y servicios (CEPAL, 2018a).

Tipos y calidad de los datos

Los datos creados por las entidades públicas y privadas, los ciudadanos, la academia y los dispositivos digitales pueden ser de diferente tipo y presentar atributos particulares. Estos dependen, por un lado, de su entorno de aplicación, la política o el modelo comercial (OCDE, 2019a) y, por otro lado, de cómo y dónde se almacenan o son accedidos (DAMA International, 2017). Si bien no existe una categorización de tipos de datos estándar o correcta, en el Cuadro 1.2 se presenta un ejemplo de clasificación de los datos del sector público considerando tres aspectos: sus atributos para la gestión y uso (grandes, abiertos y personales), su formato de acceso (estructurados, no estructurados y semiestructurados) y los actores involucrados: Gobiernos, ciudadanos y empresas. Las combinaciones de intercambios son múltiples y pueden ser entre el Gobierno y los ciudadanos (*Government to Citizen* [G2C]); entre el Gobierno y las empresas (*Government to Business* [G2C]); entre diferentes Gobiernos (*Government to Government* [G2C]); entre los ciudadanos (*Citizen to Citizen* [C2C]); entre las empresas (*Business to Business* [B2B]); y entre las empresas y los consumidores (*Business to Consumer* [B2C]).

Cuadro 1.2
Tipos de datos

Atributos para gestión y uso

Macrodatos (*big data*)	Grandes conjuntos de datos que se asocian a características como volumen, variedad, velocidad, viscosidad, volatilidad y veracidad, las cuales se precisan más adelante como dimensiones de los datos. Estas características se conocen como las «V» del Big Data. Este tipo de datos requieren nuevas formas de procesamiento para permitir una mejor toma de decisiones, la generación de conocimientos y la optimización de procesos.	**Ejemplos:** › Contenido web e información de redes sociales. › Datos biométricos, como huellas digitales, escaneo de la retina, reconocimiento facial o genética. › Datos de sensores o medidores que capturan algún evento en particular (velocidad, temperatura, presión, variables meteorológicas, etc.).
Datos abiertos	Conjuntos de datos que se encuentran disponibles para que cualquier persona pueda acceder a ellos, usarlos y compartirlos fácilmente, con el fin de generar valor político, social o económico en procesos de transparencia activa, rendición de cuentas, participación ciudadana, investigación e innovación, entre otros.	**Ejemplos:** › Estadísticas nacionales. › Datos de matrículas de estudiantes. › Datos de calidad del aire y del agua. › Datos de mapas nacionales. › Datos del gasto público. › Datos de propiedad de la tierra. › Datos de registro de empresas.
Datos personales	Datos o información relacionada con una persona y que permiten su identificación. Es decir, que se puede reconocer, directa o indirectamente, por referencia a un identificador como el nombre, la ubicación o uno o más factores de tipo físico, fisiológico, genético, económico o de identidad cultural o social.	**Ejemplos:** › Nombre, número de identificación y edad. › Domicilio, número telefónico. › Estado de salud. › Origen étnico y racial. › Características físicas.

Formato		
Estructurados	Datos que definen atributos, longitud, tamaño y que se almacenan en formato de tabla, hoja de cálculo o en bases de datos relacionales[a].	**Ejemplos:** › Datos que se ingresan en un formulario parametrizado.
No estructurados	Datos sin estructura formal y que se almacenan como documentos de texto, PDF, videos, imágenes y correos electrónicos.	**Ejemplos:** › Documentos de texto o en PDF. › Imágenes, audio o video. › Correos electrónicos.
Semiestructurados	Datos con etiquetas[b], pero sin una estructura formal, como en una base de datos.	**Ejemplos:** › Datos agrupados y representados a través de los formatos HTML, XML o JSON, para etiquetar y organizar documentos e intercambiar datos entre aplicaciones web y servidores.
Actores involucrados		
Government to Citizen **(G2C)**	Datos que se intercambian entre el gobierno y el ciudadano.	**Ejemplos:** › Datos de servicios públicos, transporte o circuitos cerrados de televisión. › Datos capturados a través de sensores que forman parte del internet de las cosas (IoT, por sus siglas en inglés).
Government to Business **(G2B)**	Datos que se intercambian entre el gobierno y las empresas.	**Ejemplos:** › Datos sobre financiación, obligaciones legales, pago de impuestos, licitaciones, etc.
Government to Government **(G2G)**	Datos que se intercambian entre las entidades públicas.	**Ejemplos:** › Datos de servicios de información entre entidades públicas, compras públicas, licitaciones o provisión de servicios centralizados.
Citizen to Citizen **(C2C)**	Datos que se intercambian entre los ciudadanos.	**Ejemplos:** › Datos de redes sociales. › Datos de comunicaciones en correo electrónico, mensajes de texto y voz.
Business to Business **(B2B)**	Datos que se intercambian en el marco de transacciones comerciales entre empresas.	**Ejemplos:** › Datos de la cadena de valor. › Datos financieros y de recursos humanos.
Business to Consumer **(B2C)**	Datos que se intercambian entre las empresas y los consumidores finales en transacciones relativas a productos y servicios al cliente.	**Ejemplos:** › Datos de comercio electrónico. › Datos del cliente. › Datos de servicios financieros. › Datos de servicios de salud.

Notas: [a] Las bases de datos relacionales almacenan y permiten el acceso a puntos de datos relacionados entre sí. [b] Una etiqueta es una palabra clave asignada a un dato almacenado en un repositorio que proporciona información para dar sentido o describir dicho dato (una imagen digital o un clip de vídeo, por ejemplo) y facilitar su recuperación; son, por tanto, un tipo de metadato.

Fuente: Elaboración propia basada en Curry, *et al.* (2012); Naser (2008); OCDE (2019a), y van Ooijen *et al.* (2019).

Para facilitar la generación de valor a partir del procesamiento de los datos y determinar la confiabilidad en la toma de decisiones, es necesario que dichos datos cuenten con características que satisfagan las expectativas y necesidades de información de sus usuarios, las cuales se centran básicamente en la calidad de los datos.

Varias dimensiones permiten determinar el nivel de calidad de los datos a partir de una medición objetiva (completitud o validez) o de acuerdo con un contexto o interpretación subjetiva (usabilidad o fiabilidad). Tales dimensiones se determinan según las percepciones de los usuarios de los datos, el contexto de aplicación y los objetivos que defina cada organización, de manera que no existe un único conjunto de dimensiones a considerar. No obstante, es posible señalar las dimensiones usadas con mayor frecuencia para la gestión de datos (DAMA International, 2017; Myers, 2019):

- **Accesibilidad.** Grado en el que los datos y metadatos[1] se pueden acceder cuando se necesitan, cuánto tiempo se retienen y cómo se controla su acceso.
- **Completitud.** Se refiere a si todos los datos están presentes o si hay suficientes datos.
- **Consistencia.** Grado en que los datos son o no son equivalentes entre los sistemas o en su ubicación de almacenamiento.
- **Facilidad para encontrar.** Grado en que datos y metadatos pueden ser encontrados por la comunidad después de su publicación, mediante herramientas de búsqueda.
- **Exactitud.** Grado en el que los datos representan correctamente entidades de la vida real.
- **Integridad.** Grado en que los conjuntos de datos están lógicamente relacionados.
- **Interoperabilidad.** Datos y metadatos que utilizan estándares abiertos para permitir su intercambio y su reutilización.
- **Oportunidad.** Vigencia de los datos y frecuencia con la que cambian.
- **Validez.** Los valores de los datos son correctos y se ajustan a un estándar en cuanto a formato, tipo de dato, valores posibles o rangos.
- **Reusabilidad.** Capacidad de los datos y metadatos de ser reutilizados al quedar clara su procedencia y las condiciones de reutilización.

1 Los metadatos incluyen información sobre cómo se creó el dato, su composición, sus usos previstos y cómo se ha mantenido a lo largo del tiempo.

Gestión de los datos y cadena de valor

La gestión de datos es el ejercicio que orienta el desarrollo, ejecución y supervisión de los planes, políticas, programas y prácticas que controlan, protegen, entregan y mejoran el valor de los datos como activos de las organizaciones. Su propósito es garantizar que los datos sean adecuados para cubrir las necesidades de información de las organizaciones públicas o privadas.

En el Cuadro 1.3 se describe el alcance y contexto de los componentes de la gestión de datos, cuyo eje central es el gobierno de datos. Los demás componentes son aspectos que pueden implementarse en el sector público en diferentes momentos, de acuerdo con las necesidades y requerimientos particulares de cada organización.

Cuadro 1.3
Componentes de la gestión de datos

Gobierno de los datos	Ejercicio de la autoridad, el control y la toma de decisiones (planificación, seguimiento y ejecución) sobre la gestión de los activos de los datos.
Arquitectura de los datos	Identificación de las necesidades de datos que tiene la entidad para diseñar y mantener los planes maestros que guían la integración de datos, el control de los activos y la alineación de la inversión en datos con la estrategia organizacional.
Modelado de datos y diseño	Proceso iterativo para descubrir, analizar y representar los requerimientos de datos, en un modelo que puede ser conceptual, lógico y físico.
Almacenamiento de datos y operaciones	Supone el diseño, la implementación y el soporte de los datos almacenados para maximizar su valor durante su ciclo de vida.
Seguridad de los datos	Incluye el diseño, la planificación, el desarrollo y ejecución de políticas de seguridad y los procedimientos para proveer la autenticación apropiada, autorización, acceso y auditoría de los datos y activos de información.
Integración de los datos e interoperabilidad	Administra el movimiento y consolidación de los datos dentro de las aplicaciones y organizaciones y entre estas.
Gestión de documentos y contenido	Incluye la planeación, implementación y las actividades de control para la gestión del ciclo de vida de los datos y la información encontrada de cualquier forma o por cualquier medio.
Datos maestros y de referencia	Administra los datos compartidos para conocer objetivos organizacionales, reducir riesgos asociados con la redundancia de datos, asegurar una alta calidad y reducir los costos de la integración de datos.
Data *warehousing*[a] e inteligencia de negocio	Incluye la planificación, implementación y procesos de control para proveer datos que soporten las decisiones y conocimiento de trabajadores involucrados en la realización de reportes, consultas y análisis.
Gestión de metadatos	Incluye la planeación, implementación y actividades de control para permitir el acceso a metadatos[b] integrados de alta calidad.
Calidad de datos	Incluye la planeación, implementación y actividades de control que aplican a los datos técnicas de gestión de calidad, a fin de asegurar que sean aptos para su uso y satisfacer las necesidades de los consumidores de datos.

Notas: [a.] Data *warehousing* es el proceso de construir y usar un repositorio de datos (*data warehouse*) con datos de múltiples fuentes que pueden ser extraídos y analizados para su aprovechamiento en las actividades de las organizaciones. [b.] Los metadatos son descripciones estandarizadas de las características de un conjunto de datos.

Fuente: Tomado de DAMA Internacional (2017).

Para obtener valor de los datos de forma continua y mejorar tanto la toma de decisiones como el desempeño de las entidades del sector público y los efectos a nivel social y económico, es necesario que los gestores de esas entidades conozcan la cadena de valor de los datos. Este es un instrumento de gestión para supervisar y evaluar la secuencia de los procesos que transforman y agregan valor a los datos, desde su creación hasta su uso e impacto. Su visualización facilita el análisis de la planificación de los procesos y los recursos (datos, algoritmos, técnicas, procedimientos y personas) que intervienen en ella.

Considerando que los datos son el insumo para el uso de aplicaciones de IA, se presentan a continuación las etapas de la cadena de valor de los macrodatos (*big data*) (Curry *et al*., 2014):

- **Adquisición de datos.** Comprende la recopilación, filtrado y limpieza de los datos antes de disponibilizarlos en un repositorio o en cualquier otra solución de almacenamiento en la que se pueda realizar su análisis.
- **Análisis de datos.** Se enfoca en hacer que los datos brutos adquiridos sean susceptibles de ser utilizados en la toma de decisiones, así como en el uso específico del dominio. El análisis de datos implica explorar, transformar y modelar datos con el objetivo de resaltar aquellos relevantes y sintetizar y extraer información oculta útil, con alto potencial para atender las necesidades de la organización.
- **Curaduría de datos.** Consiste en mejorar la accesibilidad y la calidad de los datos con el fin de garantizar que estos sean confiables, detectables, accesibles, reutilizables y se ajusten al propósito y necesidades de la organización.
- **Almacenamiento de datos.** Es la gestión de datos que aparecen de forma persistente, de manera que se puedan manejar a mayor escala y se cubran las necesidades de las aplicaciones en cuanto a velocidad de acceso.
- **Uso de datos.** Proceso que cubre las actividades de la organización, se basan en los datos y necesitan acceso, análisis y herramientas para integrar los resultados en la organización.

Para obtener valor de los datos de forma continua y mejorar tanto la toma de decisiones como el desempeño de las entidades del sector público y los efectos a nivel social y económico, es necesario que los gestores de esas entidades conozcan la cadena de valor de los datos. Así podrán supervisar y evaluar la secuencia de los procesos que transforman y agregan valor a los datos, desde su creación hasta su uso e impacto.

Reconocer el valor de los datos dependerá del modo en que cada entidad reconozca las oportunidades generadas a partir de ellos. Si bien no todas las entidades realizan los procesos de la cadena de valor, comprender su alcance permitirá a los gestores públicos definir dónde concentrar los esfuerzos para mejorar el uso, el procesamiento y la explotación de los datos masivos, con el fin de comprender los problemas y necesidades de los ciudadanos, mejorar los servicios, definir y desarrollar políticas, y realizar la gestión y el seguimiento del desempeño e integridad de los Gobiernos.

Figura 1.1
Etapas de la cadena de valor del *big data*

Adquisición	Análisis	Curaduría	Almacenamiento	Uso
Actividades				
› Recopilar › Filtrar › Limpiar	› Explorar › Transformar › Modelar	› Mejorar acceso y calidad	› Proveer, mantener y gestionar los datos de manera escalable	› Explotar datos de salida
Objetivo				
Preparar los datos y disponibilizarlos en un almacén de datos	Resaltar datos relevantes, extraer y sintetizar información útil	Garantizar que son confiables, detectables, accesibles, reutilizables y adecuados	Dar acceso rápido a los datos	Comprender problemas, mejorar servicios, ajustar decisiones

Fuente: Elaboración propia.

COMPRENDER EL ALCANCE DE LA CADENA DE VALOR DE LOS DATOS PERMITIRÁ A LOS GESTORES PÚBLICOS DEFINIR DÓNDE CONCENTRAR LOS ESFUERZOS PARA MEJORAR SU USO, PROCESAMIENTO Y EXPLOTACIÓN, CON EL FIN DE COMPRENDER LOS PROBLEMAS Y NECESIDADES DE LOS CIUDADANOS, MEJORAR LOS SERVICIOS, DEFINIR Y DESARROLLAR POLÍTICAS, Y REALIZAR LA GESTIÓN Y EL SEGUIMIENTO DEL DESEMPEÑO E INTEGRIDAD DE LOS GOBIERNOS.

¿QUÉ ES LA **INTELIGENCIA ARTIFICIAL?**

La inteligencia artificial (IA) es un campo de estudio que se refiere a la creación, a partir del uso de tecnologías digitales, de sistemas capaces de desarrollar tareas para las que se considera que se requiere inteligencia humana. Una definición sencilla, acuñada por la Universidad de Stanford describe la inteligencia artificial como «toda actividad dedicada a hacer las máquinas inteligentes», agregando que la inteligencia «es aquella cualidad que permite a una entidad funcionar apropiadamente y con previsión en su ambiente» (Stone *et al*., 2016).

Más allá de su definición, la IA involucra tecnologías computacionales inspiradas por la forma en que las personas y otros organismos biológicos sienten, aprenden, razonan y toman decisiones (IEEE, 2019).

Existen muchas formas de clasificar la IA, una de ellas es según su capacidad de funcionamiento. En este sentido, se reconocen en la actualidad dos tipos de IA (Fjelland, 2020; OCDE, 2019c):

- **Inteligencia artificial general** (*Artificial General Intelligence* [AGI, por sus siglas en inglés]). Son sistemas que pueden entender y ejecutar tareas generalizadas, tener interacciones y realizar operaciones como las que haría una persona. Esto significa que tiene una mayor capacidad para procesar información y para usarla de forma rápida. Cabe anotar que los desarrollos tecnológicos aún no han alcanzado esta etapa de la IA.

- **Inteligencia Artificial específica** (*Artificial Narrow Intelligence* [ANI, por sus siglas en inglés]). Es aquella que está diseñada para el cumplimiento de una tarea o función concreta, sin poder realizar tareas adicionales o diferentes. Son sistemas no conscientes, que no son sensibles ni están impulsados por emociones. Todas las aplicaciones actuales de IA se ubican dentro de esta categoría, por lo tanto, es a esta a la que se restringen las consideraciones y análisis presentados en este documento.

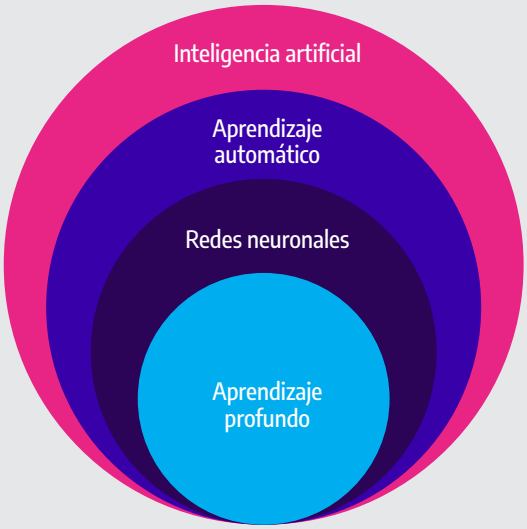
La IA puede dividirse también en **IA simbólica** o basada en reglas y la **IA no simbólica.** La primera de ellas se desarrolla a partir de reglas escritas por humanos para describir un flujo de trabajo y producir resultados, aplicando una secuencia condicional (*if-then* en inglés). También es conocida como «sistemas expertos», dado que se necesita la participación de especialistas con conocimiento de la organización, el proceso y el contexto para el establecimiento de las reglas. Por su relativa simplicidad, este tipo de IA resulta más adecuado para procesos o problemas de baja complejidad, donde participan pocos actores, las acciones a ejecutar son pocas y los cambios no son frecuentes (Berryhill *et al*., 2019).

El uso de IA simbólica puede ser un primer paso para que las entidades públicas comiencen a familiarizarse con las bases de la tecnología. A medida que esos sistemas resulten insuficientes para procesos más complejos, se pueden explorar aplicaciones más sofisticadas basadas en la IA no simbólica.

Esta última se refiere al aprendizaje automático o aprendizaje de máquinas (*machine learning*), consistente en una serie de técnicas que permiten a las máquinas aprender y hacer predicciones a partir de datos históricos, con base en la identificación de patrones, sin que sean necesarias las instrucciones de un humano. Lo más interesante de este tipo de IA es que, en lugar de recibir conocimiento a través de reglas explícitas, los sistemas se entrenan para obtener el conocimiento e inferir las reglas por sí mismos, lo que permite su aplicación en contextos donde los procesos o problemas no alcanzan a estar bien definidos. En años recientes, el aprendizaje automático se ha tornado el enfoque dominante, haciendo que con frecuencia sea tomado como un sinónimo de IA.

Dentro del aprendizaje automático se ubican las redes neuronales y como parte de estas últimas el aprendizaje profundo (*deep learning* [DL, por sus siglas en inglés]). Todos son expresiones o subconjuntos de la IA (Figura 1.2).

Figura 1.2
Clasificación de la IA



Fuente: Elaboración propia adaptada de Needman (2018).

El aprendizaje automático se ocupa de construir programas informáticos que mejoran mecánicamente con la experiencia, de manera que los sistemas de IA detectan patrones y aprenden a hacer predicciones, recomendaciones y prescripciones con base en datos y experiencias, sin necesidad de recibir instrucciones de programación explícitas. Adicionalmente, este tipo de sistemas adaptan su respuesta a nuevos datos y a nuevas experiencias, es decir, continúan aprendiendo y mejorando su desempeño con el tiempo (NITI Aayog, 2018).

Existen cuatro formas principales de aprendizaje automático en los sistemas de IA según su finalidad: supervisado, no supervisado, por refuerzo y por ensamble. El Cuadro 1.4 detalla los distintos tipos de aprendizaje y sus funcionalidades.

Existen cuatro formas principales de aprendizaje automático en los sistemas de IA según su finalidad: supervisado, no supervisado, por refuerzo y por ensamble.

Cuadro 1.4

Resumen de los tipos de aprendizaje automático por finalidad

Aprendizaje no supervisado

¿Qué es?
Un algoritmo que explora patrones y características de un conjunto de datos.

¿Cuándo usarlo?
Cuando se tienen claras las características de los datos y se desea que el algoritmo descubra patrones de comportamiento para etiquetarlos.

¿Cómo funciona?
El algoritmo recibe unos datos no etiquetados, de los que infiere una estructura e identifica agrupaciones posibles de los datos que exhiben un comportamiento similar.

Principales subcategorías y técnicas usadas	Aplicaciones actuales
Agrupación: los datos se agrupan de la mejor manera posible según un criterio determinado. Por ejemplo, *K-Means*, DBSCAN y *Mean-Shift*, entre otros.	Segmentación de mercado, puntos similares en mapas, compresión de imágenes, etiquetado automático de datos, detección de comportamiento anormal.
Reducción de dimensionalidad: se agrupan características específicas en otras de más alto nivel. Por ejemplo, análisis de componentes principales, descomposición en valores singulares, asignación por *Dirichlet* latente (LDA, por sus siglas en inglés), análisis semántico latente (LSA), entre otros.	Sistemas de recomendación, visualizaciones, modelización de tópicos y búsqueda de documentos, análisis de imágenes falsas, gestión de riesgo.
Reglas de asociación: búsqueda de patrones en flujos de pedidos. Por ejemplo, *Apriori*, *Euclat*, *FP-growth*, entre otros.	Predicción de ventas y descuentos, análisis de asociación de bienes para ventas, ubicación de productos en salas de venta, análisis de patrones de navegación en la web.

Aprendizaje supervisado

¿Qué es?
Un algoritmo que descubre las relaciones entre unos datos de entrada y otros de salida, que han sido proporcionados por un supervisor, quien tiene el conocimiento sobre esos datos y los ha etiquetado.

¿Cuándo usarlo?
Cuando se tienen claras las características de los datos, se sabe cómo etiquetarlos, se conoce el tipo de comportamiento que se desea predecir y se desea que el algoritmo haga predicciones de nuevos datos.

¿Cómo funciona?
Un supervisor (normalmente un humano) etiqueta los datos de entrenamiento y define la variable que se desea predecir.

Se alimenta el algoritmo con los datos de entrenamiento para encontrar la relación entre las variables de entrada y de salida. Una vez que el sistema descubre estas relaciones, recibe nuevos datos para hacer predicciones a partir de ellos.

Principales subcategorías y técnicas usadas	Aplicaciones actuales
Clasificación: los datos son divididos con base en una categoría discreta a predecir. Entre ellos, *Naive Bayes*, árboles de decisión, regresión logística, K Vecinos más Cercanos, máquinas de soporte vectorial, redes neuronales.	Filtrado de *spam* en correos electrónicos, detección de idioma, análisis de sentimientos, reconocimiento de escritura a mano, detección de fraudes.
Regresión: los datos son divididos a partir de unos datos continuos a predecir. Por ejemplo, regresiones lineales, polinomiales, regresión de Lasso, redes neuronales, entre otros.	Predicciones de precios de acciones, análisis de demanda y volumen de ventas, diagnósticos médicos.

Aprendizaje por refuerzo

¿Qué es?
Un algoritmo que aprende a desarrollar una tarea por un método simple de recompensas según sus acciones y, en consecuencia, requiere interactuar con un ambiente de prueba.

¿Cuándo usarlo?
Cuando no se tienen conjuntos de datos muy grandes o sólo se puede aprender del ambiente interactuando con él.

¿Cómo funciona?
El algoritmo toma una acción en el ambiente y recibe una recompensa si la acción produjo un aumento en la maximización de las recompensas disponibles.

El algoritmo se optimiza para la mejor serie de acciones, corrigiéndose a sí mismo con el tiempo.

Principales subcategorías y técnicas usadas	Aplicaciones actuales
Algoritmos genéticos, como SARSA, *Q-learning*, *Deep Q-Networks* (DQN), entre otros.	Vehículos autónomos (en combinación con muchas otras técnicas), juegos, aspiradoras robots, transacciones automatizadas (*trading*), gestión de recursos empresariales.

Ensambles

¿Qué es?
Un arreglo de algoritmos clásicos que si operan de manera individual pueden resultar poco eficientes, pero que al ser combinados mejoran su desempeño significativamente.

¿Cuándo usarlo?
Cuando se tienen claras las características de los datos y la calidad es un problema significativo.

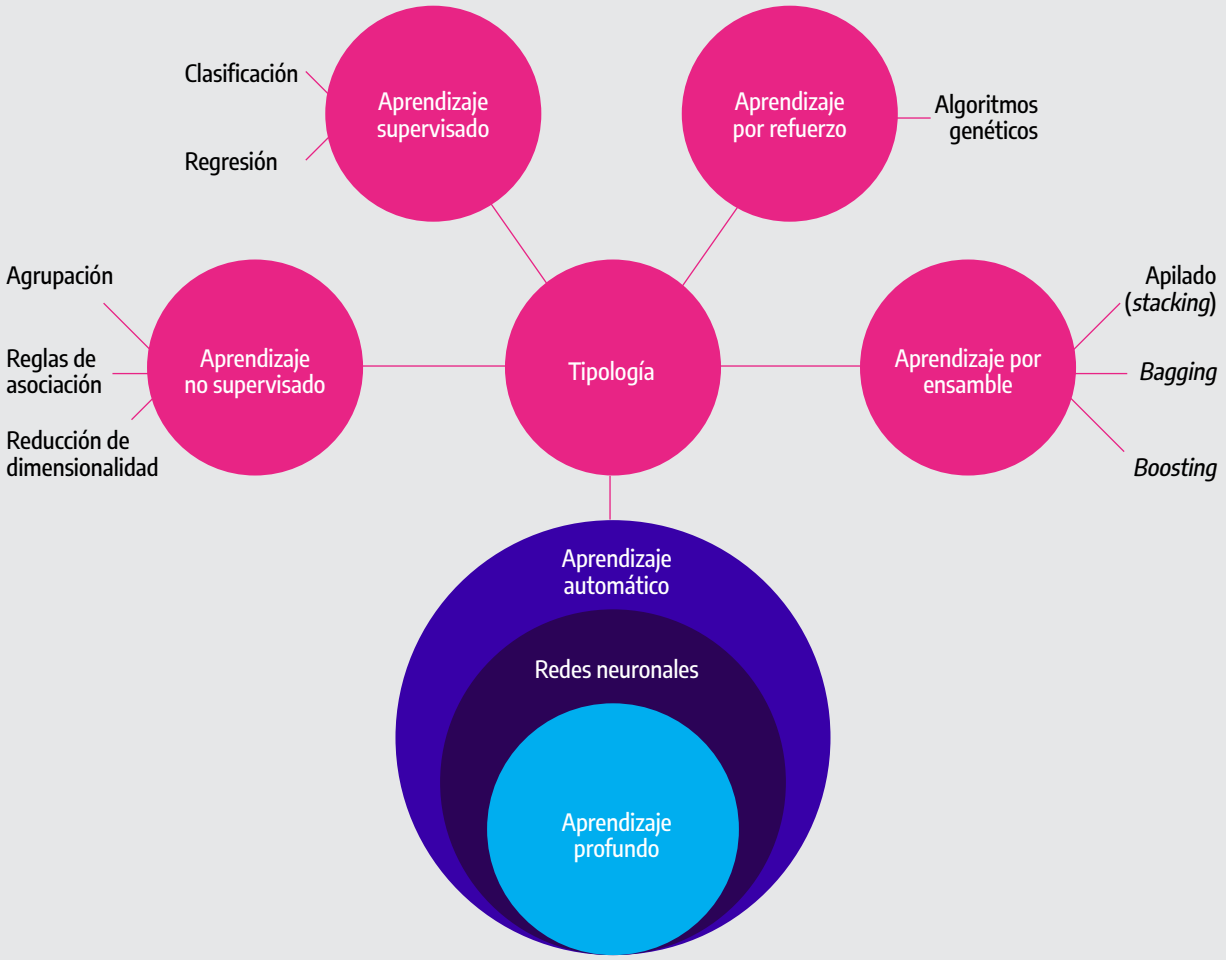
¿Cómo funciona?
Se parte de un conjunto de algoritmos básicos con baja eficiencia y se disponen en un arreglo que los combina buscando maximizar su eficiencia. El arreglo de algoritmos entrega la decisión final.

Principales subcategorías y técnicas usadas	Aplicaciones actuales
Apilado (*Stacking*): diferentes algoritmos colocados en paralelo entregan su salida; todas estas salidas constituyen las entradas para un modelo final que toma la última decisión.	Constituye la base para el *Bagging*.
Bagging (*Bootstrap Aggregating*): funciona como el apilado, pero cada algoritmo es entrenado con diferentes subconjuntos del conjunto de datos de entrenamiento original. *Random Forest* ha sido el más usado.	Aplicaciones de teléfonos inteligentes para detectar rostros en tiempo real.
Boosting: los algoritmos básicos son entrenados uno a uno de manera secuencial, usando subconjuntos de datos de entrenamiento (priorizando aquellos ejemplos de entrenamiento en los que el modelo antecesor falló).	Clasificación en búsquedas web.

Notas: [a] Un algoritmo es un conjunto de reglas o procesos a ser seguidos especialmente por un computador en cálculos u otras operaciones de solución de problemas.

Fuente: Adaptado de NITI Aayog (2018) y Chui y McCathy (2020).

Figura 1.3
Tipos de aprendizaje automático



Fuente: Elaboración propia.

Las redes neuronales artificiales son modelos computacionales, inspirados en las redes neuronales del cerebro humano, que posibilitan el aprendizaje automático. Como su nombre indica, se basan en neuronas o nodos que operan por capas conectadas entre sí y con funciones diferentes. Una vez que se introducen parámetros a una red neuronal, esta entra en una fase de aprendizaje o entrenamiento, que le permite aumentar su precisión y reducir el margen de error.

Las redes neuronales han recibido gran atención en los últimos años, dado que pueden ser utilizadas para todas las finalidades descritas en el Cuadro 1.4 (aprendizaje supervisado, no supervisado, por refuerzo y ensamble). También han generado interés por el auge del aprendizaje profundo, basado en redes neuronales que se construyen con múltiples capas (más de tres) y que, sumadas a las actuales capacidades de procesamiento disponibles, permiten una alta capacidad de aprendizaje en tareas muy complejas. Estas últimas, denominadas redes neuronales profundas, pueden procesar un mayor rango de recursos de datos, requieren menos preprocesamiento de los datos por parte de los humanos y pueden producir mejores resultados que las aproximaciones tradicionales[2] . Esta área ha sido principalmente utilizada para tratar datos no estructurados, como texto, voz, imágenes o video. Una de las particularidades de estos sistemas es que, en muchos casos, sus desarrolladores no logran conocer totalmente la forma en que esta tecnología procesa los datos y toma decisiones, lo que se convierte en un reto para su transparencia y explicabilidad (Stone *et al.*, 2016).

2 En estas redes cada capa de nodos se entrena en un conjunto distinto de características según el resultado de la capa anterior. Cuanto más se avanza en la red neuronal, más complejas son las características que sus nodos pueden reconocer, dado que agregan y recombinan características de la capa anterior.

Principales aplicaciones de IA

Irónicamente, la IA sufre la llamada «paradoja extraña» o «el efecto IA»: a medida que la población se apropia de las nuevas aplicaciones creadas, estas dejan de ser consideradas IA (Stone *et al.*, 2016). Ejemplos de sistemas de IA que hoy pueden no ser considerados como tales incluyen la publicidad personalizada y las sugerencias para hacer nuevos contactos en redes sociales, aplicaciones de movilidad, que permiten trazar la mejor ruta posible de un punto de origen a un punto de destino con información de tráfico en tiempo real, o la identificación automática de correos electrónicos no deseados.

Algunas de las aplicaciones con mayor desarrollo en la actualidad requieren la combinación de varias herramientas de IA e incluso precisan de otras áreas del conocimiento, siendo esencialmente transdisciplinarias y haciendo más compleja esta tecnología. En el Cuadro 1.5 se presentan ejemplos de esas aplicaciones.

Cuadro 1.5
Aplicaciones de IA

Aplicaciones	Definición y áreas que integra
Visión computacional o visión artificial	Área de estudio surgida alrededor de 1950, enfocada en sistemas computacionales con capacidad de entendimiento e interpretación de la información visual, partiendo de imágenes estáticas o de videos (Bebis *et al.*, 2003). Incluye áreas como: › procesamiento de imágenes; › visión robótica; › imaginería médica; › bases de datos de imágenes; › reconocimiento de patrones; › gráficas computacionales; › realidad virtual; Sus aplicaciones incluyen robótica, fabricación, medicina y censado remoto.
Procesamiento del habla y del lenguaje natural	Esta área de la ciencia y la tecnología se encarga de aplicar técnicas computacionales al entendimiento y a la generación de lenguaje humano, sea escrito o hablado. Incluye los siguientes temas de interés: (IEEE, 2015): › reconocimiento de voz; › síntesis de texto a voz; › comprensión del lenguaje hablado; › traducción de voz a voz; › gestión de diálogos hablados; › indexación del habla; › extracción de información; › reconocimiento del hablante y del lenguaje; Las anteriores áreas pueden basarse en aprendizaje automático, con particular énfasis en aprendizaje profundo. Es además transversal a la lingüística, el audio y la filología. Sus aplicaciones a nivel industrial incluyen robots conversacionales, biometría de voz, clasificación de documentos, automatización de resúmenes de textos y asistentes del hogar que reciben instrucciones a partir de la voz.

Vehículos autónomos	Este es un campo bastante amplio que abarca diferentes tipos de vehículos (siendo los más comunes los automóviles y drones) y diferentes niveles de autonomía (se ha estandarizado una clasificación entre 0, siendo este el nivel en que el vehículo debe ser completamente operado por un humano, y 5, nivel donde el vehículo no requiere intervención humana [SAE, 2018]). Incluye áreas como: › sensado remoto (radar, detección y medición de distancia por infrarrojos [LIDAR, del inglés *light detection and ranging*]); › percepción, mapeo y localización; › comunicaciones vehiculares (*vehicle to vehicle* [V2V, por sus siglas en inglés]); › robótica y visión artificial; › interfaces hombre-máquina; › aspectos legales y de seguridad (Yurtsever *et al.*, 2020).

Fuente: Elaboración propia con base en Bebis *et al.* (2003), IEEE (2015) y Yurtsever *et al.* (2020).

Capacidades de la IA

Gracias a capacidades que superan en muchos aspectos el desempeño humano —por ejemplo, para el procesamiento de grandes volúmenes de datos—, la IA ha demostrado su utilidad en varios campos. Esa capacidad permite obtener no solo mejores resultados, sino procesos más eficientes y rápidos, entre ellos, la reducción de retrasos y tiempos de respuesta, la disminución de costos, la gestión de recursos limitados, el desarrollo de tareas repetitivas y rutinarias, el mejoramiento de proyecciones y predicciones, y la ejecución de tareas dispendiosas, como la revisión de miles de documentos e informes para extraer contenido relevante. En el Recuadro 1.1 se destacan cinco capacidades de la IA que pueden ser direccionadas para propósitos específicos, de acuerdo con los intereses y roles de distintos tipos de organización.

Recuadro 1.1

Capacidades transversales de la IA

Automatización
La IA tiene la capacidad de llevar la automatización a otro nivel, permitiendo ejecutar un alto volumen de tareas repetitivas, rutinarias y de optimización de procesos de forma automática y sin necesidad de participación humana.

Precisión
Entrenados adecuadamente, los algoritmos de IA pueden desempeñar ciertas tareas con mayor precisión y exactitud que las personas, principalmente porque su capacidad para procesar grandes volúmenes de datos de manera simultánea y responder rápidamente, excede cualquier capacidad humana.

Detección
En tareas que requieren gran nivel de atención y agudeza, como la detección de errores en sistemas o reportes, o la advertencia de fraudes o robos de información, los algoritmos de la IA pueden llegar a ser de gran utilidad. Además, a diferencia de los humanos, que pueden ver alteradas sus actuaciones por condiciones externas e incluso por sus emociones en determinado momento o por las distintas visiones del mundo, la IA tiene la capacidad de actuar de manera lógica (eso sí, dentro de los parámetros para los que ha sido programada), capturando detalles que pasarían desapercibidos para las personas.

Predicción
La IA constituye un apoyo para la toma de decisiones de diferentes maneras. Una de ellas es reduciendo el tiempo dedicado al procesamiento y análisis de datos que, basado en técnicas de simulación y modelación, pueden informar sobre tendencias y posibles consecuencias de ciertas decisiones. Así mismo, mediante el uso de la IA, es posible elaborar predicciones más exactas, a menor costo y en un mayor número de áreas (diagnósticos médicos, moratoria de créditos, riesgos en seguros, transporte y logística, etc.). Las predicciones se convierten en insumo clave al ser una ayuda determinante para disminuir la incertidumbre que implica la toma de decisiones.

Control y optimización de procesos
La IA hace posible reunir diferentes tipos de datos de diversas fuentes para obtener una mirada integradora que permita identificar posibles errores o ajustes en flujos de información o trabajo y, de acuerdo con ello, tomar medidas correctivas, mejorando la eficiencia de los sistemas.

Fuente: Elaboración propia.

/1

LA IA
EN EL SECTOR PÚBLICO

Si bien las ventajas que representa la adopción de sistemas de IA están mejor documentadas en el sector privado que en el sector público, las capacidades que ofrece esta tecnología pueden ser aprovechadas también por este último en diversos frentes. En este caso se destacan tres áreas que cubren buena parte de las responsabilidades y entidades de gobierno:

1. Mejorar la formulación, ejecución y evaluación de las políticas públicas.
2. Mejorar el diseño y la entrega de servicios a los ciudadanos y las empresas.
3. Mejorar la gestión interna de las instituciones estatales.

Adicionalmente, la tecnología puede ser direccionada para temas específicos, como la salud, el transporte público, la defensa nacional, la educación o la administración de la justicia.

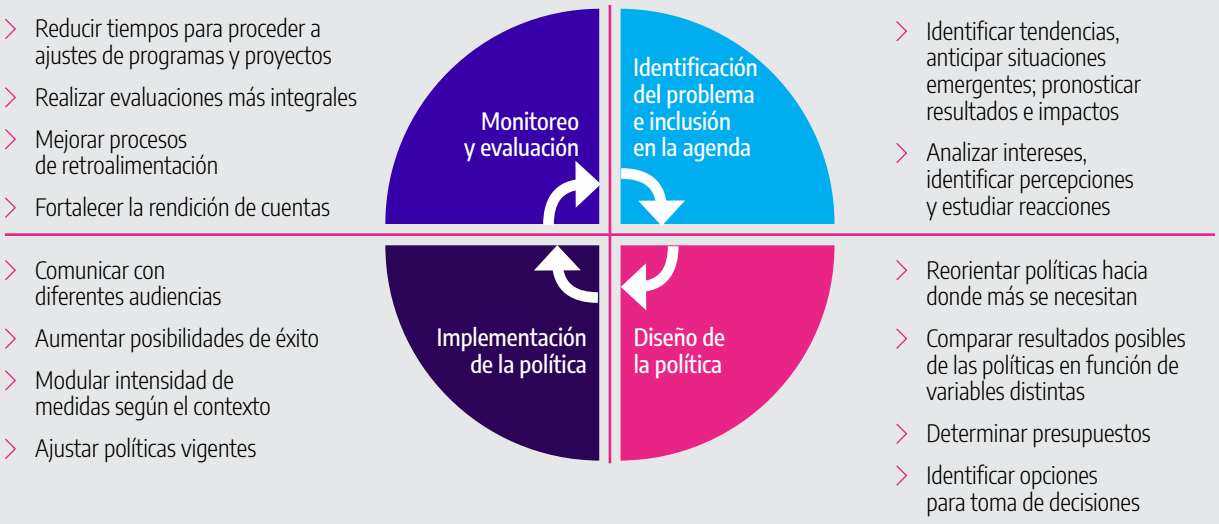
IA para mejorar la formulación, ejecución y evaluación de las políticas públicas

Las oportunidades generadas por la IA para las políticas públicas pueden identificarse más fácilmente observando las etapas que contempla el ciclo de dichas políticas, modelo que, a pesar de sus limitaciones, es el de mayor divulgación e influencia para su análisis y entendimiento. De modo simplificado, el ciclo contempla cuatro etapas: identificación del problema e inclusión en la agenda política, diseño, implementación y evaluación.

EL SECTOR PÚBLICO PUEDE APROVECHAR LAS VENTAJAS QUE REPRESENTA LA ADOPCIÓN DE SISTEMAS DE IA PARA:

1. MEJORAR LA FORMULACIÓN, EJECUCIÓN Y EVALUACIÓN DE LAS POLÍTICAS PÚBLICAS.
2. MEJORAR EL DISEÑO Y LA ENTREGA DE SERVICIOS A LOS CIUDADANOS Y LAS EMPRESAS.
3. MEJORAR LA GESTIÓN INTERNA DE LAS INSTITUCIONES ESTATALES.

Figura 1.4
Oportunidades que ofrece la IA en las diferentes etapas del ciclo de las políticas públicas



Fuente: Elaboración propia.

Identificación del problema e inclusión en la agenda

La IA puede ayudar en esta etapa de dos formas concretas. La primera es identificando tendencias y anticipando situaciones emergentes que merecen la atención e intervención de las entidades públicas, tanto para direccionarlas en el corto plazo como para prevenir complicaciones a futuro. Esta información permite a los Gobiernos estar mejor preparados y actuar de manera más proactiva. Así mismo, permite pronosticar posibles resultados e impactos de las intervenciones, lo que constituye un insumo valioso para decidir la conveniencia o no de una política específica.

La segunda forma es captando y analizando los intereses y preocupaciones de los ciudadanos o diferentes grupos de interés, por ejemplo, los expresados en redes sociales o en sondeos de opinión. Con el uso de técnicas como el procesamiento del lenguaje natural es posible identificar cómo la ciudadanía percibe o interpreta ciertos acontecimientos o cuáles son las reacciones ante medidas o posiciones de la administración pública, lo que puede facilitar un alineamiento entre la agenda de gobierno y las necesidades e intereses de la población.

Diseño de políticas

La IA puede contribuir en el direccionamiento de las políticas hacia individuos, empresas o territorios en condiciones específicas o con necesidades más urgentes. De esa manera, evita el desperdicio de recursos y aumenta la probabilidad de obtener los resultados esperados. Puede, por ejemplo, utilizarse para identificar individuos en mayor riesgo de deserción escolar o poblaciones más vulnerables a ciertas enfermedades y, con base en ello, establecer medidas específicas (Centre for Public Impact, 2017). Así mismo, los algoritmos pueden combinar datos y hacer análisis complejos, logrando integrar los objetivos de políticas e instituciones que actúan en diferentes áreas, lo que permitiría tener intervenciones más integrales y mejor estructuradas para atender problemas complejos (van Ooijen *et al.*, 2019).

Por otro lado, en el momento de elegir entre diferentes alternativas para la intervención pública, la IA permite comparar los posibles resultados de cada una de acuerdo con alteraciones de diferentes variables, así como determinar el presupuesto y los recursos necesarios en cada caso. De esta manera, los tomadores de decisiones podrán elegir las opciones más adecuadas para el contexto y los intereses específicos de las entidades y los Gobiernos.

Implementación

Uno de los ámbitos en los que la IA puede impactar en materia de implementación de políticas es en la comunicación con diferentes audiencias, entendiendo que no todos los segmentos de la población perciben las políticas de la misma manera. Adaptar el alcance, el tipo y las formas de interacción con la ayuda de la IA servirá para aumentar las probabilidades de éxito de las intervenciones (Centre for Public Impact, 2017).

La posibilidad de acompañar el proceso de implementación en tiempo real permite también modular la intensidad de las políticas en respuesta a cambios en el contexto. Por ejemplo, el análisis de los datos con IA facilitará flexibilizar las medidas para mejorar las condiciones de tráfico en las ciudades en ciertas temporadas u horarios, así como intensificar los controles sobre emisiones de ciertos agentes contaminantes en las zonas urbanas. Incluso situaciones inesperadas, como desastres naturales o epidemias, pueden ser incorporados en los algoritmos de IA para ajustar o reorientar las políticas vigentes que así lo ameriten (direccionamiento de ayudas y recursos, atención en centros de salud y alivios fiscales, entre otros).

Evaluación

El mayor impacto de la IA en la etapa de evaluación de las políticas en el corto plazo será la disminución de los tiempos necesarios para llevar a cabo sus actualizaciones o ajustes, proporcionando acceso a información valiosa en tiempo real para tomar decisiones sobre la necesidad de redireccionar, continuar o finalizar programas o proyectos. Esto además lleva a entender la evaluación, no como una actividad que sucede luego de la implementación o al final del ciclo de la política, sino como un proceso que puede desarrollarse de manera continua (Valle-Cruz *et al.*, 2020). Es importante, no obstante, mantener una visión sistémica de las intervenciones, evitando enfocarse en resultados muy específicos o de corto plazo, que no necesariamente dan cuenta de los propósitos generales de una política o de su impacto en el largo plazo.

También la posibilidad de considerar e integrar información de múltiples fuentes para realizar evaluaciones más integrales y completas puede ayudar a tener mejores procesos de retroalimentación, así como a fortalecer la rendición de cuentas de las entidades responsables de las políticas.

La IA puede impactar en la implementación de políticas, por ejemplo, en situaciones inesperadas, como desastres naturales o epidemias, para ajustar o reorientar las políticas vigentes que así lo ameriten (direccionamiento de ayudas y recursos, atención en centros de salud y alivios fiscales, entre otros).

La IA para mejorar el diseño y la entrega de servicios a los ciudadanos y las empresas

Como ya se señaló, la IA tiene una gran utilidad para identificar intereses, preocupaciones y percepciones de diferentes actores con el propósito de priorizar problemas a ser incluidos en la agenda de gobierno. No obstante, la oportunidad de generar análisis y herramientas que favorezcan un mejor entendimiento del comportamiento ciudadano o de ciertos grupos se extiende más allá de la definición y priorización de los problemas públicos hacia áreas que tienen que ver con la experiencia del acceso a servicios como el transporte, la salud, la educación, la seguridad o la justicia, o la manera en que grupos o ciudadanos se ven afectados por la actuación de las entidades públicas.

Gracias a la alta penetración de los dispositivos móviles, las plataformas sociales y los medios de comunicación, donde individuos y organizaciones expresan con frecuencia sus posiciones, las autoridades tienen la posibilidad de obtener información relevante sobre situaciones concretas que experimentan los ciudadanos en su vida diaria y que pueden afectar su bienestar, o revelar demandas específicas de ciertos actores. Esta información es fundamental para adaptar el diseño de los servicios a las maneras de pensar, sentir y actuar de las personas u organizaciones de acuerdo con sus realidades, lo que resulta muy útil, por ejemplo, para proveer servicios y asistencia a minorías o poblaciones específicas. En este sentido, el conocimiento de los usuarios puede conducir al diseño de servicios personalizados, que permitan, tanto a ellos como a las entidades proveedoras, ahorrar tiempo y recursos, con la definición de las rutas óptimas de atención a partir de técnicas de IA, como las redes neuronales.

Por otro lado, sistemas de IA, como los robots conversacionales, pueden hacer más eficientes las interacciones con los ciudadanos, dando respuestas rápidas a cuestiones o solicitudes puntuales en sus versiones más básicas, o más sofisticadas, las cuales incluyen el aprendizaje automático, abordando interacciones más complejas. Esta capacidad de ofrecer orientación y respuestas de manera ágil posibilita mejorar los niveles de satisfacción de los ciudadanos con el desempeño de las entidades públicas.

Cabe destacar aquí que es fundamental que los ciudadanos conozcan el uso que se está haciendo de los datos que ellos mismos producen en el diseño, desempeño y mejoramiento de los servicios públicos a los que tienen acceso. Solo así se podrá generar la confianza necesaria entre la población y las entidades públicas para que el intercambio de información se mantenga en el tiempo, facilitando el seguimiento del uso que se hace de los servicios públicos y, con ello, robusteciendo los sistemas de IA y sus procesos de aprendizaje, de manera que respondan mejor a situaciones específicas, necesidades y expectativas (van Ooijen *et al.*, 2019).

La IA para mejorar la gestión interna de las instituciones estatales

Las oportunidades ofrecidas por la IA para las etapas del ciclo de política o para el diseño y prestación de servicios públicos se extienden también al funcionamiento y gestión de las entidades del Estado. Este apoyo es posible no solo porque facilitan el cumplimiento de sus objetivos y responsabilidades, sino porque permiten hacerlo incrementando los niveles de eficiencia y productividad. Disponer de técnicas y herramientas que conjuguen las capacidades atribuidas a la IA, como la predicción, la automatización, la optimización o el control, representa un potencial enorme para mejorar el desempeño de cualquier organización, sea pública o privada.

Los sistemas de IA pueden apoyar la asignación y gestión de recursos financieros, ayudando a identificar y prevenir fraudes y el desvío o ineficiencias en la asignación y uso de dinero público, entre otros problemas. Así mismo, el procesamiento de solicitudes, requerimientos, análisis o decisiones puede hacerse de manera más rápida, ahorrando tiempo a las entidades y sus usuarios. En el caso de activos o infraestructura, es posible el mantenimiento preventivo, la corrección de fallas o la programación de su uso de acuerdo con la demanda mediante aplicaciones de IA, logrando una utilización más eficiente (van Ooijen *et al*., 2019).

Por otro lado, la IA puede contribuir tanto a la generación y actualización de la regulación como al refuerzo de su cumplimiento. En el primer caso, la analítica de datos o las aplicaciones de aprendizaje automático ayudan a identificar vacíos o contradicciones normativas que deban resolverse, así como los aspectos más críticos en los que se debe enfocar la inspección o la aplicación de sanciones. La construcción de modelos predictivos permite detectar tempranamente fraudes u otras violaciones a la regulación. Recurriendo a estas aplicaciones, los esfuerzos de entidades regulatorias y de control pueden estar mejor dirigidos, priorizando aquellas situaciones donde es más pertinente y necesaria su intervención y dejando en un segundo plano actuaciones que no revisten gran relevancia o impacto.

De manera particular cabe resaltar la posibilidad de aumentar sustancialmente la agregación de valor del trabajo público. Dada su capacidad de automatización, la IA puede asumir actividades y juicios humanos repetitivos y rutinarios, facultando a los funcionarios que se ocupan de este tipo de tareas a dedicar su tiempo, conocimientos y capacidades a explorar actividades de mayor valor y complejidad, donde la creatividad, el criterio, las habilidades emocionales y la perspectiva humana son necesarios.

Una forma de avanzar en la automatización es dividiendo los procesos en etapas o actividades, para identificar aquellas que puedan ser automatizadas. Con ello, se dejarán en manos de las personas las tareas restantes o incluso la supervisión del trabajo automatizado. Por ejemplo, en el proceso de atención al ciudadano, es posible automatizar la recepción, clasificación y redireccionamiento de solicitudes, las cuales, dependiendo de su nivel de complejidad, pueden ser respondidas de manera automática o asignadas al responsable directo. Igualmente, es posible automatizar la entrada de datos a un sistema mediante el reconocimiento automático de la escritura a mano, el reconocimiento de voz o el procesamiento del lenguaje natural para aumentar la rapidez y asertividad de las respuestas a las demandas ciudadanas.

Otras capacidades de la IA, como la precisión y la predicción, sirven para complementar habilidades humanas y así obtener mejores resultados en una tarea o proceso, como, por ejemplo, en los diagnósticos médicos, la recomendación de tratamientos o la definición de cursos de acción para atender una emergencia.

Si se gestionan de la manera adecuada, estos cambios en el alcance del trabajo permitirán aprovechar otras capacidades humanas que enriquezcan los procesos públicos de innovación, beneficiando a empleados, entidades y Gobiernos en general.

El Cuadro 1.6 presenta el marco propuesto por Eggers *et al.* (2017) para ayudar a entidades públicas a identificar oportunidades para implementar sistemas de IA, considerando las condiciones de que disponen para el desarrollo de sus funciones. Estas oportunidades de aplicar la IA pueden resultar viables, valiosas o vitales de acuerdo con su contribución al desarrollo de procesos y tareas específicas.

Cuadro 1.6

Marco para la evaluación de oportunidades de implementación de la IA en agencias gubernamentales

Oportunidad	Situación actual	Ejemplos
Viable	Habilidad baja a moderada; se necesita cierta percepción humana para completar todo o parte del trabajo.	Procesamiento de formularios, servicios al cliente de primer nivel, operaciones de almacén, clasificación de correo electrónico, gestión de archivos.
	Tareas que requieren grandes conjuntos de datos.	Consejo sobre inversiones, diagnósticos médicos, monitoreo de fraude usando aprendizaje automático.
	Tareas basadas en reglas o en la experiencia.	Programación de operaciones de mantenimiento, organización de horarios para el transporte público, cumplimiento de la regulación.
Valiosa	Empleados altamente calificados pueden dedicarse a actividades de mayor valor.	Preparación de informes presupuestales, dirección o realización de pilotos, tabulación de datos fiscales, seguimiento de gastos.
	Alto costo del recurso humano.	Gestión de los seguros de salud: determinación de elegibilidad, respuesta a preguntas de usuarios, detección de amenazas de seguridad.
	Habilidades escasas; el mejoramiento del desempeño tiene un alto valor.	Diagnóstico médico, vigilancia aérea, predicción de delitos.
Vital	El desempeño estándar de la industria requiere tecnologías cognitivas.	Licencias de conducción o renovación del pasaporte en línea, defensa cibernética, investigación criminal, predicción del clima.
	El trabajo humano es insuficiente para ampliar de escala la actividad o el servicio.	Detección del fraude, emisión de patentes y protección de derechos de propiedad intelectual, atención de desastres, minería de textos.
	Grandes retrasos; actividades que requieren el uso de IA.	Análisis de reportes históricos, aplicaciones de patentes, retrasos en atención a reclamaciones, vehículos autónomos y drones.

Fuente: Eggers *et al.* (2017).

Si bien las capacidades de la IA presentan un gran potencial para el sector público, ya sea para la optimización del proceso de la política, la provisión de servicios, la gestión de las entidades o el logro de objetivos en áreas de intervención específicas también es cierto que existe una alta incertidumbre respecto a los efectos y la evolución que pueden tener las aplicaciones de la tecnología en diferentes ámbitos. Hasta ahora, las experiencias con el desarrollo y uso de la IA, tanto a nivel público como privado, han llamado la atención sobre algunos aspectos que pueden conducir a resultados indeseados o problemáticos para los ciudadanos, las organizaciones, los Gobiernos y la sociedad en general. De llegar a materializarse, pueden deteriorar la confianza de los diferentes actores en la tecnología y, en consecuencia, minar su legitimidad para el uso público, asuntos que se abordan a continuación.

/1

POTENCIALES RIESGOS DE LA IA EN EL SECTOR PÚBLICO

Las expectativas generadas en los últimos años alrededor de la IA, impulsadas principalmente por las inversiones y apuestas del sector privado, han supuesto un reto para los Gobiernos. Por un lado, estos deben generar rápidamente políticas y condiciones para estimular la innovación, manteniendo límites éticos; por el otro, hacen esfuerzos para comprender la tecnología y no quedarse atrás en su adopción.

Las grandes compañías, usando de manera estratégica la inmensa cantidad de datos que obtienen de sus clientes, están cada vez mejor informadas para tomar decisiones y agilizar sus procesos, mejorar la calidad y el direccionamiento de sus productos y servicios. Quedarse atrás en este sentido representa para el sector público el riesgo de ser superado en su capacidad de actuar de manera estratégica y atender de manera rápida y eficiente las necesidades sociales. A medida que las aplicaciones de IA se expanden hacia áreas sensibles, como la defensa nacional, la ciberseguridad o la bioingeniería, la actuación de los Gobiernos cobra mayor relevancia (Centre for Public Impact, 2017). No obstante, mantener la prudencia ante las promesas de la IA puede ser la posición más sensata y productiva que adopten los Gobiernos en su intento de lograr mayores beneficios (MIRI, s. f.).

Quizá el mayor reto que enfrentan los Gobiernos ante la IA es encontrar un equilibrio entre la explotación de una tecnología que le permitirá dar un gran salto en los frentes anteriormente descritos y el establecimiento de límites a la misma para garantizar el bienestar social. Para avanzar en esta dirección, es necesario que el sector público preste especial atención a los aspectos que se relacionan a continuación.

Privacidad y confidencialidad

La privacidad pasa por el derecho que tienen los individuos a establecer límites sobre la información que de ellos se divulga, a no ser observados y a que se mantenga su confidencialidad. En este sentido, el creciente volumen de datos generado por unos y capturado por otros diariamente representa un riesgo para mantener ese derecho, sobre todo porque gran parte de esa captura está sucediendo a través de dispositivos y procedimientos que no son conocidos o autorizados por los propietarios de los datos, como cámaras y sensores en lugares públicos, aplicaciones para teléfonos móviles o redes sociales. No saber cuándo está ocurriendo la captura y tratamiento de datos, ni por parte de quién y mucho menos para qué, despoja a los ciudadanos y las organizaciones de la posibilidad de adoptar una posición al respecto.

Siendo los datos personales y colectivos el pilar fundamental de la IA, su análisis, divulgación, uso y reutilización pueden generar resultados o conclusiones que los propietarios de la información no quieren que sean divulgados o empleados para ciertos propósitos. Por ejemplo, a partir de la triangulación de información de una misma persona en diferentes conjuntos de datos es posible

identificar su identidad, su estado de salud o sus tendencias políticas. Estas informaciones, al ser integradas en algoritmos para la toma de decisiones automatizada, pueden llevar a su identificación y, a partir de allí, conducir a situaciones de discriminación, o exponer comportamientos de personas o grupos sin su autorización.

Para generar confianza, tanto en las instituciones públicas como en la tecnología misma, es clave que las personas sientan que no pierden su derecho a la privacidad. Por esta razón, es fundamental que los Gobiernos garanticen que los sistemas de IA diseñados e implementados se ajustan a las normas y regulaciones vigentes sobre protección de datos en cada país. Así mismo, es preciso que los ciudadanos conozcan sus derechos, la regulación aplicable y cómo pueden realizar cualquier reclamación en caso de considerarlo necesario.

Transparencia y explicabilidad

El tratamiento de grandes volúmenes de datos que realizan los algoritmos de IA resulta complejo y difícil de entender para la mente humana. Cuanto más sofisticado es el modelo usado, menores son las posibilidades de participación y entendimiento. Algoritmos de aprendizaje automático, como las redes bayesianas[3] y los árboles de decisión[4], son relativamente más comprensibles que, por ejemplo, las redes neuronales profundas o los algoritmos genéticos[5], cuyo proceso de aprendizaje es tan autónomo que resulta realmente difícil determinar qué parámetros utilizan para tomar decisiones (Bostrom & Yudkowsky, 2015). Esta dificultad para entender cómo y porqué un sistema de IA genera un resultado o toma una decisión los convierte en una «caja negra», cuyo contenido es desconocido hasta para sus programadores (Stone *et al.*, 2016).

Cuando una decisión tomada o informada por un sistema de IA tiene implicaciones en la vida de las personas o de grupos —por ejemplo, autorizando la libertad condicional de un preso, asignando subsidios públicos o realizando diagnósticos médicos—, la necesidad de entender las razones que la generaron y, con ello, permitir su refutación en caso de considerarse equivocada o injusta, exige que el proceso de toma de decisión sea transparente y explicable, características cruciales para garantizar la confianza pública (Brookfield Institute, 2018).

En todo caso, no basta que las entidades usuarias divulguen los algoritmos utilizados y su forma operativa. Es necesario que esa información sea entendible para todos los actores involucrados (programadores o diseñadores, órganos de regulación, usuarios finales o afectados). No menos importante es que estos actores sean informados de antemano del uso de sistemas de IA, su propósito, sus capacidades y limitaciones. Cuando se trate de algoritmos demasiado complejos, la posibilidad de explicarlos puede reforzarse con mecanismos de trazabilidad y auditoría y la divulgación de sus alcances (Berryhill *et al.*, 2019).

3 Las redes bayesianas son modelos gráficos probabilísticos que permiten modelar un fenómeno mediante un conjunto de variables y las relaciones de dependencia entre ellas. Dado este modelo, se puede hacer inferencia bayesiana, es decir, estimar la probabilidad posterior de las variables no conocidas a partir de las variables conocidas. Estos modelos tienen diversas aplicaciones, entre ellas, la clasificación, la predicción y el diagnóstico. Además, pueden dar información relevante sobre cómo se relacionan las variables del dominio, las cuales pueden ser interpretadas en ocasiones como relaciones de causa-efecto (Sucar, 2015).

4 Los árboles de decisión son un tipo de aprendizaje automático supervisado donde los datos se dividen continuamente de acuerdo con un parámetro determinado. El árbol se estructura con base en nodos de decisión y hojas. Las hojas son las decisiones o los resultados finales y los nodos de decisión es donde se dividen los datos al ser aplicado el parámetro (Decision Trees for Classification, s. f.).

5 Los algoritmos genéticos son algoritmos de búsqueda que actúan sobre una población de posibles soluciones. Se basan en la mecánica de la genética y la selección de poblaciones. Las posibles soluciones están codificadas como «genes». Se pueden producir nuevas soluciones «mutando» a los miembros de la población actual y «acoplando» dos soluciones para formar una nueva solución. Las mejores soluciones se seleccionan para reproducirse y mutar y las peores se descartan. Son métodos de búsqueda probabilísticos; esto significa que los estados que exploran no están determinados únicamente por las propiedades de los problemas (Shapiro, 2001).

Una alternativa para superar el riesgo que representa la falta de «explicabilidad» de los algoritmos, particularmente del aprendizaje automático, es la inteligencia artificial explicable o XAI (*eXplainable AI*) (ver el Recuadro 1.2).

Recuadro 1.2

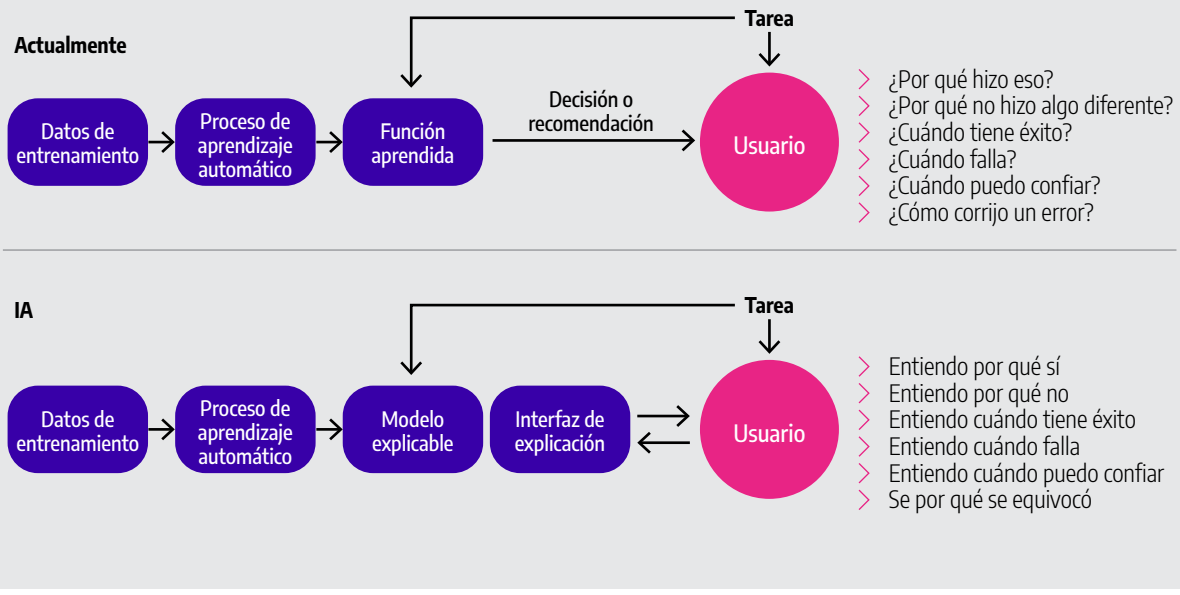
La IA explicable

La IA explicable o XAI (*eXplainable AI*) propone una serie de técnicas y herramientas para ayudar a entender e interpretar los resultados de los modelos, permitiendo mejorar también su desempeño.

«El conjunto de técnicas de aprendizaje automático 1) produce modelos más explicables mientras mantienen un alto nivel de rendimiento del aprendizaje (p. ej., precisión de la predicción), y 2) permite a los humanos comprender, confiar adecuadamente y gestionar eficazmente el surgimiento de sistemas de IA» (Barredo-Arrieta *et al.*, 2020).

Figura 1

Concepto de IA explicable



Fuente: DARPA (s. f.).

De acuerdo con la Figura 1, la IA explicable genera nuevos modelos de aprendizaje de máquinas que, además de ser susceptibles de interpretación, ofrecen una explicación de sus resultados o decisiones, incluyendo los parámetros que condujeron a ella y admitiendo ajustes, si fueran necesarios (Berryhill *et al.*, 2019).

Fuente: Barredo-Arrieta *et al.* (2020); Berryhill *et al.* (2019) y DARPA (s. f.).

Inclusión, equidad o representatividad

Los algoritmos de IA pueden arrojar resultados inexactos o erróneos con el riesgo de que conduzcan a la discriminación o la exclusión. Esto puede suceder por distintas vías. Una de ellas es porque los datos con los que se ha entrenado el algoritmo (enseñando patrones, tendencias o respuestas correctas) presentan sesgos[6], es decir, excluyen información importante, reflejan prejuicios sociales que se introducen durante su recolección o etiquetado o no son representativos y, por lo tanto, no son adecuados para hacer generalizaciones. Cuando este tipo de datos se introducen en un algoritmo, sus limitaciones se extienden a todo el ciclo de vida del sistema de IA, haciendo que sus predicciones o decisiones mantengan o refuercen esos sesgos y así amplifiquen

6 Consultar la sección «El sesgo en la inteligencia artificial», en el Capítulo 2, para información complementaria.

disparidades o situaciones de exclusión existentes en el mundo real. Como consecuencia, ciertos grupos o individuos pueden ver afectado su acceso a recursos o servicios, el nivel de vigilancia al que están expuestos, la forma como son tratados por el Gobierno e incluso su capacidad para ser tomados en cuenta en un entorno que enfatiza las tecnologías (Brookfield Institute, 2018). En este último caso, los resultados de algunos sistemas pueden reafirmar o profundizar la brecha digital existente.

Otra forma de inducir resultados discriminatorios o excluyentes en sistemas de IA es a partir del diseño de los algoritmos. Este diseño es realizado por humanos con sesgos o prejuicios, conscientes e inconscientes, que terminan siendo integrados en los parámetros que se definen para alimentar el sistema. En este caso, el entendimiento de ciertas realidades o la visión de quien diseña el algoritmo respecto a las necesidades o características de personas o comunidades pueden llevar a enfatizar o priorizar algunas variables sobre otras (por ejemplo, el color de la piel, el nivel de educación, el estrato socioeconómico o el lugar de residencia). Así mismo, la interpretación que se hace de los resultados de un sistema de IA dependerá en ocasiones de las ideas, paradigmas y juicios preconcebidos de quien los utiliza. Paradójicamente, la IA puede operar en la dirección contraria, ayudando a disminuir la influencia de sesgos del mundo real en la toma de decisiones, en la medida que procesa unos datos de entrada exactamente como se ha programado, reduciendo así los ruidos y las inconsistencias que tienden a estar presentes en las decisiones humanas.

Teniendo en cuenta la manera en que ideas y valores pueden influenciar los sistemas de IA, uno de los mayores retos para los Gobiernos es promover acuerdos sociales alrededor de esos sistemas, su entendimiento, alcance o relevancia para poblaciones específicas, así como sobre lo que se espera de la IA.

Seguridad

Técnicamente hablando, los sistemas de IA se desarrollan usando *software* y *hardware*, los cuales no siempre funcionan correctamente, pudiendo causar fallas en esos sistemas. Más aún, los errores generados en los algoritmos por los datos o modelos sesgados que se mencionaron antes, la falta de un correcto mantenimiento, su uso en situaciones no deseadas, la violación de la privacidad o el aprendizaje de comportamientos no seguros una vez han comenzado a operar, son otros factores que pueden comprometer la seguridad de los usuarios y los sistemas mismos (Brookfield Institute, 2018).

A medida que los algoritmos de IA incrementan la eficiencia y capacidad de muchos procesos, también introducen nuevas vulnerabilidades. A diferencia de los actuales modelos de ciberseguridad, que se enfocan en el control de accesos no autorizados, las debilidades de la IA no están solo en los puntos de entrada al sistema, sino en sus interacciones con el mundo real. En ese sentido, pueden ser atacados manipulando su capacidad de aprender o actuar sobre lo aprendido (Elish y Watkins, 2019). Tanto los algoritmos de IA como el *software* y *hardware* que están en su base son propensos a errores y susceptibles a la manipulación. Los fallos de este tipo pueden representar serios riesgos para los individuos, organizaciones y países.

En conclusión, encarar los riesgos generados por la IA implica reconocer que esta es parte de contextos y sistemas sociales más amplios y, por tanto, no puede concebirse al margen de los actores y procesos sociales que la rodean. La explotación y entendimiento de la tecnología, particularmente por los Gobiernos, no puede enfocarse solo en los aspectos técnicos. También pasa por considerar principalmente las respuestas y actitudes sociales, por mantener su responsabilidad de garantizar el respeto de los derechos humanos y por generar diálogos y acuerdos alrededor de lo que espera la sociedad de la IA. Solo de esa manera será posible construir la confianza necesaria para lograr su plena adopción.

/2

USO RESPONSABLE

de la inteligencia artificial en el **sector público**

Este capítulo examina los retos relacionados con el uso ético y responsable de la IA por el sector público, centrándose en cuatro aspectos fundamentales: condiciones para un uso efectivo, orientación de los recursos humanos, retos culturales y legitimidad. Además, discute el riesgo de introducir sesgos en los sistemas de IA[7] , los tipos de sesgos existentes y la forma de minimizarlos. A continuación, se explica la importancia de establecer y aplicar un conjunto de principios éticos, considerando especialmente que se trata de una tecnología en constante y rápida evolución y la necesidad de atender intereses y opiniones diversos. La última sección aborda cómo definir y poner en práctica un marco ético, detallando alternativas para su cumplimiento, incluidas la gobernanza de los datos y la regulación, para cerrar analizando lo que significa un ecosistema de confianza basado en un marco regulatorio para esta tecnología.

7 Los términos sistema de IA, modelo de IA, algoritmo de IA o solución de IA son usados de manera equivalente en este capítulo.

LOS GRANDES DESAFÍOS DE LA IA **PARA EL SECTOR PÚBLICO**

El despliegue de la IA plantea a las entidades públicas un conjunto de desafíos derivados de la necesidad de que los sistemas de IA sean justos, eficientes y eficaces. Aquí se presentan los desafíos referentes a cuatro aspectos críticos del diseño y operación de sistemas de IA: el uso efectivo de los datos y la tecnología, las capacidades humanas, la cultura de lo público, y la legitimidad y confianza.

Uso efectivo de los datos y la tecnología

Para alcanzar un despliegue eficaz y ético de sistemas de IA que apoyen la toma de decisiones en el sector público, es indispensable acceder a datos precisos, que mantengan la privacidad y cumplan las normas éticas para su uso (Berryhill *et al.*, 2019). Por ello, se hace necesario que las entidades públicas de orden nacional y local definan e implementen políticas y estrategias de gestión de los datos que, de acuerdo con el contexto de aplicación de una solución de IA, permitan:

- Establecer lineamientos y estándares que orienten a los interesados en la implementación de programas de gestión de datos y de la tecnología de IA.
- Aumentar el uso y análisis de datos.
- Eliminar obstáculos a la disponibilidad, calidad, pertinencia, interoperabilidad, seguridad y privacidad de los datos.

En términos de disponibilidad, los Gobiernos deben asegurarse de que cuentan con fuentes de datos internas y externas accesibles, oportunas y constantes para su aprovechamiento a través de la IA. Para facilitar el uso de datos de gobierno a funcionarios, diseñadores de política y proveedores, es necesario disponer de diccionarios o catálogos de datos que faciliten su identificación, selección y descubrimiento (Berryhill *et al.*, 2019; Desouza, 2018). De igual modo, para compartir datos con otras entidades, es importante mantener esfuerzos para la publicación de conjuntos de datos relevantes en portales dedicados a ello y que se ajusten a las necesidades de los usuarios. Cabe mencionar, además, que los Gobiernos deben supervisar sus activos de datos y los métodos de recolección y evaluación de los datos existentes, de manera que puedan identificar eventuales ausencias de datos en las muestras a utilizar, las posibles variaciones en el tiempo y los sesgos que se puedan producir.

El éxito o fracaso en el análisis de datos a partir de los sistemas de IA depende en gran medida de la calidad de los datos. Para mantener esa calidad, es clave que las entidades evalúen y mejoren los métodos de recopilación de los datos estructurados. Se deben también establecer mecanismos que faciliten la interacción entre el Gobierno y las comunidades de usuarios, con el fin de comprender cómo perciben y miden su calidad en las diferentes dimensiones (ver el apartado «Tipos y calidad

de los datos» en el Capítulo 1). Para mejorar la calidad de los datos, se puede usar una herramienta como la tercerización masiva (*crowdsourcing*), que permite la colaboración abierta entre distintos interesados para analizar conjuntos de datos, con el propósito de verificar su calidad y corregir errores. Su implementación ayudará a los usuarios a disminuir los tiempos de depuración y limpieza, así como a aumentar la confianza en los datos a los que se acceden.

Para dar una respuesta apropiada a las cuestiones que se busca resolver mediante sistemas de IA, se requiere determinar la pertinencia de los datos, razón por la cual es preciso identificar los que se necesitan, revisar si están disponibles y analizar si sus tipos y licencias de uso son adecuados para encontrar soluciones a la situación de interés y las necesidades de los interesados. Así mismo, es preciso que los responsables del procesamiento y análisis de datos se aseguren de utilizar fuentes adecuadas, que permitan obtener muestras representativas para el entrenamiento de los algoritmos de acuerdo con el escenario o situación que se desee abordar a través de la IA.

Para lograr un aprovechamiento eficaz de los datos en iniciativas de IA en el sector público, se debe invertir en la modernización de la infraestructura tecnológica y los sistemas de información heredados, de manera que estos sean confiables y seguros (Desouza, 2018). Estas inversiones son clave para soportar las etapas de la cadena de valor de los datos (adquisición, análisis, curaduría, almacenamiento y uso), descritas en el Capítulo 1. Simultáneamente, se debe mejorar el intercambio e interoperabilidad de los datos entre las diferentes áreas de las entidades públicas del orden nacional, regional y local. Estas mejoras deben abordar aspectos técnicos (actualización de los sistemas de información, gestión de la calidad de datos, definición y adopción de estándares de intercambio comunes y de formatos de publicación); semánticos (documentación de metadatos); culturales (gestión del cambio para fomentar el valor de los datos y los beneficios de su uso); organizacionales (ajustes de procesos y creación de espacios para realizar experimentos tempranos del uso de IA); y humanos (conocimiento y habilidades digitales entre los funcionarios públicos) (OCDE, 2019d).

PARA DAR UNA RESPUESTA APROPIADA A LAS CUESTIONES QUE SE BUSCA RESOLVER MEDIANTE SISTEMAS DE IA, SE REQUIERE DETERMINAR LA PERTINENCIA DE LOS DATOS, RAZÓN POR LA CUAL ES PRECISO IDENTIFICAR LOS QUE SE NECESITAN, REVISAR SI ESTÁN DISPONIBLES Y ANALIZAR SI SUS TIPOS Y LICENCIAS DE USO SON ADECUADOS PARA ENCONTRAR SOLUCIONES A LA SITUACIÓN DE INTERÉS Y LAS NECESIDADES DE LOS INTERESADOS.

Recursos humanos

Un requisito fundamental para el despliegue de la IA en el sector público es la presencia de empleados que entiendan la tecnología y el potencial que esta ofrece, no solo a nivel técnico, sino también a nivel operativo y directivo. A nivel técnico, la ausencia de capacidades puede traducirse en una dependencia de los proveedores externos, haciendo más lento y limitado el uso de la tecnología. En el caso de funciones directivas, la falta de nociones técnicas, el desconocimiento del potencial que ofrece o de aspectos legales y éticos involucrados en su uso pueden obstaculizar proyectos institucionales orientados a su adopción. De igual manera, es necesaria la presencia de líderes dispuestos a trabajar de manera diferente y a usar su influencia para eliminar obstáculos y apoyar la exploración de la IA a fin de mejorar el desempeño del sector público.

De acuerdo con el Foro Económico Mundial, la lista de ocupaciones con mayor demanda en toda la economía la encabezan los analistas y científicos de datos, los especialistas en IA y aprendizaje automático, y los especialistas en macrodatos (*big data*). Para los Gobiernos, ese auge de la demanda implica una fuerte competencia a la hora de atraer ese talento (WEF, 2020b). A eso se suma que la adquisición de habilidades para desarrollar soluciones de IA toma tiempo, por lo que la capacitación de los empleados actuales, que es la otra alternativa, debe darse en el marco de una estrategia de desarrollo de recursos humanos de largo plazo, que tenga en cuenta la clasificación de los puestos de trabajo, los programas de capacitación y el tipo de perfiles y habilidades demandados (Desouza, 2018).

La capacitación de los empleados públicos incluye la actualización y perfeccionamiento de ciertas habilidades, que serán necesarias para desempeñar las nuevas funciones que emerjan a medida que se adopte la IA. Algunas de ellas podrán exigir una interacción directa con los sistemas desarrollados, la interpretación de sus resultados o su monitoreo. Al mismo tiempo, la automatización implicará la reubicación de un número significativo de trabajadores, para lo cual es importante una articulación entre el sector público, las empresas y las instituciones, de manera que se facilite esa reacomodación de la fuerza laboral.

Cultura y procesos públicos

En el sector público, existen pocos incentivos a la innovación. A diferencia de lo que ocurre en el sector privado, los empleados públicos no suelen ser estimulados a tomar riesgos o experimentar. Por el contrario, en entornos institucionales muy rígidos o excesivamente regulados, hacer las cosas de manera diferente o equivocarse en la exploración de nuevas rutas de intervención puede acarrear sanciones. En el caso de los líderes organizacionales, la poca disposición a correr riesgos o la ausencia de una mentalidad innovadora puede perjudicar su compromiso con la adopción de nuevas tecnologías transformadoras como la IA y generar un ambiente de poca apertura al cambio (Ubaldi *et al.*, 2019).

El desconocimiento del alcance y las posibilidades que ofrece la tecnología, sumado al temor que para muchos funcionarios despierta la automatización de algunas de las tareas que realizan, también suscita resistencia a su incorporación en el proceso público y los cambios que pueda generar.

Otro aspecto que constituye una barrera para la adopción de la IA son los procesos de adquisición o compra pública. En primer lugar, el hecho de que muchos desarrolladores consideren los algoritmos como activos de propiedad intelectual impide que las entidades puedan

eventualmente adaptarlos o actualizarlos para mantenerlos relevantes y pertinentes (WEF, 2020a). En segundo lugar, los procesos de contratación lentos, complejos y muy específicos dificultan la provisión de soluciones de manera oportuna y entorpecen la participación de proveedores, particularmente empresas pequeñas o emergentes (*startups*).

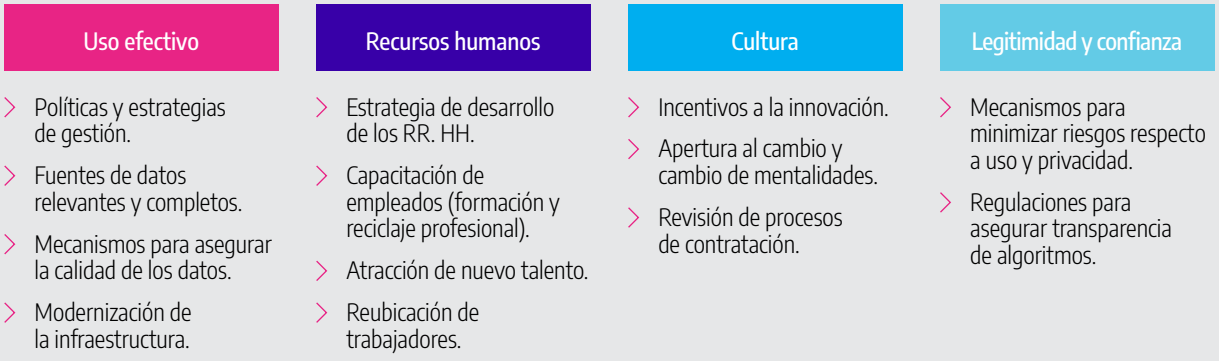
Legitimidad y confianza pública

Lograr que la ciudadanía acepte, confíe y respalde el uso de la IA por parte de los Gobiernos es uno de los mayores retos, si no el principal, para el sector público. Para conseguirlo, este último debe sortear dos situaciones: por un lado, tiene que ofrecer garantías a los ciudadanos ante usos indebidos o posibles perjuicios derivados del uso de algoritmos y, por el otro, debe aumentar su propia eficiencia, optimizar el ciclo de las políticas públicas y la atención a los ciudadanos usando justamente algoritmos. Esto es, «gobernar algoritmos mientras se gobierna con algoritmos» (Kuziemski y Misuraca, 2020).

En el centro de ese reto está la capacidad de los Gobiernos para generar respuestas y mecanismos que minimicen los riesgos y preocupaciones que se derivan del uso de la IA, tanto en el ámbito público como en el privado, alrededor de asuntos como la privacidad y la transparencia. El uso no autorizado de datos personales o su divulgación constituyen una amenaza para la protección del derecho a la privacidad y generan cuestionamientos sobre el alcance que tiene la tecnología y cómo puede ser usada en contra de grupos o individuos. Así mismo, la complejidad de ciertos algoritmos, cuándo se usan y para qué son asuntos que exigen claridad por parte del Gobierno de manera que su actuación y decisiones puedan justificarse y entenderse.

Al mismo tiempo, situaciones de discriminación o exclusión, entre otros efectos indeseables causados por decisiones basadas en IA, muchas veces sobre individuos o grupos que no cuentan con las herramientas necesarias para pronunciarse ante tratos injustos, son factores que debilitan la confianza ciudadana en la tecnología. Con frecuencia es la presencia de sesgos lo que conduce a dichas situaciones discriminatorias. Debido a su relevancia, este aspecto se considera en mayor detalle a continuación.

Figura 2.1
Retos de la IA en el sector público y medidas para mitigarlos



Fuente: Elaboración propia.

EL SESGO EN LA **INTELIGENCIA ARTIFICIAL**

Los sesgos en los sistemas de IA se refieren a errores, inexactitudes o anomalías que se presentan en los algoritmos con los que opera, pudiendo tener impactos no deseados en individuos y grupos, principalmente en términos de discriminación o exclusión; de ahí que se considere que la existencia de algún tipo de sesgo en un sistema de IA determina si este es o no es justo. En este sentido, la presencia de sesgos representa un serio riesgo para las entidades públicas.

Cuando se habla de un sistema de IA justo, se hace referencia a que no tiene ningún tipo de prejuicio o favoritismo hacia un individuo o grupo en particular por sus características o atributos (Mehrabi *et al.*, 2019). Ahora bien, para determinar si un sistema de IA es justo, es necesario considerar el contexto histórico y social en el cual dicho sistema será desplegado y utilizado (Silberg y Manyika, 2019).

Los sesgos en los sistemas de IA pueden darse principalmente por dos razones. La primera de ellas, por suposiciones o prejuicios que se introducen durante el diseño y desarrollo del algoritmo, el cual es realizado por humanos que tienen sus propios sesgos cognitivos y sociales, los cuales terminan siendo integrados en los parámetros que se definen para el sistema. La segunda, por suposiciones o prejuicios sociales o históricos, introducidos durante la recolección o etiquetado de los datos que se emplean para entrenar el sistema, o por su falta de representatividad, es decir, la ausencia de datos.

Los sesgos pueden incorporarse en cualquier momento del ciclo de vida de los sistemas de IA (ver la Figura 2.2, más adelante) y ser de distintos tipos, como se muestra a continuación (Suresh y Guttag, 2020):

- **Sesgo histórico.** Surge cuando los valores u objetivos que se codifican en un algoritmo no reflejan la realidad actual; pueden infiltrarse incluso cuando se tiene un adecuado muestreo y selección de características para su adquisición.
- **Sesgo de representación.** Ocurre cuando un grupo no está adecuadamente representado en la muestra de la cual se van a obtener los datos, haciéndolos inapropiados para hacer generalizaciones.
- **Sesgo de medición.** Se produce al elegir y medir las características y etiquetas particulares de cada conjunto de datos, pudiendo omitir factores importantes o introducir «ruido».
- **Sesgo de agregación.** Se da durante la construcción del modelo, cuando distintas poblaciones se combinan de manera inapropiada. En muchas aplicaciones, la población de interés es heterogénea y es poco probable que un solo modelo se adapte a todos los subgrupos.

- **Sesgo de evaluación.** Aparece durante la iteración y evaluación del modelo. Puede darse cuando se extraen conclusiones falsas para un subgrupo, basándose en la observación de otros subgrupos diferentes, o cuando no se utilizan las métricas apropiadas para la forma con la que se usará el sistema.

- **Sesgo de implementación.** Surge una vez que se ha puesto en marcha el modelo, cuando el sistema se usa o se interpreta de manera inapropiada.

Además de mantener, reafirmar o amplificar disparidades o situaciones de exclusión existentes en el mundo real, los sistemas de IA sesgados pueden, por ejemplo, afectar la obtención de subsidios y otras ayudas sociales o el acceso a servicios públicos, como el de educación, salud o la administración de la justicia. Ese fue el caso del algoritmo utilizado en el sistema judicial de los Estados Unidos (COMPAS, por sus siglas en inglés), que mostró tendencia a etiquetar a personas de piel negra con mayores probabilidades de reincidencia, en oposición a las personas de piel blanca (Spielkamp, 2017).

A pesar de la importancia de prevenir y minimizar los sesgos en los sistemas de IA, no existe todavía una fórmula exacta para hacerlo, pero es posible realizar acciones concretas que ayuden a prevenirlos y mitigar su efecto. En primer lugar, es necesario invertir tiempo y recursos para auditar los datos, especialmente cuando información sensible, como la edad, el género o la raza, forman parte del conjunto de datos empleados para entrenar el sistema de IA (Clausen, 2020; McKenna, s. f.). Sin embargo, la auditoría de los datos no es suficiente, ya que, como se mencionó, los sesgos pueden presentarse en varias etapas del proceso, afectando de manera significativa las decisiones tomadas por el algoritmo. Por lo tanto, para implementar en la práctica modelos no sesgados, es necesario realizar un monitoreo constante durante todo el ciclo de vida del sistema y contar con un modelo de gobernanza claro (Clausen, 2020).

Por otro lado, es necesario tratar de garantizar que no se estén introduciendo sesgos y prejuicios propios del desarrollador en el diseño y construcción de los algoritmos. Para ello, es recomendable contar con grupos multidisciplinarios y heterogéneos que permitan minimizar este riesgo (Shin, 2020; McKenna, s. f.). De igual forma, someter los algoritmos a la evaluación de distintos grupos antes de una etapa de producción puede ayudar a identificar los sesgos potenciales (Shin, 2020).

Desde el punto de vista técnico, es posible hablar de tres métodos para mitigar los sesgos presentes en los sistemas de IA:

1. Métodos de preprocesamiento enfocados en los datos.

2. Métodos de procesamiento enfocados en los algoritmos de aprendizaje automático.

3. Métodos de posprocesamiento enfocados en el modelo de aprendizaje automático (*machine learning*)[8].

8 Parara un mayor detalle de estos métodos, consultar Ntoutsi *et al.* (2020).

Recuadro 2.1

Principio de no discriminación

El Instituto Alan Turing del Reino Unido incluyó en la "Guía para el diseño e implementación responsable de sistemas de IA en el sector público" el principio de no discriminación en los términos que se indican a continuación.

> Los diseñadores y usuarios de sistemas de IA, que procesan datos sociales o demográficos relacionados con características de sujetos humanos, patrones sociales o formaciones culturales, deben priorizar la mitigación de sesgos y la exclusión de influencias discriminatorias en los resultados e implementaciones de sus modelos. Priorizar la no discriminación implica que los diseñadores y usuarios de los sistemas de IA aseguren que las decisiones y comportamientos de sus modelos no generen impactos discriminatorios o inequitativos en las personas y comunidades afectadas. Esto implica que estos diseñadores y usuarios garanticen que los sistemas de IA que están desarrollando e implementando:
>
> - Están entrenados y probados sobre conjuntos de datos apropiadamente representativos, relevantes, precisos y generalizables (equidad de datos).
> - Poseen arquitecturas de modelo que no incluyen variables objetivo, características, procesos o estructuras analíticas (correlaciones, interacciones e inferencias) que no son razonables, moralmente objetables o injustificables (equidad de diseño).
> - No tienen impactos discriminatorios o inequitativos en la vida de las personas a las que afectan (equidad de resultado).
> - Son implementados por usuarios suficientemente capacitados para implementarlos de manera responsable y sin sesgos (equidad en la implementación).

Notas: Traducción de los autores.

Fuente: Leslie (2019, p. 14).

Algunas herramientas disponibles actualmente para reducir los sesgos en los sistemas de IA son las siguientes (Kantarci, 2021):

- **AI Fairness 360.** Librería de código abierto de IBM para detectar y mitigar sesgos en algoritmos de aprendizaje automático no supervisado. Dicha librería contaba hasta 2020 con 34 contribuyentes en GitHub[9]. AI Fairness 360 permite a los programadores probar sesgos en modelos y conjuntos de datos con una serie completa de métricas, además de mitigar los sesgos con la ayuda de un paquete de 12 algoritmos, entre los que se encuentran Learning Fair Representations, Reject Option Classification y Disparate Impact Remover.
- **IBM Watson OpenScale.** Realiza una verificación y mitigación de sesgos en tiempo real cuando el sistema de IA está tomando las decisiones.
- **What-if tool.** Usando esta herramienta de Google es posible probar el desempeño del sistema en situaciones hipotéticas, analizar la importancia de diferentes características de los datos y visualizar el comportamiento de múltiples modelos y subconjuntos de datos de acuerdo con diferentes métricas de equidad.

9 GitHub es una plataforma de desarrollo colaborativo de *software* (https://github.com/).

/2

ÉTICA DE LA IA
Y DE LOS DATOS

El concepto de ética hace referencia al estudio de la moralidad, que se entiende como un sistema de reglas y valores que guían la conducta humana, junto con los principios para evaluar esas reglas. En consecuencia, cuando se dice que un comportamiento es ético, no estamos diciendo necesariamente que sea un «buen» comportamiento, sino que se ajusta a ciertos valores (WEF, 2019a).

Aplicada a la IA, la ética se entiende como el conjunto de reglas y valores que se ajustan a lo que se considera correcto y aceptable para guiar su desarrollo y uso bajo conductas morales. A esos valores y reglas, se suman principios y otros mecanismos que definen deberes y obligaciones básicos para el despliegue de sistemas de IA, que, además de éticos, sean justos y seguros (Leslie, 2019).

Esta discusión es central para los países y la comunidad internacional, dado el acelerado ritmo de avance de esta tecnología, que permitirá procesar una cantidad de datos cada vez mayor, de manera más precisa y en diversos campos de aplicación, con efectos aún desconocidos sobre la vida de las personas y el orden social. De esta manera, la responsabilidad de minimizar cualquier impacto negativo o perjudicial de la IA aumenta en la medida en que crecen las expectativas de sus beneficios y se expande su adopción.

Dos elementos son, por tanto, fundamentales para definir la ética de la IA: de un lado, el reconocimiento de los daños que puede causar a los individuos y a la sociedad, sea por el abuso, el mal uso, problemas de diseño o efectos no deseados de la tecnología, y, de otro lado, lo que se considera como deseable o beneficioso (reglas, valores, principios, técnicas) tanto a nivel individual como colectivo, en oposición a dichos daños. Respecto al primer elemento, los potenciales daños que estaría generando la IA y que han sido apuntados de modo recurrente incluyen:

- **Brechas de responsabilidad.** Decisiones que pueden afectar negativamente la vida de personas y grupos son tomadas de manera autónoma por sistemas de IA. Si bien los datos y modelos son diseñados por humanos, los procesos de aprendizaje que suceden de forma automática y la intervención de distintas personas durante el ciclo de vida de los sistemas dificultan la asignación de responsabilidades legales y, de esa manera, obstaculizan la presentación de recursos por parte de quienes se ven perjudicados, afectando el ejercicio de sus derechos.

- **Resultados no transparentes o justificables.** Además de procesar cantidades inmensas de datos, muchos algoritmos de IA utilizan métodos de gran complejidad para la mente humana (p. ej., redes neuronales profundas), dificultando el entendimiento y las razones que generan una decisión o resultado específico. Cuando se trata de decisiones polémicas que conllevan discriminación, injusticia o desigualdad, la falta de transparencia de los sistemas de IA es problemática y deteriora la confianza en el uso de la tecnología.

- **Violaciones a la privacidad.** La captura y procesamiento de datos personales en sistemas de IA pueden llevarse a cabo sin el debido consentimiento de sus titulares, mientras que un tratamiento inadecuado puede conducir a la divulgación de información personal. No estar informado de cuándo ocurre dicha captura y tratamiento, ni por parte de quién y mucho menos para qué, impide a los ciudadanos decidir sobre el uso que se hace de sus propios datos. Además, cuando un sistema de IA direcciona información a una persona sin que esta lo sepa, puede estar coartando la capacidad que tiene el individuo de tomar decisiones autónomas (Leslie, 2019).

- **Aislamiento y desintegración de las conexiones sociales.** Cuanto más automatizadas, personalizadas y dependientes de sistemas de IA sean las decisiones humanas, menores serán las necesidades de interacción entre individuos, lo que tendría efectos aún desconocidos sobre nuestra manera de pensar y relacionarnos. Una excesiva personalización mediada por algoritmos, sumada a la multiplicación de los sesgos en ellos implícitos, pueden potenciar la polarización o distorsionar la realidad al reducir la pluralidad de opiniones y la exposición a visiones del mundo diferentes a la propia.

- **Resultados no confiables, inseguros o manipulados.** Un manejo inadecuado de los datos, problemas en el diseño de los algoritmos (incluyendo los sesgos), la falta de un correcto mantenimiento, su uso en situaciones no deseadas o el aprendizaje de comportamientos no seguros, entre otros factores, pueden comprometer la confiabilidad, validez y seguridad de los algoritmos y sus resultados. Además, los errores que pueden presentarse con el *software y hardware* que forman parte de los sistemas de IA, los hacen susceptibles a la equivocación y la manipulación, representando riesgos para individuos, organizaciones y países.

- **Discriminación.** El uso de la tecnología puede conducir a la discriminación o la exclusión de individuos o grupos como resultado de los sesgos presentes en los datos y las estructuras e historia que estos representan (inclusive la ausencia de ellos). La discriminación también puede surgir como consecuencia de los prejuicios conscientes e inconscientes de quienes diseñan, implementan e interpretan los sistemas de IA. Otro riesgo es que refuerce y amplifique los sesgos y prejuicios existentes en la sociedad, con efectos sobre la igualdad y la equidad.

- **Pérdida de empleos.** Existe una preocupación sobre el desplazamiento laboral que puede ocasionar el uso de la IA, que trae además incertidumbre sobre la generación de nuevas oportunidades de trabajo, teniendo en cuenta el mayor nivel de automatización y digitalización. Se espera, además, que los impactos en el mercado laboral sean mayores para las mujeres, las minorías y la población con menor nivel educativo, lo que puede traducirse en una profundización de las brechas sociales y económicas existentes.

- **Impacto ambiental.** La creciente demanda de minerales utilizados en la elaboración de baterías para dispositivos electrónicos tiende a acelerar la tasa de agotamiento de esos recursos. Así mismo, el consumo de aparatos electrónicos y la obsolescencia programada de muchos de ellos aumentarán el volumen de basura electrónica y materiales tóxicos en el ambiente. A lo anterior se suman la huella de carbono generada en el entrenamiento de la IA por cuenta de los grandes requerimientos de energía para el procesamiento de grandes volúmenes de datos (Bird *et al.*, 2020).

Las diferentes posiciones respecto a lo que se considera deseable o beneficioso para los individuos y la sociedad frente a los posibles perjuicios de la IA pueden conciliarse mediante la creación de marcos éticos para el diseño y despliegue de esta tecnología. En los últimos dos años, Gobiernos, organismos internacionales, el sector privado y la sociedad civil han publicado un número considerable de marcos éticos. En ellos, es notoria la búsqueda de un enfoque centrado en lo humano y la influencia del marco internacional de los derechos humanos (Fjeld *et al.*, 2020), el cual presenta una visión consensuada de lo que se considera deseable y conveniente para la humanidad de modo general.

A pesar de sus diferentes orígenes, los marcos éticos para la IA tienden a converger en la definición de un conjunto básico de principios, cuyo significado y alcance pueden variar considerablemente, demostrando que estos se adaptan a los valores compartidos por una comunidad, a los contextos culturales, geográficos y organizacionales e inclusive a regulaciones o marcos sectoriales. A pesar de esas diferencias, es posible señalar un conjunto de principios adoptados más frecuentemente por varias instituciones internacionales y en América Latina, representados en el Cuadro 2.1.

Cuadro 2.1
Principios éticos para el despliegue de la IA

Principio	OCDE	Comisión Europea, Grupo independiente de expertos de alto nivel sobre IA	IA América Latina	IEEE, Iniciativa global sobre ética de sistemas autónomos e inteligentes
Privacidad			●	
Responsabilidad	●		●	●
Seguridad y protección	●	●	●	●
Transparencia y "explicabilidad"	●	●	●	●
Equidad y no discriminación	●	●	●	
Control humano de la tecnología		●		
Promoción de las capacidades, valores y derechos humanos	●			●
Autonomía humana		●	●	●

Fuente: Elaboración propia.

Una vez establecidos los principios éticos para el despliegue de la IA, el paso siguiente es generar alternativas para llevarlos a la práctica, lo cual se logra combinando medidas técnicas y no técnicas. Las primeras incluyen el cumplimiento de la ética «desde el diseño», la cual puede enfocarse de diferentes maneras (WEF, 2019a):

- Enfoque de arriba hacia abajo (*top-down*). Los principios éticos se programarían directamente en el sistema de IA[10], por ejemplo, mediante la incorporación en la arquitectura del sistema de las normas que debería seguir en todo momento (lista blanca) y de las restricciones sobre determinados comportamientos o estados que el sistema jamás debería infringir (lista negra), así como combinaciones de ambas (Grupo Independiente de Expertos de Alto Nivel sobre Inteligencia Artificial, 2019).

10 En un enfoque de abajo hacia arriba (*bottom-up*) se espera que los algoritmos aprendan a tomar decisiones éticas a partir de la observación del comportamiento humano, lo que los expone a comportamientos comunes, pero no necesariamente éticos (WEF, 2019a).

- Enfoque casuístico. Las máquinas estarían programadas para reaccionar específicamente en cada situación en la que pudieran tener que tomar una decisión ética, lo que implicaría anticipar todos los posibles escenarios.

- Enfoque dogmático. Las máquinas serían programadas en línea con una escuela de pensamiento ético específica (lo que las haría especialmente rígidas).

- Implementación de la IA a un metanivel técnico. Supone el desarrollo de un sistema de monitoreo impulsado por la IA que controle el cumplimiento de leyes y reglas éticas predefinidas a un metanivel («IA guardiana»). Tal IA podría interferir técnicamente en el sistema y corregir directamente decisiones ilegales o poco éticas, o informar de esa situación a la autoridad correspondiente.

Otras alternativas técnicas para el cumplimiento de los principios éticos incluyen contar con métodos de explicación, como los que se vienen investigando en la «IA explicable»; la realización de ensayos y validaciones constantes para hacer un seguimiento minucioso de la estabilidad, solidez y funcionamiento del modelo durante su ciclo de vida; y el empleo de métricas e indicadores de calidad de servicio previamente definidos, que permitan evaluar el desempeño del sistema, tales como rendimiento, usabilidad, fiabilidad, seguridad y facilidad de mantenimiento (Grupo Independiente de Expertos de Alto Nivel sobre Inteligencia Artificial, 2019).

Las alternativas no técnicas contemplan, entre otras, la revisión y adaptación de la normatividad vigente, la expedición de códigos de conducta generales y sectoriales, la educación y concientización de quienes diseñan, implementan, usan y se ven afectados por la IA, certificaciones, marcos de gobernanza y, por supuesto, regulación. Estos dos últimos se tratan en las secciones siguientes.

En el ámbito del sector público y específicamente en lo que se refiere a la ética de los datos, la OCDE ha publicado un conjunto de diez "Principios de buenas prácticas" con los que se busca apoyar la implementación de la ética en proyectos, productos y servicios digitales, de manera que la confianza ocupe un papel central en su diseño y ejecución y que se mantenga la integridad pública con acciones específicas (OCDE, 2021). Tales principios incluyen:

- Integridad en la gestión de los datos.
- Conocer y respetar los acuerdos gubernamentales pertinentes para el acceso, el intercambio y el uso de datos confiables.
- Incorporar consideraciones éticas de los datos en los procesos de toma de decisiones gubernamentales, organizacionales y del sector público.
- Supervisar y mantener el control sobre las entradas de datos, en particular los que se utilizan para informar el desarrollo y entrenamiento de sistemas de IA, y adoptar un enfoque basado en el riesgo para la automatización de decisiones.
- Ser específico sobre el propósito del uso de los datos, especialmente en el caso de los datos personales.
- Definir límites para acceder, compartir y usar datos.
- Ser claro, inclusivo y abierto.
- Publicar los datos abiertos y códigos fuente.
- Ampliar el control que los individuos y colectivos tienen sobre sus datos.
- Ser responsable y proactivo en la gestión de riesgos.

GOBERNANZA DE LA IA EN EL SECTOR PÚBLICO

La innovación tecnológica tiene el potencial de aumentar la productividad, el crecimiento económico y, en general, el bienestar de las personas. Sin embargo, como se ha mencionado anteriormente, puede tener consecuencias negativas que obligan a cuestionar si las regulaciones actuales en los muchos ámbitos de aplicación son adecuadas. El reto de las instituciones de gobierno es definir políticas públicas que prevengan, mitiguen y corrijan los posibles efectos negativos de las nuevas tecnologías sin restringir las posibilidades de emprendimiento que ofrecen ni impedir sus potenciales beneficios (OCDE, 2018a).

La gobernanza tecnológica se define como el «proceso de ejercer la autoridad política, económica y administrativa en el desarrollo, difusión y operación de tecnología en las sociedades» (OCDE, 2018a). La gobernanza implica la orientación y definición de objetivos y la elección de los mecanismos para alcanzarlos, la regulación de su funcionamiento y la verificación de sus resultados, entre otros aspectos (Stirling *et al.*, 2019). En el caso particular de la IA, este es un tema especialmente relevante por tratarse de una tecnología de naturaleza cambiante, con niveles de madurez variable, según la técnica usada y el caso de uso, y una alta incertidumbre respecto a cuáles son y serán sus límites y efectos reales. Esto obliga a que los marcos de gobernanza sean objeto de revisión, ajuste y actualización permanente a medida que la tecnología y sus aplicaciones evolucionan, tanto en el ámbito público como en el privado.

Uno de los principales retos para la gobernanza de la IA de manera general y para su uso en el sector público en particular tiene que ver con el «dilema del control». Este se refiere a la dificultad que existe para tomar decisiones sobre la orientación y configuración de una tecnología o sus aplicaciones durante sus primeras etapas de desarrollo, cuando aún se carece de evidencia suficiente para identificar la necesidad de cambio o redireccionamiento. Entretanto, cuando se tienen evidencias consistentes sobre el desempeño de la tecnología y se hace innegable la necesidad de intervenciones para cambiar su trayectoria, estas son ya muy complejas y costosas (Collingridge, 1980).

La capacidad de identificar riesgos de manera oportuna y de actuar en consecuencia, superando el dilema del control, implica, para las entidades que usan la tecnología, adoptar posiciones más flexibles, construir capacidad de respuesta y contar con mecanismos que así lo favorezcan. Actuando en esa dirección, han ido ganando terreno los nuevos modelos de gobernanza más enfocados en los procesos de desarrollo y la adopción de la tecnología que en los resultados propiamente dichos, principalmente en el caso de las tecnologías emergentes. El Foro Económico Mundial (WEF, por sus siglas en inglés), por ejemplo, propone, en el marco de la Cuarta Revolución Industrial (4RI), la implementación de una gobernanza ágil, entendida como la disposición para navegar rápidamente el cambio generado por las tecnologías, aceptarlo y aprender de él mientras se contribuye al valor real o percibido por el usuario final. También hace referencia al diseño adaptativo de políticas, centrado en el ser humano, inclusivo y sostenible, que reconoce que el desarrollo de políticas no se limita a los Gobiernos, sino que es un esfuerzo de múltiples actores (WEF, 2018a).

Otros enfoques de gobernanza aplicables a la IA incluyen la gobernanza reflexiva, la gobernanza anticipatoria, la gobernanza adaptativa, la gobernanza distributiva y la gobernanza mixta. Todas ellas se caracterizan por su flexibilidad, dinamismo, apertura y temporalidad, abriendo espacios a la experimentación, el aprendizaje, la reflexión y el cambio, entre otras cualidades[11].

Características de la gobernanza responsable de la IA

A la luz de las consideraciones anteriores y del potencial transformador de la IA, es importante contemplar un marco de gobernanza que sea anticipatorio, inclusivo, adaptable y con propósito[12]. Este marco sirve para orientar no solo la actuación del sector público, sino también el despliegue general de la tecnología en la sociedad. A continuación se describe el alcance de cada una de estas características.

- La **anticipación** hace referencia a la exploración de la tecnología, sus propósitos y posibles trayectorias de desarrollo con el objetivo de aproximarse de manera *ex ante* a sus implicaciones y riesgos. Para hacerlo, puede recurrirse a formas participativas de evaluación prospectiva y tecnológica para trazar futuros deseables o a otras técnicas de anticipación, como la exploración de horizontes, la planificación de escenarios o la evaluación de visiones. También pueden definirse casos de uso en los que se determine la no aplicación de la IA —por ejemplo, para el desarrollo de armas— y casos de uso en los que se establezca la necesidad de ciertas condiciones —por ejemplo, autorizaciones de los usuarios y procedimientos de seguridad (WEF, 2019a)[13].

- La **inclusión** se refiere a la vinculación activa y al diálogo entre diversos públicos y actores desde las primeras fases de los desarrollos tecnológicos para obtener aplicaciones más útiles, pertinentes y justas. Este acceso a diferentes perspectivas es fundamental para erradicar los sesgos, construir algoritmos representativos de la diversidad existente en la sociedad y conectar las necesidades y expectativas sociales con las soluciones tecnológicas. Se pueden usar diversas herramientas en este plano para percibir y explorar motivaciones, dudas, visiones, experiencias, resistencias y dilemas ante la IA. Algunas son adecuadas para tratar un tema con plazo definido, como los grupos focales, o para prolongarlo más tiempo y obtener más detalles, como en las consultas públicas, conferencias, paneles de ciudadanos, laboratorios de valor y encuestas de opinión (RRI Tools, s. f.).

- La **adaptación** implica tener la capacidad de cambiar visiones, comportamientos y estructuras organizacionales para dar respuesta a nuevas circunstancias, perspectivas, conocimientos y normas. Esta capacidad es la que permite alinear la acción con las necesidades expresadas por las partes interesadas. En algunos casos, los mecanismos para impulsar la adaptación incluyen la aplicación del principio de precaución, una moratoria o un código de conducta, la ampliación de procesos de evaluación de tecnologías y prospectivos, el diseño sensible al valor (*value-sensitive-design*) o técnicas como la activación por etapas (*stage gate*) (RRI Tools, s. f.; Stilgoe *et al.*, 2013)[14].

11 Para ampliar, ver Kuhlmann *et al.* (2019).

12 Estos atributos de la gobernanza se inspiran en el enfoque de la investigación e innovación responsables (RRI, por sus siglas en inglés), que busca anticipar y evaluar las posibles implicaciones y expectativas sociales respecto a la investigación y la innovación, con el propósito de hacerlas éticamente aceptables, sostenibles y socialmente deseables. Para ello, señala que los procesos de investigación e innovación deben ser diversos e inclusivos, anticipativos y reflexivos, abiertos y transparentes, con capacidad de respuesta y adaptativos. Para ampliar, consultar la web siguiente: https://www.rri-tools.eu/es/about-rri#wha

13 Otras herramientas para conducir ejercicios de anticipación y evaluación de riesgos pueden consultarse en el Catálogo de Acciones Engage2020 (http://actioncatalogue.eu/) o en el kit de herramientas de riesgos y oportunidades futuras publicado por el Foresight Horizon Scanning Center del Reino Unido (IRISS, s. f.).

14 La eAnthology de Engage2020 ofrece diversos recursos para diseñar procesos de participación ciudadana en los procesos de innovación. Disponible en: http://engage2020.eu/media/Engage2020_withVideo.pdf.

- El **propósito** tiene que ver con la dirección que se da a la tecnología como resultado de ejercicios de anticipación, inclusión y adaptación, que le imprimen a la misma un sentido en razón a lo que socialmente se considera deseable. Esta característica de la gobernanza es determinante para orientar el desarrollo de la tecnología en función de los ciudadanos y su bienestar.

Consideraciones, mecanismos y herramientas para la gobernanza de la IA en entidades públicas

Los riesgos planteados por la IA y sus preocupaciones éticas son los principales impulsores en la construcción de marcos de gobernanza en el sector público que garanticen el diseño e implementación de sistemas de IA justos, confiables y transparentes, o cualquier otro atributo que sea definido en contextos específicos. La creación de esos marcos de gobernanza implica desplegar un amplio conjunto de mecanismos, herramientas y prácticas que, desde distintos frentes (técnico, de política pública, legal, de relación con actores interesados u otros), apunten al uso de la tecnología en favor del interés común.

Como se mencionó previamente, muchos países han definido sus principios éticos para el despliegue de la tecnología en todos los sectores, creando con ellos las bases de su gobernanza. Tales principios se convierten en un primer referente que las entidades públicas pueden adaptar, complementar o priorizar a fin de garantizar una correspondencia con sus funciones y la visión que se tiene de esta tecnología.

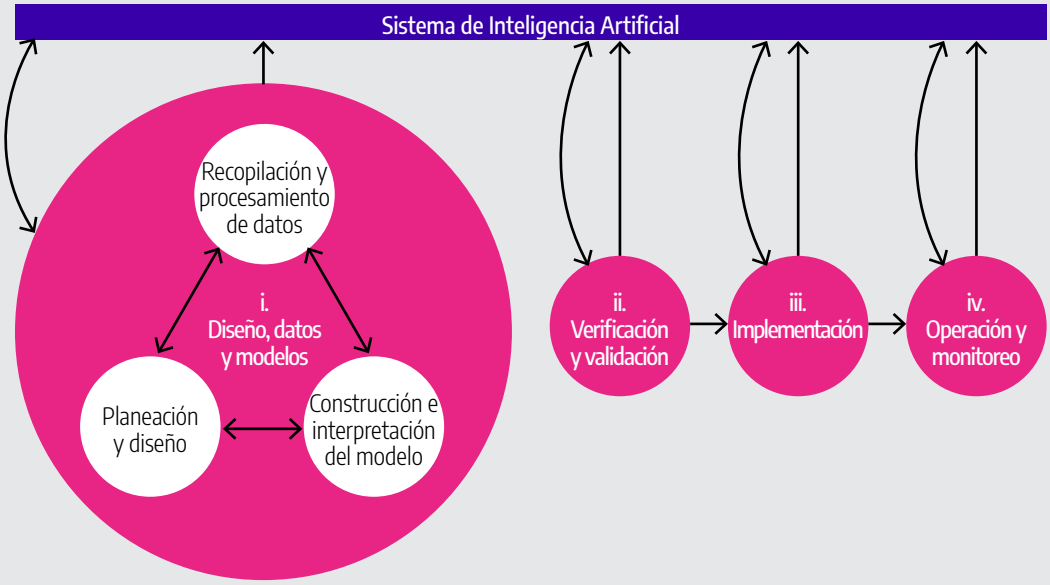
En el Reino Unido, por ejemplo, se ha definido que cualquier proyecto de IA en el sector público debe ser: i) éticamente permisible; ii) justo y no discriminatorio; iii) digno de confianza pública; y iv) justificable. Para facilitar el logro de esos objetivos se definen además unos valores —respetar, conectar, cuidar y proteger— y unos principios de vía rápida —equidad, responsabilidad, sostenibilidad y transparencia—. Los primeros buscan apoyar, respaldar y motivar un ecosistema responsable de diseño y uso de datos. Los principios, por su parte, buscan orientar un diseño y uso responsable de los sistemas de IA (Leslie, 2019).

En Nueva Zelanda, por su parte, las entidades públicas se comprometieron mediante la Carta de Algoritmos (*Algorithm Chapter*) a gestionar cuidadosamente los algoritmos, logrando un equilibrio entre privacidad y transparencia, evitando sesgos y reflejando sus leyes, como una forma de generar confianza ciudadana frente al uso de algoritmos de IA en el sector público. Al suscribir esa carta, las entidades se comprometen, con acciones específicas, a: i) mantener la transparencia, explicando claramente cómo los algoritmos informan las decisiones; ii) ofrecer un beneficio público claro; iii) enfocarse en las personas; iv) asegurarse de que los datos son adecuados para el propósito que se tiene; v) garantizar que la privacidad, la ética y los derechos humanos están protegidos, y vi) conservar la supervisión humana (Gobierno de Nueva Zelanda, 2020).

Una forma de avanzar en la estructuración de la gobernanza, cuando ya se tienen directrices generales proporcionadas por los principios éticos para la IA en el sector público, es considerar primero el proceso de desarrollo e implementación de un sistema de IA[15] (ver Figura 2.2). Después, se definen, en cada una de sus etapas, los procedimientos, protocolos y herramientas necesarios para facilitar la operacionalización de los principios durante todo el ciclo de los sistemas. Esos principios pueden divulgarse mediante guías y herramientas para la auditoría de algoritmos y la evaluación de impactos y riesgos, entre otros. En las secciones siguientes se abordan tres aspectos específicos de la construcción de la gobernanza: la gobernanza de los datos, la evaluación de riesgos e impactos y las estructuras y medidas para la operación de sistemas de IA.

15 Algunas entidades pueden optar por tomar como referente el ciclo de vida del desarrollo de *software* (SDLC, por sus siglas en inglés) o diseñar su propio proceso.

Figura 2.2
Ciclo de vida de los sistemas de IA



Fuente: OCDE (2019c).

Teniendo claro el ciclo de vida y ajustándose a los riesgos y posibles impactos individuales y públicos de los sistemas de IA, la gobernanza debe dar cuenta, entre otros, de los siguientes aspectos (Leslie, 2019):

- Los miembros del equipo y los roles involucrados en cada etapa y acción del proceso.
- Las etapas en las que es necesario intervenir y tener consideraciones específicas para cumplir los objetivos de gobernanza o principios del sector público.
- Plazos explícitos para las acciones de seguimiento, reevaluación y monitoreo necesarias.
- Protocolos claros y bien definidos para el registro (*logging activity*) y para establecer mecanismos que aseguren la posibilidad de auditar los procesos de principio a fin.

Es importante mencionar que las medidas definidas por las entidades, así como la rigurosidad en su cumplimiento, estará mediada por diversos factores que incluyen:

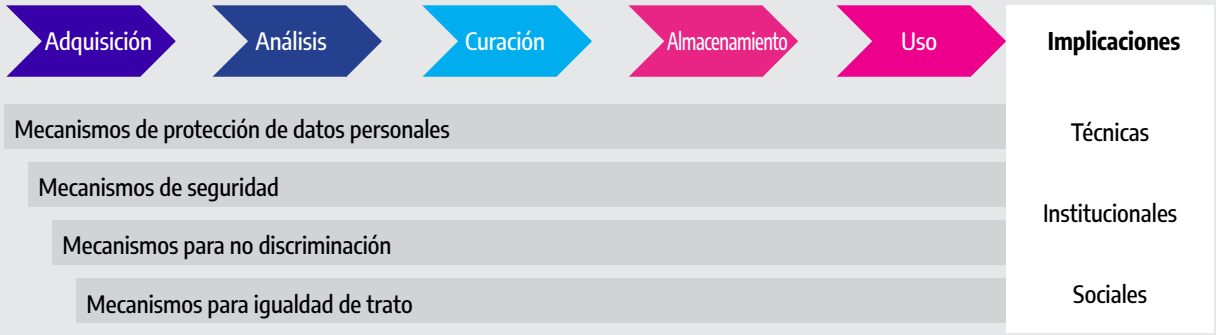
- La naturaleza, complejidad y nivel de riesgo de los sistemas de IA.
- El uso (o no) de la IA para la toma de decisiones y en funciones centrales de la entidad.
- La severidad de los impactos esperados sobre individuos, empresas y comunidades.
- Los recursos disponibles.
- La reglamentación existente.

La gobernanza de los datos

La gobernanza de los datos es vista como la manera de ejercer control sobre la calidad de los datos, el cumplimiento de los requerimientos legales y éticos relacionados con los datos y su utilización para la creación de valor público, garantizando resultados confiables. Es especialmente relevante teniendo en cuenta que los datos que nutren la IA provienen de diferentes fuentes, lo que incrementa la dependencia entre organizaciones y dificulta la asignación de responsabilidades. Además, errores en los datos pueden traducirse en decisiones sesgadas o ilegales, altos riesgos financieros y crisis políticas, entre otros, lo que tiene serias implicaciones para las organizaciones involucradas, los ciudadanos, las empresas y la sociedad en general. Para lograr su propósito, un buen esquema de gobernanza de datos para la IA requiere, entre otros, mecanismos de protección de datos personales, seguridad, no discriminación e igualdad en el trato, cubriendo toda la cadena de valor del dato y abordando las implicaciones técnicas, institucionales y sociales del intercambio de datos (Janssen *et al.*, 2020).

El presenta consideraciones relevantes para estructurar el marco de gobernanza de datos, especialmente aquellos usados por los sistemas de IA.

Figura 2.3
Gobernanza de datos



Fuente: Elaboración propia.

La gobernanza de los datos es vista como la manera de ejercer control sobre la calidad de los datos, el cumplimiento de los requerimientos legales y éticos relacionados con los datos y su utilización para la creación de valor público, garantizando resultados confiables.

Cuadro 2.2
Resumen de los principios para la gobernanza de datos

Principio	Descripción
Evaluar la calidad y el sesgo de los datos	Cuando los datos son usados en sistemas algorítmicos, es necesario evaluar su calidad y sesgos.
Detectar patrones cambiantes	Cuando los resultados de los algoritmos cambian, se debe verificar su validez e investigar las razones de tales cambios.
Información requerida	Minimizar la cantidad de datos que se comparten (solo lo necesario); por ejemplo, respuestas a preguntas en lugar de conjuntos de datos completos.
Recompensa por detectar errores	Las recompensas podrían usarse para alentar a las personas a detectar errores y problemas y reportarlos.
Informar al compartir	Cuando los gobiernos comparten datos sobre una persona u organización, estos deben ser informados para garantizar la transparencia y evitar el uso indebido.
Separación de datos	Separar los datos personales de los no personales y los datos sensibles de los no sensibles.
Control ciudadano de los datos	Empoderar a las entidades públicas, los ciudadanos y las organizaciones para que tengan el control y verifiquen la precisión de sus datos.
Recopilación de datos en la fuente	Para garantizar su exactitud y saber cómo se recaban.
Minimizar la autorización para acceder a los datos	Si una de las partes no necesita datos, no se le debe otorgar acceso.
Almacenamiento distribuido de datos	Los sistemas distribuidos son menos vulnerables y evitan combinar datos fácilmente sin autorización.
Administradores de datos	Asignar administradores de datos para formalizar la rendición de cuentas por la gestión de los recursos de información mientras se adhiere a la separación de preocupaciones.
Separaciones de preocupaciones	Las responsabilidades respecto a los datos deben distribuirse de tal manera que una única persona no pueda hacer un mal uso o abusar de los mismos.
Utilidad	Los datos deben reconocerse como un activo valioso que puede ser utilizado para la IA.

Fuente: Janssen *et al.* (2020).

Evaluación de riesgos e impactos

La evaluación de los posibles impactos que tendría el uso de un algoritmo de IA sobre individuos, empresas y comunidades, asociado a la clara identificación de los riesgos involucrados, es un requisito indispensable para determinar las acciones necesarias que garantizarán el cumplimiento de los principios éticos de la IA, entre ellas el nivel de participación humana. Estos análisis deben ser monitoreados durante el ciclo de vida del sistema de IA y ajustados cuando sea necesario.

Una primera evaluación debe contemplar las necesidades de los usuarios, las comunidades afectadas, los potenciales riesgos y sesgos en el sistema, así como escenarios de consecuencias no deseadas. Para estos últimos, es útil realizar el análisis del impacto específico sobre la protección de datos y la igualdad, considerando cómo el uso del sistema interactúa con mecanismos de supervisión, revisión y otras salvaguardas (WEF, 2020a).

Varios países han avanzado en la elaboración de herramientas específicas para la identificación de riesgos e impactos, entre ellos Canadá, Uruguay y Nueva Zelanda, cuyos casos se presentan a continuación. También el WEF ha diseñado, en colaboración con el Gobierno del Reino Unido y las empresas Deloitte y Splunk, una herramienta para la evaluación del riesgo de la IA, en el marco de una guía para la adquisición de soluciones y servicios de IA por el sector público, que está disponible para consulta en línea[16].

Canadá

El Gobierno canadiense ha definido la **Evaluación del impacto algorítmico** como pieza fundamental de su enfoque para el uso de la IA en el sector público. Su propósito es ayudar a las instituciones a comprender y reducir mejor los riesgos asociados con los sistemas de decisión automatizados y, a partir de ello, determinar las medidas pertinentes de acuerdo con las diferentes situaciones (Gobierno de Canadá, 2019). Para realizar la evaluación, se determina el nivel de riesgo que representa una decisión automatizada de acuerdo con su nivel de impacto sobre: i) los derechos de individuos o comunidades, ii) la salud o el bienestar de los individuos o comunidades, iii) los intereses económicos de los individuos, entidades o comunidades, y iv) la sostenibilidad continua de un ecosistema.

Los posibles niveles de impacto son cuatro:

- **Nivel I.** Las decisiones conducirán con frecuencia a impactos que son reversibles y breves.
- **Nivel II.** Las decisiones conducirán con frecuencia a impactos que serán probablemente reversibles y de corto plazo.
- **Nivel III.** Las decisiones conducirán con frecuencia a impactos que pueden ser difíciles de revertir y son continuos.
- **Nivel IV.** Las decisiones conducirán con frecuencia a impactos que son irreversibles y perpetuos.

Cada uno de estos niveles de impacto da lugar a requerimientos específicos en términos de revisión de pares, notificaciones, intervención humana en las decisiones, explicaciones requeridas, pruebas, monitoreo, entrenamiento, planes de contingencia y, finalmente, aprobación para la operación del sistema[17].

Uruguay

El Gobierno del Uruguay, en el marco de su Estrategia de IA para el Gobierno Digital, diseñó el **Estudio de impacto algorítmico (EIA)** para analizar los sistemas automatizados de apoyo a la toma de decisiones. Esta herramienta, dirigida a gerentes de proyectos o equipos que lideran proyectos vinculados a la temática, permite identificar los aspectos importantes y que merecen mayor atención o tratamiento. Se trata de un cuestionario con preguntas que buscan, por un lado, una caracterización del sistema, obteniendo información sobre la razón para llevarlo a cabo, el impacto social que se espera, las capacidades del sistema y el área en el que se estaría aplicando (servicios de salud, intereses económicos, asistencia social, entre otros); por otro lado, procuran una evaluación del impacto en términos del uso del sistema para la toma de decisiones (apoyo o reemplazo de la decisión humana), reversibilidad de los efectos de la decisión y su duración. Se indaga también por el origen y tipo de datos que alimentan el sistema y los actores interesados. Finalmente se pregunta por medidas para reducir y atenuar los riesgos relacionados con la calidad de los datos y la imparcialidad procesal.

16 Se puede acceder al libro de trabajo y los demás módulos de la guía en el sitio web del WEF (https://www.weforum.org/reports/ai-procurement-in-a-box#read-more-about-ai-and-public-procurement).

17 La herramienta, que forma parte de la Directiva sobre la toma de decisiones automatizada, puede consultarse en el sitio web del Gobierno de Canadá (https://www.tbs-sct.gc.ca/pol/doc-eng.aspx?id=32592).

Nueva Zelanda

Todas las agencias gubernamentales de Nueva Zelanda deben realizar un **Informe de evaluación del algoritmo,** que se fundamenta en el uso de una matriz de riesgo. En esta matriz se identifica la posibilidad de un resultado adverso no deseado (con tres categorías, probable, ocasional o improbable) y su nivel relativo de impacto (bajo, moderado o alto). Con el cruce de estas variables se establece el nivel general de riesgo representado por un color. Así, las casillas en verde simbolizan un riesgo bajo, las casillas amarillas un riesgo moderado y las casillas en rojo un riesgo alto (ver Figura 2.4). A su vez, los colores de las casillas determinan el curso de acción: con riesgo bajo puede aplicarse la carta de algoritmos, con riesgo moderado debe aplicarse la carta y con riesgo alto su aplicación es obligatoria (Gobierno de Nueva Zelanda, 2020)[18].

18 La Carta de Algoritmos está disponible en el sitio web del Gobierno (https://data.govt.nz/assets/data-ethics/algorithm/Algorithm-Charter-2020_Final-English-1.pdf).

Figura 2.4
Matriz de riesgo para sistemas de IA de Nueva Zelanda

Nota: El punto azul representa un riesgo bajo, el punto gris, riesgo moderado y el punto fucsia, riesgo alto.

Fuente: Gobierno de Nueva Zelanda (2020).

En cuanto a la **evaluación de riesgos**, el Cuadro 2.3 presenta los aspectos clave a considerar, de acuerdo con el WEF (2020a).

Lo más importante de las evaluaciones de riesgos e impactos es que conduzcan al diseño de estrategias para atenderlos y gestionarlos de la mejor manera (por ejemplo, intervención humana, revisión de pares, documentación y explicación del algoritmo, pruebas, consultas a la ciudadanía). Los funcionarios responsables de la toma de decisiones deben conocer particularmente los riesgos y las medidas para mitigarlos y, si es el caso, dar fundamento a la continuidad o cancelación de un proyecto (WEF, 2020a).

Cuadro 2.3
Aspectos clave de la evaluación de riesgos

Datos	› Sensibilidad. Cuanto más confidenciales los datos, mayores puntos de chequeo deben incluirse. Es necesario considerar si los datos pueden identificarse o revelar información personal.
	› Calidad. Ante dudas sobre la calidad de los datos, se recomienda incorporar garantías adicionales para evitar sesgos y reducir el riesgo del proyecto. Para asegurar la representatividad es posible que se requieran medidas específicas. Se deben considerar, además, sesgos sociales que podrían reflejarse en los datos.
	› Consentimiento. Es necesario tener autorización para usar los datos en el área de aplicación específica del sistema de IA y hay que asegurar que ese uso no sea inferido a partir de otros consentimientos.
Área de uso	› Escrutinio público. Un proyecto en sectores de servicios con especial sensibilidad, como los de salud, asistencia social, empleo, finanzas, justicia e inmigración, entre otros, exigen mayores consideraciones y precauciones.
Impacto socioeconómico	› Actores involucrados. Cuanto mayor sea el impacto potencial en individuos, empresas o comunidades, más importantes serán las consideraciones éticas y el análisis de la solución de IA.
	› Alcance del impacto. Es importante tener en cuenta cuántas personas serán impactadas y qué tan alto y probable es dicho impacto. El riesgo se incrementa cuando son grupos vulnerables.
Consecuencias financieras para la agencia e individuos	› Alcance del impacto financiero. Cuanto mayores sean las posibles consecuencias financieras, más se deben abordar todas las áreas relacionadas con las consideraciones específicas de la IA.
	› Tipos de impactos financieros. Pueden incluir aspectos monetarios, acceso al crédito, oportunidades económicas, escolaridad o formación, seguros y certificaciones.
Impacto del sistema de IA en sus procesos, empleados y actividades principales	› Impacto en las actividades centrales. A mayor dependencia tecnológica, mayor el riesgo. Además de los riesgos técnicos, deben mitigarse los riesgos reputacionales.
	› Impacto en las funciones de la agencia. Considerar si se reemplaza una función o solo se mejoran o aumentan las actuales.
	› Pérdida de empleos. Mayor automatización puede conducir a mayor pérdida de empleos, aumentando los riesgos y sensibilidad respecto al despliegue de la IA.
	› Intervenciones humanas. Pocos controles y contrapesos generan mayor riesgo. Es necesario que las soluciones que surgen como resultado de la IA se puedan explicar e interpretar.

Fuente: Basado en WEF (2020a).

Estructuras y medidas para la operación de sistemas de IA

Las estructuras de gobernanza son indispensables para que las entidades públicas puedan supervisar y controlar lo que ocurre durante el ciclo de vida de los sistemas de IA. Pueden ser estructuras creadas para atender específicamente las cuestiones relacionadas con la tecnología o resultar de adaptaciones o actualizaciones de estructuras existentes. En cualquiera de los casos, cada entidad deberá identificar y atender las necesidades que se deriven de su situación particular. Para hacerlo, pueden guiarse por las consideraciones y prácticas planteadas en el Cuadro 2.4, con base en el «Marco de gobernanza modelo de inteligencia artificial» y de la «Guía para su implementación y autoevaluación» (ISAGO), desarrollada por el Centro para la Cuarta Revolución Industrial del WEF, en colaboración con la Autoridad de Desarrollo de Medios de Información y Comunicaciones (IMDA) y la Comisión de Protección de Datos Personales (PDPC) de Singapur (WEF *et al.*, 2020).

Cuadro 2.4

Consideraciones prácticas para el establecimiento de estructuras de gobernanza de la IA

Sobre el desarrollo de estructuras adecuadas	
Para asuntos de supervisión, considere la pertinencia de las siguientes medidas	› Determinar qué es más útil y práctico para la entidad, adaptar las estructuras de gobernanza, riesgo y cumplimiento existentes o crear unas específicas para la IA. › Tener una estructura de gobierno experimental para probar e implementar soluciones de IA antes de establecer estructuras definitivas. › Establecer un comité con representantes de las áreas relevantes para validar la estructura de gobernanza de la IA. › Establecer un proceso donde el director de cada área desarrolle y asuma la responsabilidad de los controles y políticas respectivos, con la supervisión de expertos de la misma entidad. › Establecer controles y contrapesos. Se puede establecer un equipo interno para supervisar metodologías, algoritmos e implementación de la IA y un equipo separado o externo para hacer una validación. En caso de surgir diferencias o preocupaciones, deberán llevarse a cabo nuevas pruebas y validaciones.
Para implementar la estructura, considere	› La creación de un comité o junta presidido por la alta dirección e incluya a los directores o gerentes de las diferentes áreas y equipos. › Buscar que la alta dirección establezca expectativas o directrices claras para el gobierno de la IA dentro de la organización. › Decidir entre la toma de decisión de manera centralizada o descentralizada. • El enfoque centralizado se recomienda para sistemas considerados de alto riesgo o potencialmente polémicos, siendo necesaria su presentación ante la alta dirección o el comité de ética, si existe. • Para el enfoque descentralizado, recomendado para sistemas de menor riesgo, se pueden establecer casos de uso de IA permitidos y no permitidos con base en sus impactos potenciales y riesgos, que sirvan como referente a las diferentes áreas para avanzar o no con las soluciones de IA. › Revise periódicamente los procesos y estructuras de gobernanza.

Sobre roles y responsabilidades claros para el despliegue ético de la inteligencia artificial

Dentro de las funciones y responsabilidades que deben ser asignadas, se destacan	› El uso de un marco de gestión de riesgos y aplicación de medidas de control de riesgos que permita: i) evaluar y gestionar los riesgos de la implementación de la IA, incluyendo cualquier impacto perjudicial para las personas; ii) decidir sobre el nivel adecuado de participación humana en la toma de decisiones mejoradas por IA; iii) gestionar el proceso de entrenamiento y selección del modelo de IA. › El mantenimiento, monitoreo, documentación y revisión de los modelos de IA que han sido implementados, con el propósito de tomar medidas correctivas cuando sea necesario. › La revisión de canales de comunicación e interacciones con las partes interesadas para brindar información y canales de retroalimentación efectivos. › Garantizar que el personal que se ocupa de los sistemas de IA esté debidamente capacitado para, por ejemplo, interpretar los resultados y las decisiones del modelo de IA, detectar y gestionar sesgos, y reconocer y entender los beneficios, riesgos y limitaciones al usar la IA.

› Asegúrese de que todos los funcionarios se comprometan con la implementación de buenas prácticas:
  - A nivel estratégico, la junta o comité directivo es responsable de los riesgos y los principios éticos, mientras que sus miembros los traducen en estrategias.
  - En la implementación, además de la supervisión de la alta dirección, los líderes de proyectos y funcionarios deben asumir la responsabilidad por los mismos. Las funciones y responsabilidades para gestionar los riesgos y asegurar el cumplimiento de las normas vigentes tienen que estar claramente establecidas y documentadas. Los equipos jurídicos deben apoyar la puesta en marcha con la verificación de restricciones legales o requisitos existentes para el despliegue de la IA, así como estar atentos a recibir retroalimentación o atender dudas o cuestionamientos sobre asuntos éticos.

› Establezca responsabilidades diferenciadas para el personal estratégico y técnico. El primero debe ser responsable de definir las metas y verificar que los sistemas de IA estén orientados a su cumplimiento bajo las normas establecidas. El segundo debe ser responsable de las prácticas con los datos, la seguridad, la estabilidad y el manejo de errores.

› Asegúrese de que las personas involucradas en los procesos de gobernanza de la IA son plenamente conscientes de sus funciones y responsabilidades, tienen la formación adecuada y cuentan con los recursos y orientaciones necesarios para desempeñar dichas funciones.

› Revise periódicamente la descripción de los puestos de trabajo para aquellos roles que incluyen el despliegue de IA.

› Considere tener un equipo multidisciplinario que ofrezca una perspectiva amplia sobre el impacto de la implementación de la IA en la organización y las personas.

Fuente: WEF *et al.* (2020).

Las estructuras de gobernanza son indispensables para que las entidades públicas puedan supervisar y controlar lo que ocurre durante el ciclo de vida de los sistemas de IA. Pueden ser estructuras creadas para atender específicamente las cuestiones relacionadas con la tecnología o resultar de adaptaciones o actualizaciones de estructuras existentes.

UN ECOSISTEMA DE CONFIANZA: **MARCO REGULATORIO PARA LA IA**

La regulación se refiere al conjunto de reglas, normas y sanciones formales e informales que buscan moldear el comportamiento de las personas para lograr un objetivo o meta de política. Su objetivo es brindar certeza y gestionar los riesgos, al tiempo que permite que los beneficios se distribuyan de manera equitativa (WEF, 2020c).

Los sistemas regulatorios tradicionales comprenden instrumentos legalmente vinculantes —como leyes, decretos, resoluciones— e instrumentos no vinculantes —como guías y códigos de práctica, autorregulación, estándares y certificaciones, entre otros—. Estos instrumentos buscan influenciar el comportamiento de una forma menos directa y se clasifican dentro de las medidas no técnicas para implementar los principios éticos de la IA. En el caso de las tecnologías emergentes, la complejidad técnica y la velocidad de los cambios exponen diferentes limitaciones de estos sistemas regulatorios, entre las que figuran la poca flexibilidad y la poca proactividad. Esta situación está llevando a los Estados a explorar nuevos enfoques regulatorios que sean más ágiles y colaborativos, que implican, entre otros factores, una buena combinación de ambos tipos de instrumentos (vinculantes y no vinculantes), interdisciplinariedad y participación de diferentes actores.

Los siguientes son los principales desafíos que encuentra la regulación de IA en la actualidad (WEF, 2020c):

- **Regulación vigente.** Parte de la regulación existente permite a muchos países manejar ciertos aspectos de la IA, por ejemplo, los relacionados con la privacidad; sin embargo, es difícil encontrar cobertura de todos los aspectos que pueden derivarse de la evolución de la IA en el largo plazo.

- **Regular en la medida justa.** Vacíos regulatorios generan incertidumbre y pueden conducir a mayor discriminación; por otra parte, un exceso de regulación puede frenar u obstaculizar la innovación y dificultar la interoperabilidad global.

- **Tensiones sobre la forma de regular.** Por un lado, se sabe que la IA tendrá un gran impacto en la vida de las personas y, por el otro, aún no se conoce el alcance de ese impacto, sus retos y oportunidades. Las diferentes visiones sobre las áreas y actividades que deben regularse varían ampliamente entre regiones.

- **Cambios en la tecnología.** La IA evoluciona rápidamente, haciendo las formas tradicionales de regulación poco prácticas. Situaciones como la falta de consenso sobre qué dirección seguir para avanzar o el temor a obstaculizar la innovación impiden una rápida respuesta regulatoria. De cualquier forma, anticipar los impactos de las tecnologías emergentes no es sencillo, por lo que es necesario vincular a los actores que se verán afectados con los procesos de decisión.

Alternativas para abordar la regulación

Leyes y normas (*hard law*)

El desarrollo de una regulación específica se justifica en la medida en que los riesgos que plantea esta tecnología son nuevos o no están contemplados por la regulación actual. Antes de apresurarse a crear respuestas o acciones regulatorias técnicas y legales específicas que cubran los daños directos o indirectos asociados a los sistemas de IA, es necesario considerar hasta qué punto las leyes y normas (*hard law*) existentes pueden regularlos adecuadamente o el tipo de ajustes que serían necesarios para que eso ocurra. Por ello, es importante que la regulación vigente sea objeto de una revisión y auditoría permanente por parte de expertos, teniendo en cuenta que los individuos, empresas o comunidades que puedan estar siendo afectados no siempre son conscientes de ello o desconocen los argumentos y mecanismos para reclamar. También puede suceder que lo que se ve como daños individuales sean en realidad daños a grupos específicos, lo que muchas veces no es interpretado de esta manera y, por lo tanto, no se actúa en consecuencia.

Un buen ejemplo de esta aproximación lo ofrece el Gobierno de Canadá, donde la Comisaría de Protección de Datos Personales abrió un proceso de consulta a expertos sobre 11 propuestas para reformar la Ley de Protección de la Información Personal y los Documentos Electrónicos (PIPEDA, por sus siglas en inglés). El objetivo de la propuesta era reforzar la protección de la privacidad y otros derechos, particularmente en relación con los datos, en el desarrollo y la implementación de la IA, así como su futura regulación[19].

No obstante lo anterior, y a pesar del creciente número de países que ha decretado una estrategia nacional para la IA, la mayoría de los Estados han sido cautelosos en lo que respecta a la creación de instrumentos vinculantes específicos. Los principales avances en esa dirección se han dado en relación con las leyes de uso y protección de datos personales, las cuales tienen una incidencia directa sobre gran parte de los desarrollos de la IA, como lo demuestra el caso canadiense. En ese mismo sentido, el Reglamento General de Protección de Datos (RGPD) de la Unión Europea establece los requisitos específicos para empresas y organizaciones sobre captura, almacenamiento y gestión de los datos personales, los cuales se aplican tanto a las organizaciones europeas que tratan datos personales de ciudadanos de los países miembros como a las organizaciones que tienen su sede fuera de los mismos y cuya actividad se dirige a personas que viven en la UE[20].

Instrumentos no vinculantes

Hasta ahora, las acciones tendientes a regular la IA se han concentrado en el desarrollo de instrumentos no vinculantes (*soft law*). Entre ellas figuran guías[21], marcos éticos y herramientas para el análisis de riesgos e impactos, como los mencionados en secciones previas y que pueden también considerarse instrumentos de autorregulación.

19 Los detalles de la consulta y sus propuestas se pueden ver en la web de la Comisaría (https://www.priv.gc.ca/en/about-the-opc/what-we-do/consultations/completed-consultations/consultation-ai/pos_ai_202001/).

20 Disponible en el sitio de la Unión Europea (https://europa.eu/youreurope/business/dealing-with-customers/data-protection/data-protection-gdpr/index_es.htm).

21 Un ejemplo es la «Guía para la construcción y el uso de la IA en el sector público», publicada en el año 2019 por el Servicio Digital del Gobierno y la Oficina de Inteligencia Artificial del Reino Unido (GDS y OAI, 2019). Este documento ofrece orientaciones en cinco frentes, que van desde el entendimiento de la tecnología y la evaluación de su pertinencia, pasando por su planeación e implementación hasta la comprensión de aspectos éticos y de seguridad. Cada uno de estos frentes plantea definiciones, acciones concretas y ejemplos en un lenguaje concreto y sencillo para los encargados de tomar decisiones relacionadas con la IA en el sector público.

Una alternativa interesante, que viene surgiendo para replantear los modelos de regulación actuales ante la incertidumbre y expectativa generada por el avance de la IA y sus posibles impactos en la sociedad, es la creación de ambientes controlados donde puedan ponerse a prueba sistemas de IA. Esta opción ofrece a los reguladores la oportunidad de entender la tecnología e identificar las necesidades regulatorias que estimularían la innovación, al tiempo que garantizarían la protección de los ciudadanos y la adecuación a los marcos éticos de la IA. Estos espacios de experimentación, conocidos como *sandboxes* regulatorios, surgieron en principio para estimular la innovación en la industria financiera mediante una regulación más flexible y ágil; posteriormente, se han extendido a otros ámbitos de la economía digital, específicamente a los datos y la IA. Su propósito es facilitar la prueba de innovaciones a pequeña escala (productos, servicios, modelos de negocio) en un entorno controlado, similar al del mercado, suspendiendo temporalmente reglas, disposiciones o requisitos obligatorios a cambio de que se incorporen salvaguardas adecuadas para aislar al mercado de los riesgos de la innovación (Zetzsche Buckley *et al.*, 2017; FCA, 2017).

Países como Reino Unido, Noruega y Singapur han creado estos espacios de experimentación y pilotaje. En el caso de Noruega, el *sandbox* está orientado a soluciones de inteligencia artificial que usan datos personales, buscando que sean éticas y responsables; por su parte, la Oficina del Comisionado de Información del Reino Unido ofrece asesoramiento y apoyo a organizaciones que están desarrollando productos y servicios que utilizan datos personales de formas innovadoras y seguras de modo general. En el caso de Singapur, el *sandbox* es específicamente para vehículos autónomos.

En Colombia, la Consejería Presidencial para Asuntos Económicos y Transformación Digital presentó en 2020 una propuesta de modelo conceptual para el diseño de *sandboxes* regulatorios en inteligencia artificial. El modelo plantea, entre otros, un conjunto de principios para orientar a los reguladores en el diseño de este instrumento, una serie de acciones para avanzar en el diseño de un *sandbox* regulatorio transversal y las etapas mínimas de un modelo general para su diseño. Estas últimas se basan en la propuesta de la Oficina del Comisionado de Información del Reino Unido y sus etapas aparecen esquematizadas en la Figura 2.5.

Figura 2.5
Etapas para el diseño de un espacio de experimentación (*sandbox*) regulatorio

Fuente: Guío (2020a).

Agencias regulatorias

Las agencias regulatorias pueden adoptar diversas formas, responsabilidades y atribuciones. Por ejemplo, pueden limitarse a emitir directrices, códigos de conducta y mejores prácticas y atender solicitudes de asesoramiento, o pueden exigir ciertas acciones, como la realización de controles de seguridad de los sistemas de IA. Por otro lado, pueden estar enfocadas en una tecnología o grupo de tecnologías, una industria, un problema o un objetivo de política. Entre sus ventajas, se consideran: i) la vinculación de especialistas en el área específica, en lugar de profesionales «generalistas»; ii) mayor flexibilidad para realizar investigaciones independientes y tomar decisiones basadas en consideraciones sociales más amplias que los hechos específicos a los que deben sujetarse, por ejemplo, los tribunales; iii) la posibilidad de actuar de manera más proactiva que reactiva y tomar decisiones de mayor alcance (Scherer, 2016).

Una de las discusiones más importantes a la hora de definir la creación de una nueva agencia para la IA es la de su alcance. Por un lado, está la opción de crear una única agencia (transversal) a nivel nacional, que opere en todos los campos de uso y disciplinas, incluyendo el sector público y el privado. El mayor reto aquí residiría en la gran cantidad de usos y aplicaciones que puede llegar a tener la tecnología y, de ahí, los requerimientos de conocimiento especializado y contextual que puedan surgir. Así mismo, conlleva el peligro de poner todos los algoritmos dentro de una misma categoría, sin considerar niveles de riesgo y la necesidad de supervisión.

Por otro lado, está la opción de crear agencias especializadas (verticales), encargadas de la supervisión de los algoritmos en sus diferentes campos de aplicación, siguiendo el modelo usado para determinadas industrias, como la farmacéutica o de alimentos, la educación o el sistema financiero. Aquí, uno de los principales factores a contemplar sería el riesgo de traslapar responsabilidades o roles con los reguladores ya existentes a nivel sectorial. En ese caso, podría delegarse a estos la regulación de la IA en sus respectivos campos, con el inconveniente potencial de incrementar sustancialmente su carga de trabajo y exigir una capacitación intensiva del personal vinculado a temas de IA.

Otras propuestas para la regulación de la IA incluyen la creación de un órgano específico como custodio de IA, que monitoree y solicite explicaciones sobre la forma en que los algoritmos toman decisiones. Otros organismos especializados serían un Consejo Nacional de Robótica, con capacidad técnica para realizar recomendaciones; una Comisión para aprendizaje automático (*machine learning*), con capacidad tecnológica para crear sus propios algoritmos e inspeccionar el desarrollo tecnológico, pero sin poder de certificación o aprobación; o un Consejo Nacional de Seguridad de Algoritmos, con acceso a la información necesaria, directivos nombrados de forma rotatoria que sean independientes de las empresas reguladas, monitoreo constante y capacidad de hacer aplicar sus recomendaciones (Abdala *et al.*, 2019).

Las preguntas planteadas en el Error: Reference source not found pueden ayudar en el proceso de exploración y diseño del tipo de agencia que sería más adecuado de acuerdo con el contexto (WEF, 2020c).

Cuadro 2.5
Preguntas clave para explorar el diseño de opciones

Justificación para el cambio	› ¿Cuál es el contexto? › ¿Cuál es la razón fundamental para decidir crear una agencia? › ¿Qué problema se está abordando?
Visión	› ¿Cuál es la situación que se quiere a futuro? › ¿Cómo le contará esta historia a su público? › ¿Qué resultados espera lograr con la creación de esta agencia?
Poder y mandato	› ¿De dónde provendría el mandato o autoridad de la agencia? › ¿Qué poderes necesita la agencia para desempeñar sus funciones? ¿Cuál es la naturaleza y alcance de estos poderes? › ¿Involucrará a los sectores público y privado? Si lo hace, ¿cómo se abordarán las diferentes obligaciones legales y comportamientos de cada uno?
Forma y funciones	› ¿Cuál será el rol de la agencia? ¿Dónde se ubicará en el sistema actual? ¿Cuáles son sus vínculos clave? › ¿Qué actividades realizará la agencia inicialmente? ¿Cómo evolucionará con el tiempo? › ¿Qué forma tomará esta nueva agencia?
Recursos	› ¿Qué financiación y recursos se necesitan para ejecutar el nuevo modelo? › ¿De dónde vendrá este recurso? › ¿Tiene acceso a la experiencia adecuada? ¿Puede acceder a la experiencia externa?
Gobernanza y rendición de cuentas	› ¿Qué tipo de liderazgo necesita esta agencia? › ¿Cómo se responsabilizará la agencia de su trabajo? › Podría ser necesario que presente públicamente un resumen de sus actividades, incluyendo proyectos específicos desarrollados. El informe podría hacer recomendaciones para respaldar el desarrollo y la implementación de la IA. › ¿Cómo coopera la agencia con socios internacionales para alcanzar el potencial de la IA y anticipar su impacto en la humanidad?

Fuente: WEF (2020c).

Para finalizar, se puede concluir que la construcción de marcos regulatorios para la IA se encuentra en una etapa inicial, lo que presenta una oportunidad para que estos se discutan y compartan con todos los actores involucrados. Avanzar en este sentido es crucial para definir una regulación centrada en la ética, la seguridad, la justicia y la transparencia, que fortalezca la confianza de la ciudadanía en la tecnología y su uso por parte del Estado. Los avances logrados requerirán además revisiones y control público frecuente para mantener su relevancia, aplicabilidad y efectividad.

Bibliografía

Abdala, M., Eussler, S. L. y Soubie, S. (2019). «La política de la inteligencia artificial: sus usos en el sector público y sus implicaciones regulatorias». *Documento de trabajo n.º 185*. CIPPEC. https://www.cippec.org/wp-content/uploads/2019/10/185-DT-Abdala-Lacroix-y-Soubie-La-pol%C3%ADtica-de-la-Inteligencia-Artificial-octubre-2019.pdf.

AI Multiple. (2021, febrero 13). Bias in AI: What it is, Types & Examples, How & Tools to fix it. https://research.aimultiple.com/ai-bias/

Barredo-Arrieta, A., Días-Rodríguez, N., Del Serc, J., Bennetot, A., Tabik, S., Barbado, A. y García, S. (2020). «Explainable artificial intelligence (XAI): concepts, taxonomies, opportunities and challenges toward responsible AI». *Information Fusion*, 58, 82-115. https://doi.org/10.1016/j.inffus.2019.12.012.

Bebis, G., Egbert, D. y Shah, M. (2003). «Review of computer vision education». *IEEE Xplore* [en línea]. Transactions on Education, 46(1), 2-21. Recupeardo de: https://doi.org/10.1109/TE.2002.808280.

Berryhill, J., Heang, K. K., Clogher, R. y McBride, K. (2019). «Hello, world: artificial intelligence and its use in the public sector». *OECD Working Papers on Public Governance*, n.º 36. https://doi.org/10.1787/726fd39d-en.

Bird, E., Fox-Skelly, J., Jenner, N., Larbey, R. y Weitkamp, E. (2020). *The ethics of artificial intelligence: Issues and initiatives*. Parlamento Europeo. http://op.europa.eu/en/publication-detail/-/publication/3a046f26-88f7-11ea-812f-01aa75ed71a1/language-en.

Bostrom, N. y Yudkowsky, E. (2014). «The ethics of artificial intelligence». En K. Frankish y W. Ramsey (eds.), *The Cambridge handbook of artificial intelligence* (pp. 316-334). Cambridge: Cambridge University Press. doi:10.1017/CBO9781139046855.020.

Brookfield Institute (2018). *Intro to AI for policymakers: understanding the shift*. https://brookfieldinstitute.ca/intro-to-ai-for-policymakers.

Centre for Public Impact (2017). *Destination unknown: exploring the impact of Artificial Intelligence on government*. https://www.centreforpublicimpact.org/ai-government-working-paper/.

CEPAL (2018a). *Monitoreo de la agenda digital para América Latina y el Caribe eLAC2018* (CEPAL). Santiago. https://www.cepal.org/es/publicaciones/43444-monitoreo-la-agenda-digital-america-latina-caribe-elac2018.

Chui, M. y McCarthy, B. (2020). «An executive's guide to AI». McKinsey & Company [en línea]. https://www.mckinsey.com/business-functions/mckinsey-analytics/our-insights/an-executives-guide-to-ai.

Clausen, K. G. (2020, marzo 11). «How to tackle bias in AI». 2021.AI [en línea]. https://2021.ai/how-to-tackle-bias-in-ai/.

Collingridge, D. (1980). *The social control of technology*. Printer. https://www.cambridge.org/core/article/social-control-of-technology-by-david-collingridge-new-york-st-martins-press-1980-pp-i-200-2250/648B7ECDDB00120BCAB13F17E4076C08.

Comisión Europea (2021). «Proposal for a regulation laying down harmonised rules on artificial intelligence. Shaping Europe's digital future». COM/2021/206 final. https://digital-strategy.ec.europa.eu/en/library/proposal-regulation-laying-down-harmonised-rules-artificial-intelligence.

Curry, E., Freitas, A., Thalhammer, A., Fensel, A., Ngonga, Ermilov, I. ... y Ui Hassan, U. (2012). «Big data technical groups». White paper. BIG Consortium.

DAMA International (2017). *DAMA-DMBOK: Data management body of knowledge*. Technics Publications.

Davenport, T. H. y Prusak, L. (2000). «Working knowledge: how organizations manage what they know». Ubiquity, vol. 2000, n.º agosto. Recuperado de: https://doi.org/10.1145/347634.348775.

Desouza K. (2018). *Delivering artificial intelligence in government: challenges and opportunities*. IBM Center for Business of Government. http://www.businessofgovernment.org/sites/default/files/Delivering%20Artificial%20Intelligence%20in%20Government.pdf.

Eggers W., Schatsky D. y Viechnicki P (2017b). «How artificial intelligence can transform the government. Executive Summary». *Deloitte Insights*. https://www2.deloitte.com/us/en/insights/focus/cognitive-technologies/artificial-intelligence-government-summary.html.

Eggers, W. D., Schatsky, D. y Viechnicki, P. (2017a). *AI-augmented government. Using cognitive technoloogies to redesign public sector work*. Deloitte University Press. https://www2.deloitte.com/content/dam/insights/us/articles/3832_AI-augmented-government/DUP_AI-augmented-government.pdf.

Elish, M. C. y Watkins, E. A. (2019). «When humans attack». *Points. Daya & Society*. 14 de mayo de 2019. https://points.datasociety.net/when-humans-attack-re-thinking-safety-security-and-ai-b7a15506a115.

FCA (2017). Regulatory sandbox lessons learned report. Financial Conduct Authority. https://www.fca.org.uk/publication/research-and-data/regulatory-sandbox-lessons-learned-report.pdf.

Fjeld, J., Achten, N., Hilligoss, H., Nagy, A., & Srikumar, M. (2020). Principled Artificial Intelligence: Mapping Consensus in Ethical and Rights-Based Approaches to Principles for AI. SSRN Electronic Journal. https://doi.org/10.2139/ssrn.3518482

Fjelland, R. (2020). «Why general artificial intelligence will not be realized». *Humanities and Social Sciences Communications*, 7(1), 1-9. https://doi.org/10.1057/s41599-020-0494-4.

Floridi, L., & Cowls, J. (2019). A Unified Framework of Five Principles for AI in Society. Harvard Data Science Review, 1(1). https://doi.org/10.1162/99608f92.8cd550d1

GDS y OAI (2019). *A guide to using AI in the public sector*. Government Digital Service y Office for Artificial Intelligence. Actualizado el 18 de octubre de 2021. https://www.gov.uk/government/collections/a-guide-to-using-artificial-intelligence-in-the-public-sector.

Gobierno de Canadá (2019, febrero 5). «Directive on automated decision-making». *Government of Canada* [en línea]. How government Works. Policies, directives, standards and guidelines. Treasury Board of Canada Secretariat. https://www.tbs-sct.gc.ca/pol/doc-eng.aspx?id=32592.

Gobierno de Nueva Zelanda (2020, julio). «Algorithm charter for Aotearoa New Zealand». *Data.Govt.Nz*. https://data.govt.nz/use-data/data-ethics/government-algorithm-transparency-and-accountability/algorithm-charter.

Guío, A. (2020). Modelo conceptual para el diseño de regulatory sandboxes & beaches en inteligencia artificial. Documento borrador para discusión. Consejería Presidencial para Asuntos Económicos y Transformación Digital de Colombia. Recuperado de: https://dapre.presidencia.gov.co/AtencionCiudadana/DocumentosConsulta/consulta-200820-MODELO-CONCEPTUAL-DISENO-REGULATORY-SANDBOXES-BEACHES-IA.pdf

Grupo Independiente de Expertos de Alto Nivel sobre Inteligencia Artificial (2019). *Directrices técnicas para una IA fiable*. Comisión Europea. https://ialab.com.ar/wp-content/uploads/2020/06/Grupo-independiente-de-expertos-de-alto-nivel-sobre-IA-creado-por-la-Comisio%CC%81n-Europea.pdf.

Guillén, M. A., López Ayuso, B., Paniagua, E. y Cadenas, J. M. (2015). «Una revisión de la cadena datos-información-conocimiento desde el pragmatismo de Peirce». Documentación de las Ciencias de la Información, 38(0), 153–177. https://doi.org/10.5209/rev_dcin.2015.v38.50814.

How To Governance—RRI Tools. (s. f.). Recuperado 20 de septiembre de 2020, de https://www.rri-tools.eu/es/how-to-pa-governance#menu-anchor-id3-content

IEEE (2015). «Speech and Language Processing Technical Committee». *IEEE Signal Processing Society*. https://signalprocessingsociety.org/community-involvement/speech-and-language-processing.

IEEE (2019). *Artificial intelligence. IEEE position statement*. https://globalpolicy.ieee.org/wp-content/uploads/2019/06/IEEE18029.pdf.

IRISS (s. f.). *Future risk and opportunities toolkit*. Institute for Research and Innovation in Social Services. https://www.iriss.org.uk/sites/default/files/future_risk_and_opportunities_card_pack.pdf.

Janssen, M., Brous, P., Estevez, E., Barbosa, L. S. y Janowski, T. (2020). «Data governance. Organizing data for trustworthy artificial intelligence». *Government Information Quarterly*, 37(3), 101493. https://doi.org/10.1016/j.giq.2020.101493.

Khtira, R., Elasri, B. y Rhanoui, M. (2017). «From data to big data: Moroccan public sector». En *ACM International Conference Proceeding Series*, art. n.º 46, pp. 1-6. New York, NY: Association for Computing Machinery. https://doi.org/10.1145/3090354.3090401.

Kuhlmann, S., Stegmaier, P. y Konrad, K. (2019). «The tentative governance of emerging science and technology – A conceptual introduction». *Research Policy*, 48, 1091-1097. Elsevier. https://doi.org/10.1016/j.respol.2019.01.006.

Kuziemski, M. y Misuraca, G. (2020). «AI governance in the public sector: three tales from the frontiers of automated decision-making in democratic settings». *Telecommunications Policy*, 44(6), 101976. https://doi.org/10.1016/j.telpol.2020.101976.

Leslie, D. (2019). *Understanding artificial intelligence ethics and safety. A guide for the responsible design and implementation of AI systems in the public sector*. The Alan Turing Institute. https://doi.org/10.5281/zenodo.3240529.

McKenna, M. (s. f.). «Machines and trust: how to mitigate AI bias». *Toptal Engineering* [en línea]. Blog. https://www.toptal.com/artificial-intelligence/mitigating-ai-bias (consulta realizada el 3 de marzo de 2021).

Mehrabi, N., Morstatter, F., Saxena, N., Lerman, K. y Galstyan, A. (2019, octubre). «A survey on bias and fairness in machine learning». *ArXiv* [en línea]. Cornell University. https://arxiv.org/abs/1908.09635.

MIRI. (S.F). Why AI Safety? Machine Intelligence Research Institute. https://intelligence.org/why-ai-safety/

Myers, D. (2019). *2019 Annual report on the dimensions of data quality. Year five-opportunity abounds for competitive advantage based on information quality*. DQ Matters. http://dimensionsofdataquality.com/download/prior_whitepapers/2019-Annual-Report-on-the-Dimensions-of-Data-Quality1008.pdf?doc_num=1008&src=1008fromcddqsite.

Naser, A. (2008). Gobierno electrónico y gestión pública. Ilpes/Cepal.

Needham, M. (2018, noviembre 12). Graph Algorithms in Neo4j: Graph Technology & AI Applications. Neo4j Graph Database Platform. https://neo4j.com/blog/graph-algorithms-neo4j-graph-technology-ai-applications/

NITI Aayog. (2018). *National Strategy for Artificial Intelligence #AIFORALL*. https://niti.gov.in/writereaddata/files/document_publication/NationalStrategy-for-AI-Discussion-Paper.pdf 02/02/2021.

Ntoutsi, E., Fafalios, P., Gadiraju, U., Iosifidis, V., Nejdl, W., Vidal, M.-E., ... y Staab, S. (2020). «Bias in data-driven artificial intelligence systems. An introductory survey». *WIREs Data Mining and Knowledge Discovery*, 10(3), e1356. https://doi.org/10.1002/widm.1356.

Nugroho, R. P., Zuiderwijk, A., Janssen, M. y de Jong, M. (2015). «A comparison of national open data policies: lessons learned». *Transforming government: people, process and policy*, 9(3), 286–308. https://doi.org/10.1108/TG-03-2014-0008.

OCDE (2018a). *OECD science, technology and innovation outlook 2018: adapting to technological and societal disruption*. OCDE. https://doi.org/10.1787/sti_in_outlook-2018-en.

OCDE (2019a). *Data in the digital age*. París: OECD Publishing. https://doi.org/10.1787/99b9ba70-en.

OCDE (2019b). «Using digital technologies to improve the design and enforcement of public policies». *OECD Digital Economy Papers*, n.º 274). https://dx.doi.org/10.1787/99b9ba70-en.

OECD. (2019b). The Pattreasuryh to Becoming a Data-Driven Public Sector. OECD. https://doi.org/10.1787/059814a7-en

OCDE (2019c). Artificial intelligence in society. París: OECD Publishing. https://doi.org/10.1787/eedfee77-en.

OCDE (2019d). «Government at a glance 2019 edition. Open government data». OECD.Stats [base de datos]. https://stats.oecd.org/Index.aspx?DataSetCode=GOV (consulta realizada el 21 de junio de 2021).

OCDE (2021a). *OECD good practice principles for data ethics in the public sector*. OECD Publishing. https://www.oecd.org/gov/digital-government/good-practice-principles-for-data-ethics-in-the-public-sector.htm.

ODLIS. (2020). ABC-CLIO. Retrieved November 17, 2020, from https://products.abc-clio.com/ODLIS/odlis_d.aspx

RRI Tools (s. f.). «How to governance». *Responsible Research and Innovation* [en línea]. Disponible en https://www.rri-tools.eu/es/how-to-pa-governance#menu-anchor-id3-content.

SAE (2018). «Taxonomy and definitions for terms related to driving automation systems for on-road motor vehicles». *SAE International* [en línea]. https://www.sae.org/standards/content/j3016_201806/.

Scherer, M. (2016). «Regulating artificial intelligence systems: risks, challenges, competencies, and strategies». *Harvard Journal of Law & Technology*, 29(2), 353-400. https://doi.org/10.2139/ssrn.2609777.

Shin, T. (2020, junio). «Real-life examples of discriminating artificial intelligence». *Towards data science*. https://towardsdatascience.com/real-life-examples-of-discriminating-artificial-intelligence-cae395a90070.

Silberg, J. y Manyika, J. (2019, junio). *Notes from the AI frontier: tackling bias in AI (and in humans)*. MckKinsey Global Institute. https://www.mckinsey.com/~/media/mckinsey/featured%20insights/artificial%20intelligence/tackling%20bias%20in%20artificial%20intelligence%20and%20in%20humans/mgi-tackling-bias-in-ai-june-2019.pdf.

Spielkamp, M. (2017, junio). «Inspecting algorithms for bias». *MIT Technology Review*. https://www.technologyreview.com/2017/06/12/105804/inspecting-algorithms-for-bias/.

Stilgoe, J., Owen, R. y Macnaghten, P. (2013). «Developing a framework for responsible innovation». *Research Policy*, 42, 1568-1580. http://dx.doi.org/10.1016/j.respol.2013.05.008.

Stirling, A., Chubb, J., Montana, J., Stilgoe, J. y Wilsdon, J. (2019). *A review of recent evidence on the governance of emerging science and technology*. Comissionmed by the Wellcome Trust. https://wellcome.org/sites/default/files/evidence-review-governance-emerging-science-and-technology_0.pdf.

Stone, P., Brooks, R. y Brynjolfsson, E. (2016). «2016 Report | One hundred year study on artificial intelligence». *AI100*. https://ai100.stanford.edu/2016-report.

Sucar, L.E. (2015). Probabilistic Graphical Models: principles and applications. Springer-Verlag. https://doi.org/10.1007/978-1-4471-6699-3

Suresh, H. y Guttag, J. V. (2020). A framework for understanding unintended consequences of machine learning. *ArXiv* [en línea]. Cornell University. http://arxiv.org/abs/1901.10002.

Ubaldi, B., Le Febre, E. M., Petrucci, E., Marchionni, P., Biancalana, C., Hiltunen, N., Intravaia, D. M. y Yang, C. (2019). «State of the art in the use of emerging technologies in the public sector». *OECD Working Papers on Public Governance*, n.º 31; https://doi.org/10.1787/932780bc-en.

Valle-Cruz, D., Criado, I., Sandoval-Almazán, R. y Ruvalcaba-Gómez, E. A. (2020). «Assessing the public policy-cycle framework in the age of artificial intelligence. From agenda-setting to policy evaluation». *Government Information Quarterly*, 37(4). https://doi.org/10.1016/j.giq.2020.101509.

Van Ooijen, C. Ubaldi, B. y Welby, B. (2019). «A data-driven public sector: enabling the strategic use of data for productive, inclusive and trustworthy governance». *OECD Working Papers on Public Governance* n.º 33. https://doi.org/10.1787/09ab162c-en

WEF (2018a). *Agile governance: reimagining policy-making in the Fourth Industrial Revolution*. Foro Económico Mundial. https://www.weforum.org/whitepapers/agile-governance-reimagining-policy-making-in-the-fourth-industrial-revolution/.

WEF (2019a). *AI governance: a holistic approach to implement ethics into AI*. Foro Económico Mundial. https://www.weforum.org/whitepapers/ai-governance-a-holistic-approach-to-implement-ethics-into-ai/.

WEF (2020a). *AI procurement in a box. Procurement guidelines*. Foro Económico Mundial. https://www.weforum.org/reports/ai-procurement-in-a-box/.

WEF (2020b). *The future of jobs report 2020*. Foro Económico Mundial. https://www.weforum.org/reports/the-future-of-jobs-report-2020/.

WEF (2020c). *Reimagining regulation for the age of AI: New Zealand pilot project*. Foro Económico Mundial. https://www.weforum.org/whitepapers/reimagining-regulation-for-the-age-of-ai-new-zealand-pilot-project/.

WEF, IMDA y PDPC. (2020). *Companion to the model AI governance framework —Implementation and self-assessment guide for organizations*. Foro Económico Mundial. https://www.pdpc.gov.sg/-/media/Files/PDPC/PDF-Files/Resource-for-Organisation/AI/SGIsago.pdf.

Yurtsever, E., Lambert, J., Carballo, A. y Takeda, K. (2020). «A survey of autonomous driving: common practices and emerging technologies». *IEEE Access*, 8, 58443-58469. https://doi.org/10.1109/ACCESS.2020.2983149.

Zafar, F., Khan, A., Suhail, S., Ahmed, I., Hameed, K., Khan, H. M., ... y Anjum, A. (2017). «Trustworthy data: a survey, taxonomy and future trends of secure provenance schemes». *Journal of Network and Computer Applications*, 94, 50–68. https://doi.org/10.1016/j.jnca.2017.06.003

Zetzsche, D. A., Buckley, R. P., Arner, D. W. y Barberis, J. N. (2017). «Regulating a revolution: from regulatory sandboxes to smart regulation». *SSRN Electronic Journal*. https://doi.org/10.2139/ssrn.3018534.

Conceptos fundamentales y uso responsable de la

INTELIGENCIA ARTIFICIAL EN EL SECTOR PÚBLICO

CAF BANCO DE DESARROLLO DE AMÉRICA LATINA