Guía de implementación de proyectos de inteligencia artificial en el sector público

NACIONES UNIDAS

CEPAL

CEPAL
LABORATORIO
DE TRANSFORMACIÓN
DIGITAL

cooperación
alemana
DEUTSCHE ZUSAMMENARBEIT

CEPAL

Si desea recibir información oportuna sobre nuestros productos editoriales y actividades, le invitamos a registrarse. Podrá definir sus áreas de interés y acceder a nuestros productos en otros formatos.

Deseo registrarme

Conozca nuestras redes sociales y otras fuentes de difusión en el siguiente link:

https://bit.ly/m/CEPAL

Guía de implementación
de proyectos de inteligencia artificial en el sector público

Naciones Unidas
CEPAL

CEPAL
LABORATORIO
DE TRANSFORMACIÓN
DIGITAL

cooperación
alemana
DEUTSCHE ZUSAMMENARBEIT

cooperación CEPAL-BMZ/GIZ.

Las Naciones Unidas y los países que representan no son responsables por el contenido de vínculos a sitios web externos incluidos en esta publicación.

Las opiniones expresadas en este documento, que no ha sido sometido a revisión editorial, son de exclusiva responsabilidad del autor y pueden no coincidir con las de la Organización o las de los países que representa.

Publicación de las Naciones Unidas
LC/TS.2026/33/-*
Distribución: L

Impreso en Naciones Unidas, Santiago
S.2600064[S]

Esta publicación debe citarse como: Clavijo, A. (2026). *Guía de implementación de proyectos de inteligencia artificial en el sector público* (LC/TS.2026/33/-*). Comisión Económica para América Latina y el Caribe.

Diagramas

La inteligencia artificial se ha consolidado como una tecnología de alto impacto con capacidad para transformar la economía y la vida diaria, al automatizar tareas, procesar grandes volúmenes de datos y mejorar la toma de decisiones. Esto ha captado la atención de los gobiernos y entidades públicas a nivel internacional, dado que las entidades públicas pueden aprovechar los beneficios de la IA y abordar problemas como automatizar procesos de admirativos o agilizar la gestión documental.

La adopción de IA en el sector público avanza con rapidez, impulsada por expectativas de eficiencia, mejor atención y reducción de costes. Pero su despliegue plantea retos técnicos y de gobernanza, como son la transparencia, legitimidad, rendición de cuentas, equidad y valor público, que no se resuelven solo con ingeniería.

La evidencia sintetizada por el Ada Lovelace Institute (Ada Lovelace Institute, s.f.) tras seis años de investigación en casos reales muestra que el éxito depende de: contextualizar la IA en sistemas sociotécnicos, aprender de forma estructurada qué funciona, alinear el despliegue con los valores y expectativas del sector público y la ciudadanía, y reformar compras y garantías para herramientas basadas (cada vez más) en modelos fundacionales. Estas lecciones subrayan déficits actuales de visibilidad sobre dónde se usa la IA, de evidencia comparativa de efectividad y de mecanismos de aseguramiento independientes (tests, monitorización postmercado, reparación).

La adopción de la inteligencia artificial por parte de las entidades públicas debe darse a través de factores habilitantes. Autores como Grimmelikhuijsen y Tangi (2024) mencionan que los factores que influyen en dicha adopción incluyen la tecnología disponible, una cultura abierta a la innovación, una estrategia clara para el uso de la IA y percepciones positivas hacia su implementación. De manera similar, Madan y Ashok (2023) señalan que el contexto tecnológico, la cultura organizacional y la capacidad de adopción son elementos importantes que inciden en el proceso. Asimismo, destacan la relevancia de la seguridad y la gobernanza de los datos, así como la necesidad de contar con equipos capacitados para comprender y manejar herramientas basadas en IA.

marcos teóricos a decisiones operativas, medibles y auditables en el sector público. La propuesta de esta guía aterriza esa evidencia en una ruta práctica de 3 fases y 10 pasos con preguntas orientadoras, puertas de decisión y entregables mínimos en cada etapa. Ayuda a ubicar cualquier proyecto —incipiente o avanzado— en la madurez real de la entidad y a conectar decisiones: define el problema y el valor público a lograr, compara IA vs. alternativas noIA, fija métricas técnicas y de equidad, establece límites de uso y salvaguardas, y documenta trazabilidad. Así resuelve tres vacíos habituales: ¿por dónde empezar?, ¿con qué evidencia avanzar? y ¿cómo gestionar riesgos y responsabilidades?; y crea un lenguaje común entre política pública, datos, TI, jurídico, compras y comunicación.

Esta guía se desarrolló en el marco del Laboratorio de Transformación Digital de CEPAL. El Laboratorio es un espacio de experimentación, aprendizaje e innovación colaborativa que reúne a formuladores de políticas públicas, especialistas en tecnología y académicos para co-crear metodologías, herramientas prácticas y soluciones tecnológicas que permitan a los gobiernos enfrentar de manera más efectiva sus desafíos de digitalización. El Laboratorio nace como parte del proyecto Transformación Digital para la Integración Regional (D4R), implementado conjuntamente por la Comisión Económica para América Latina y el Caribe (CEPAL) y la Cooperación Alemana (GIZ).

¿Para quién es esta guía?

Esta guía está dirigida a organizaciones públicas de todos los niveles —nacional, regional y local, incluidas empresas públicas— y a los roles clave que intervienen en proyectos de IA: directivos y responsables que definen propósito y valor público; equipos técnicos (datos, TI, ciencia de datos, producto) que diseñan y operan los sistemas; funciones habilitadoras (contratación, asesoría jurídica, privacidad y protección de datos, ciberseguridad, finanzas, RR. HH.) que aseguran cumplimiento y sostenibilidad; órganos de control y garantía (auditoría interna/externa, comités de ética, defensores del pueblo/reguladores) que supervisan riesgos y resultados; y profesionales de primera línea que interactúan con la ciudadanía y usan las herramientas en su trabajo diario. También es útil para proveedores y socios tecnológicos que quieran alinearse con estándares del sector público y para unidades de comunicación y participación que gestionan la licencia social de estos proyectos.

riesgos que no aparecen en los documentos formales.

¿Dónde empiezo?

- Si estás iniciando o tienes dudas de fundamentos, comienza por Fase I: Explorar (pasos 1-5).
- Si ya tienes prototipo/proyecto en marcha, realiza primero la lista de chequeo rápido Exploración.
  - Si hay casillas sin cumplir, vuelve a Fase I para cerrar brechas.
  - Si todo está en verde, continúa en Fase II: Desarrollar (pasos 6-8).
- Si ya operas un modelo/piloto productivo, entra a Fase III: Operar (pasos 9-10) y verifica garantías, monitoreo y rendición de cuentas.

Checklist

☐ Problema y valor público definidos.
☐ Patrocinio directivo y roles claros.
☐ Métrica primaria de éxito y criterios Go/NoGo acordados.
☐ Inventario de datos con base legal/legitimidad y calidad mínima viable.
☐ Principios y límites de uso (privacidad, equidad, seguridad, propósito).
☐ Mapa de riesgos y mitigaciones iniciales.
☐ Plan de experimentación (hipótesis, alcance, recursos).
☐ Equipo multidisciplinario conformado y calendario de trabajo.

Existen tres tipos de inteligencia artificial:

i) **Inteligencia artificial estrecha (*Artificial narrow intelligence*).** Es la capacidad de un sistema informático para realizar una tarea definida de manera limitada mejor que un humano.

   La IA estrecha es el nivel más avanzado de desarrollo de IA que la humanidad ha alcanzado hasta ahora, y todos los ejemplos de IA que se ven en el mundo real pertenecen a esta categoría —incluyendo vehículos autónomos y asistentes digitales personales. Esto se debe a que, incluso cuando parece que la IA está pensando por sí misma en tiempo real, en realidad está coordinando varios procesos limitados y tomando decisiones dentro de un marco predefinido. El "pensamiento" de la IA no implica conciencia ni emoción (Microsoft Azure, s.f.).

ii) **Inteligencia artificial general (*Artificial general intelligence*).** Hace referencia a la capacidad de un sistema informático para superar a los humanos en cualquier tarea intelectual. Es el tipo de IA que se ve en películas donde los robots tienen pensamientos conscientes y actúan por sus propios motivos.

   En teoría, un sistema informático que haya alcanzado la IA general sería capaz de resolver problemas profundamente complejos, aplicar juicio en situaciones inciertas e incorporar conocimientos previos en su razonamiento actual. Sería capaz de creatividad e imaginación al nivel de los humanos y podría asumir una gama mucho más amplia de tareas que la IA estrecha (Microsoft Azure, s. f.).

iii) **Inteligencia artificial superinteligente (*Artificial superintelligence*).** Es un sistema informático superinteligente tendría la capacidad de superar a los humanos en casi todos los campos, incluyendo la creatividad científica, la sabiduría general y las habilidades sociales (Microsoft Azure, s. f.).

de entrada están emparejados con los datos de salida correctos. El algoritmo aprende a partir de estos datos y realiza predicciones generalizando desde el conjunto de entrenamiento a situaciones no vistas.

Se puede aplicar en clasificación, donde la variable de salida es una categoría, como "spam" o "no spam". También como regresión, donde la variable de salida es un valor real, como "peso" o "precio" (Agarwal, 2025).

ii) **Aprendizaje no supervisado.** Este implica entrenar algoritmos con datos sin respuestas etiquetadas. Aquí, el sistema intenta aprender los patrones y la estructura de los datos sin ninguna referencia a resultados conocidos o etiquetados.

Su aplicación está en agrupamiento (clustering), que consiste en agrupar entidades similares como en la segmentación de mercados.

iii) **Asociación (*association*)**. Es un método basado en reglas para descubrir relaciones interesantes entre variables en grandes bases de datos, como para el análisis de la cesta de la compra (Agarwal, 2025).

iv) **Aprendizaje semisupervisado**. Este se sitúa entre el aprendizaje supervisado y el no supervisado. En este enfoque, los algoritmos se entrenan con un *dataset* que contiene datos etiquetados y no etiquetados. Típicamente, hay una pequeña cantidad de datos etiquetados y muchos datos no etiquetados. Los sistemas usan los datos etiquetados para aprender la estructura y hacer inferencias sobre cómo etiquetar los datos no etiquetados.

Su aplicación está en autoaprendizaje (*self-training*): un método en el que un modelo se entrena inicialmente con un pequeño conjunto de datos etiquetados y luego se usa para clasificar datos no etiquetados. Las predicciones de mayor confianza se añaden al conjunto de entrenamiento y el modelo se reentrena; por ejemplo, reconocimiento de imágenes y voz cuando solo algunos ejemplos están etiquetados.

Los problemas que aborda incluyen métodos basados en valor (*value-based*), que tienen como objetivo maximizar el valor de la función, que representa la recompensa acumulada máxima esperada alcanzable desde un estado, con aplicaciones como decisiones en tiempo real en software, por ejemplo, motores de recomendación. También incluye métodos basados en políticas (*policy-based*), que aprenden directamente la función de política que mapea estado a acción sin necesitar una función de valor, con aplicaciones como juegos y robótica (Agarwal, 2025).

Aprendizaje Profundo (*Deep Learning*). Es un tipo de aprendizaje automático. Sin embargo, el aprendizaje profundo puede analizar más tipos de información y ejecutar operaciones más complejas. El proceso en el que se asienta el aprendizaje profundo se inspira en la estructura y el funcionamiento del cerebro humano, concretamente en la manera en que las neuronas se interconectan y cooperan para procesar la información. Permite hacer predicciones más matizadas y profundas a partir de los datos (International Organization for Standardization, s. f.).

Para entender cómo logra esa capacidad, empecemos por sus piezas estructurales.

- **Algoritmo**. Lista precisa de pasos a seguir, como un programa informático. Los sistemas de IA contienen algoritmos, pero generalmente solo para algunas partes, como métodos de aprendizaje o de cálculo de recompensas (Stanford HAI, 2023).
- **Parámetros (*Parameters*)**. Valores numéricos que configuran y gobiernan un modelo (p. ej., pesos aprendidos y aspectos de arquitectura y operación) y condicionan su comportamiento y salidas (MIT Sloan EdTech, s. f.).
- **Red neuronal artificial**. Estructura informática inspirada en el cerebro biológico, formada por un gran conjunto de unidades computacionales interconectadas (“neuronas”) organizadas en capas. Los datos pasan entre estas unidades como entre neuronas en un cerebro, las neuronas artificiales procesan la información en conjunto, cada neurona artificial, o nodo, emplea cálculos matemáticos para procesar información y resolver problemas complejos. (Amazon Web Services, s. f.).

Con estas bases, surgen familias de modelos y técnicas que potencian el lenguaje y el razonamiento:

- **Modelo Transformer (*Transformer model*)**. Arquitectura que procesa oraciones en paralelo con mecanismos de atención para captar contexto y dependencias de largo alcance en el lenguaje (MIT Sloan EdTech, s. f.).
- **Gran modelo de lenguaje (*Large Language Model, LLM)***. Red neuronal que predice secuencias de palabras; puede dialogar, redactar y analizar grandes volúmenes de texto (MIT Sloan EdTech, s. f.).
- **Modelo preentrenado (*Pretrained model*)**. Es un modelo de machine learning que se ha entrenado previamente con un gran conjunto de datos para una tarea específica (normalmente de uso general) y que después se puede reutilizar o ajustar para una tarea diferente, pero relacionada (IBM, s. f.).
- **Ajuste fino (*Fine-tuning*)**. El ajuste fino permite adaptar los modelos de IA preentrenados para que funcionen mejor con tus datos y casos de uso específicos. Esta técnica puede mejorar el rendimiento del modelo y, al mismo tiempo, requiere menos datos de entrenamiento que la creación de un modelo desde cero (Microsoft, s. f.).
- **Aprendizaje por transferencia (*Transfer Learning*)**. Es una técnica de machine learning en la que el conocimiento adquirido a través de una tarea o conjunto de datos se utiliza para mejorar el rendimiento del modelo en otra tarea relacionada y/o diferente conjunto de datos (IBM, s. f.).
- **Modelo fundacional (*Foundation model*)**. Modelo de aprendizaje automático entrenado con una gran cantidad de datos para poder adaptarse con facilidad a una amplia gama de aplicaciones. Un tipo común de modelo fundacional son los grandes modelos de lenguaje, que impulsan chatbots como ChatGPT (The Alan Turing Institute, s. f.).
- **Generación aumentada con recuperación (*Retrieval Augmented Generation*, RAG)**. Método que combina un modelo de lenguaje con fuentes externas recuperadas (documentos, PDFs, bases de conocimiento) para respuestas más precisas y fundamentadas (MIT Sloan EdTech, s. f.).

se utilizan principalmente para la generación de imágenes y otras tareas de visión artificial. Las redes neuronales basadas en difusión se entrenan mediante aprendizaje profundo para "difundir" progresivamente muestras con ruido aleatorio y, posteriormente, revertir ese proceso de difusión para generar imágenes de alta calidad. (IBM, s. f.).

Los modelos dependen de los datos con los que alimentamos y evaluamos al modelo:

- **Datos estructurados (*Structured Data*)**. Los datos estructurados tienen un esquema fijo y encajan perfectamente en filas y columnas, como nombres y números de teléfono (IBM, s. f.).
- **Datos no estructurados (*Unstructured Data*)**. Los datos no estructurados no tienen un esquema fijo y pueden tener un formato más complejo, como archivos de audio y páginas web (IBM, s. f.).
- **Datos sintéticos**. Datos generados artificialmente, no por eventos del mundo real. Estos conservan las propiedades estadísticas del conjunto de datos original, es decir, su distribución o propiedades estadísticas, además pueden la privacidad (The Alan Turing Institute, s. f.).
- **Datos de entrenamiento**. Datos utilizados para entrenar un modelo de aprendizaje automático (International Organization for Standardization, s. f.).

Para producir contenido, los modelos generativos usan instrucciones y controles de variación:

- **IA generativa**. La IA generativa se refiere a los sistemas de inteligencia artificial que crean nuevos contenidos y artefactos, como imágenes, videos, texto y audio, a partir de simples peticiones de texto. A diferencia de la IA del pasado, que se limitaba al análisis de datos, la IA generativa aprovecha el aprendizaje profundo y los conjuntos de datos masivos para producir resultados creativos de alta calidad y similares a los producidos por humanos (Amazon Web Services, s. f.).
- ***Prompt***: entrada o instrucción que el usuario proporciona a un sistema de IA para obtener un resultado deseado (Coursera, s. f.).

redacción de *prompts* que induce al modelo a razonar paso a paso para mejorar exactitud y aplicabilidad, útil en matemática, lógica y decisiones multietapa (MIT Sloan EdTech, s. f.).

- **Temperatura (*Temperature*)**. Parámetro que controla cuán determinista o creativo es el output; baja = más consistencia, alta = más variación (MIT Sloan EdTech, s. f.).

Sus capacidades se materializan en aplicaciones como:

- **Procesamiento de lenguaje natural (NLP).** Subcampo de la IA y la lingüística computacional que permite a las máquinas comprender, interpretar y generar lenguaje humano (MIT Sloan EdTech, s. f.).
- **La visión artificial (*Computer Vision*)**. La visión artificial (CV, por sus siglas en inglés) es un proceso capaz de captar, procesar y analizar imágenes reales para permitir a las máquinas obtener información contextual significativa del mundo físico (Gartner, s. f.).
- **IA multimodal**. Modelos capaz de procesar y generar múltiples tipos de entrada/salida (texto, imágenes, audio, video), p. ej., describir imágenes o generar código desde diagramas (MIT Sloan EdTech, s. f.).
- **Agentes**. Entidades de IA autónomas o semi-autónomas que ejecutan tareas, toman decisiones y llaman herramientas o APIs según objetivos (p. ej., resumir documentos, enrutar tareas, razonamiento multietapa) (MIT Sloan EdTech, s. f.).
- Chatbot. Aplicación que imita la conversación humana mediante texto o voz (Coursera, s. f.).

Con esto emergen retos de precisión, ética y gobierno:

- **Alucinación (*Hallucination*)**. Situación en la que un sistema de IA produce información fabricada, sin sentido o inexacta. La información errónea se presenta con seguridad, lo que puede dificultar que el usuario humano determine si la respuesta es confiable (Carnegie Mellon University, Heinz College, 2023).

espera y de acuerdo con los valores y objetivos humanos (IBM, s. f.).

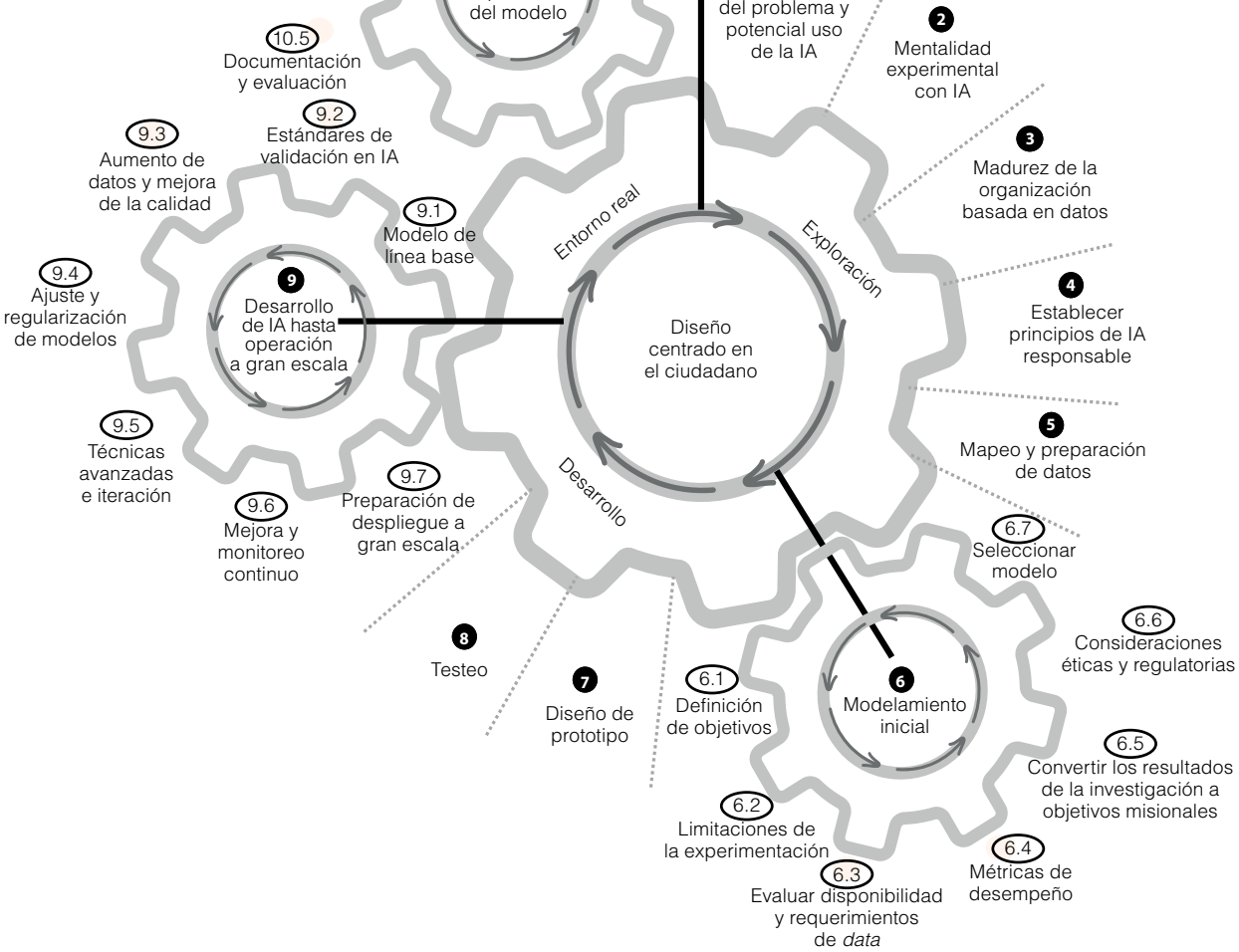
- **Explicabilidad**. Propiedad de un sistema de IA de expresar factores importantes que influyen en sus resultados de manera comprensible para humanos (International Organization for Standardization, s. f.).
- **Interpretabilidad**. La interpretabilidad a las personas a comprender y explicar mejor los procesos de toma de decisiones que potencian los modelos de inteligencia artificial (IA) (IBM, s. f.).
- **Comportamiento emergente (*Emergent behavior*)**. Habilidades inesperadas que surgen en modelos de gran escala (p. ej., programar, componer, escribir poesía o narrativa) (MIT Sloan EdTech, s. f.).

Fases de implementación de Proyectos IA

El concepto de Niveles de Madurez Tecnológica (*Technology Readiness Levels - TRL*) se utiliza para evaluar el grado de desarrollo y madurez de una tecnología, desde su concepción inicial hasta su implementación práctica en entornos reales. Originalmente adoptado por la NASA, este marco ha sido ampliamente aplicado en distintos sectores, incluido el sector público, para garantizar que las soluciones tecnológicas evolucionen de manera estructurada, segura y efectiva.

En el contexto de la inteligencia artificial (IA) en el sector público, los TRL permiten identificar el punto en que se encuentra un proyecto, orientar las inversiones, priorizar esfuerzos y reducir los riesgos asociados a la adopción de tecnologías emergentes. De esta manera, las fases de esta guía —Exploración, Desarrollo y Entorno Real— se alinean con diferentes niveles de madurez, favoreciendo la evolución progresiva desde la experimentación hasta la operación a gran escala.

Cada fase cumple un papel esencial para garantizar que los proyectos de IA sean técnicamente sólidos, éticamente responsables y socialmente valiosos, asegurando que la innovación se traduzca en valor público y confianza ciudadana.



Fuente: Elaboración propia.

La fase de Exploración corresponde a los primeros niveles de madurez tecnológica (TRL 1-3). En esta etapa, las instituciones identifican los problemas públicos que pueden abordarse con IA, evalúan su potencial, desarrollan una mentalidad experimental y establecen los principios éticos que guiarán el proceso. El objetivo principal es comprender el contexto, definir el valor público esperado y sentar las bases para un desarrollo responsable.

- Entendimiento del problema y potencial uso de la IA.
- Mentalidad experimental con IA.
- Madurez de la organización basada en datos.
- Establecer principios de IA responsable.
- Mapeo y preparación de datos.
- Modelamiento inicial (definición de objetivos, métricas, ética, selección del modelo, etc.).

1. Entendimiento del problema y potencial uso de la IA en el sector público

Introducción: ¿por qué empezar por aquí?

El reciente informe "The GenAI Divide: State of AI in Business 2025", promovido por el MIT, advierte que aproximadamente un 95 % de los pilotos de IA generativa no logran generar retorno medible. Las razones no suelen radicar en la falta de capacidades técnicas, sino en brechas de alineación estratégica, expectativas poco realistas, ausencia de integración con procesos operativos y falta de gobernanza clara. En otras palabras: la tecnología por sí sola no basta; sin un entendimiento profundo del problema, su contexto, actores y limitaciones, muchos proyectos quedan atrapados en la fase piloto sin escalar.

resolver, por qué vale la pena, quiénes participan, qué expectativas existen y cuál sería el valor público real.

a) Modelo Canvas: las tres dimensiones

El canvas de entendimiento del problema y potencial uso de la IA actúa como una herramienta de descubrimiento estructurado. Permite organizar la reflexión sobre el proyecto de IA para el sector público, asegurando que no se pase nada por alto en esta fase crítica. Las tres dimensiones son:

i) **Visión del problema**: clarifica el problema público que se busca abordar, quiénes lo sienten, en dónde ocurre y qué evidencias lo respaldan.

ii) **Visión del producto de IA**: proyecta cómo podría lucir una solución basada en IA —qué tipo de IA, qué interfaz, cómo interactuarían los usuarios— para validar si la idea tiene sentido operativo.

iii) **Potencial impacto de la IA**: evalúa si la IA es una herramienta idónea para ese problema, cuáles serían los beneficios esperados, las limitaciones, la factibilidad organizacional y el alineamiento con los actores clave.

Dimensión 1: visión del problema

Trabajar esta dimensión con cuidado asegura que el problema no sea solo un apetito tecnológico, sino una necesidad concreta con respaldo. Permite alinear al equipo con la razón de ser del proyecto, identificar actores clave, expectativas, métricas de éxito y validación desde el inicio.

Reto del sector público

- ¿Qué problema se busca resolver con datos e IA?
- Separar los síntomas de las causas.

- Clasificar necesidades y expectativas de cada actor clave.
- ¿Cómo el problema o la solución propuesta impacta a cada actor?

Validación

- ¿Cómo se sabe que este problema es real y relevante?
- Recopilar datos cuantitativos y cualitativos que soporten el planteamiento.

Métricas

- ¿Cómo se medirá el impacto de la solución?
- ¿Qué métricas indican que la posible solución al problema funcionó?

Dimensión 2: visión del producto de IA

Esta dimensión traduce el problema en una idea concreta de solución. Permite comprobar que la idea es realista dado el contexto institucional y técnico, generando claridad para estimar inversión, riesgos y plan de desarrollo.

Interfaz de usuario

- ¿Qué tipo de solución IA se imagina construir (*app*, web, *chatbot*, interno, etc.)?
- ¿Cómo interactuarán los usuarios con la solución?
- ¿Se requiere una interfaz visible o será un algoritmo?

Modelos IA

- ¿El problema que busca resolver es de clasificación, predicción, detección de patrones, optimización o generación de contenido?

Benchmark

- ¿Existen soluciones con IA al reto identificado?
- Evaluar la efectividad de las soluciones existentes.

- ¿Cuál es el costo y tiempo aproximado de implementación?

Dimensión 3: potencial impacto de la IA

Esta dimensión permite estimar si la IA es la herramienta adecuada para el problema identificado. Facilita alinear los beneficios esperados con la capacidad institucional y detectar riesgos o limitaciones que podrían comprometer la legitimidad del proyecto.

Idoneidad de la IA

- ¿Es la IA la herramienta adecuada para resolver el problema?
- ¿Existe una solución más sencilla o barata?
- ¿Cuál es la tolerancia a los errores y al no determinismo?

Beneficios

- ¿Qué procesos actuales podrían beneficiarse con automatización mediante IA?
- ¿Qué beneficios directos percibiría la ciudadanía si se aplica IA en su área de trabajo?

Limitaciones

- ¿Existen limitaciones de calidad, cobertura o interoperabilidad de los datos?
- ¿Falta de capacidad técnica o infraestructura?
- ¿Marco legal, ético o de gobernanza débil?
- ¿Poca confianza de la ciudadanía?

Factibilidad y preparación

- ¿Cómo se encuentra su entidad para implementar soluciones de IA en el sector público?
- Evaluar capacidades institucionales, recursos humanos, financiamiento e infraestructura.

Exploración

1. Visión del problema

1.1 Reto del sector público

- ¿Qué problema se busca resolver con datos o IA?
- Separar los síntomas de las causas.
- ¿Por qué es un problema?
- ¿Está alineado con los objetivos estratégicos de la entidad?

1.2 Alcance

- ¿Dónde ocurre el problema?

1.3 Actores clave

- Mapeo de actores clave directos e indirectos.
- Clasificar necesidades y expectativas de cada actor clave.
- ¿Cómo el problema o la posible solución impacta a cada actor clave?

1.4 Validación

- ¿Cómo se sabe que este problema es real y relevante?
- Recopilar datos cuantitativos y cualitativos que soporten el planteamiento del problema.

1.5 Métricas

- ¿Cómo se medirá el impacto de la solución?
- ¿Qué métricas indican que la posible solución al problema funciona?

2. Visión del producto IA

2.1 Interfaz de usuario

- ¿Qué tipo de solución IA se imagina construir?
- ¿Cómo interactuarían los usuarios con la solución?
- ¿Se requiere una interfaz visible (app, web, bot)

2.2 Modelos IA

- ¿El problema que busca resolver es de clasificación, predicción, detección de patrones, optimización o generación de contenido?

2.3 *Benchmark*

- ¿Existen soluciones con IA al reto identificado?
- Evaluar la efectividad de las soluciones existentes.

2.4 Datos

- ¿Se cuenta con datos relevantes, accesibles y suficientes para entrenar y validar un modelo de IA?
- ¿Existen datos tienen la calidad mínima necesaria (consistencia, ausencia de sesgos críticos, actualización frecuente)?

2.5 Costo y tiempo

- ¿Qué recursos tecnológicos se requieren?
- ¿Qué perfiles se necesitan? ¿Qué costo aproximado?
- Evaluar el costo y tiempo de implementación de una solución IA vs intervención humana.

3. Potencial impacto de la IA

3.1 Idoneidad de la IA

- ¿Es la IA la herramienta adecuada para resolver el problema?
- ¿Existe una solución más sencilla o mas barata?
- ¿Cuál es la tolerancia a los errores y al no determinismo?

3.2 Beneficios

- ¿Qué procesos actuales en su institución podrían beneficiarse con automatización mediante IA?
- ¿Qué beneficios directos percibiría la ciudadanía si se aplica IA en su área de trabajo?

3.3 Limitaciones

- ¿Existen limitaciones de calidad, cobertura, oportunidad o interoperabilidad de los datos?
- ¿Falta de capacidad técnica e infraestructura?
- ¿Marco legal, ético y de gobernanza débil?
- ¿Poca confianza de la ciudadanía?

3.4 Factibilidad y preparación

- ¿Cómo se encuentra su entidad para implementar soluciones de inteligencia artificial en el sector público?

3.5 Alineación actores clave

- ¿Las expectativas de todos los actores clave se encuentran alineadas con la propuesta de solución que usa IA?

Fuente: Elaboración propia.

hipótesis, midiendo resultados y aprendiendo de los errores para avanzar hacia soluciones más sólidas.

El propósito de esta etapa es fortalecer la mentalidad experimental con IA, entendida como la capacidad institucional de aprender haciendo, de aceptar la incertidumbre y de incorporar la evidencia en los procesos de decisión. Una organización con mentalidad experimental no teme al cambio, sino que lo gestiona con método, colaboración y enfoque en el valor público.

La guía se centra en cinco criterios que definen una cultura organizacional preparada para innovar con IA:

i) Explorar diferentes aproximaciones

Promueve la diversidad de ideas y soluciones. No existe una única manera de abordar los retos públicos; se deben explorar múltiples aproximaciones y metodologías antes de decidir.

Preguntas guía:

- ¿Qué alternativas existen para resolver este reto con IA?
- ¿Qué metodologías diferentes podemos probar antes de decidir?

Este criterio fomenta la creatividad y reduce la dependencia de enfoques únicos o rígidos.

ii) Aprender de las fallas

Las fallas son aprendizajes valiosos. Documentarlas, analizarlas y compartirlas permite mejorar continuamente los procesos. La experimentación con IA implica asumir riesgos calculados.

Preguntas guía:

- ¿Qué podríamos aprender de este piloto si no llegara a funcionar?
- ¿Cómo registramos y compartimos los aprendizajes?

Este criterio fortalece la resiliencia y la mejora continua del equipo.

iv) Colaboración multidisciplinaria

La innovación con IA requiere sumar capacidades diversas. Los equipos deben integrar expertos en datos, tecnología, ética, políticas públicas y ciudadanía.

Preguntas guía:

- ¿Quiénes deben participar en este experimento (técnicos, legales, sociales, ciudadanos)?
- ¿Qué valor agrega cada disciplina a la solución?

Este enfoque promueve el trabajo transversal y fortalece la legitimidad del proceso.

v) Orientación a valor público

Todo experimento con IA debe tener un propósito de servicio público. No se trata solo de eficiencia tecnológica, sino de generar beneficios concretos y medibles para la ciudadanía.

Preguntas guía:

- ¿Qué problema ciudadano resolvemos con esta prueba?
- ¿Qué riesgos sociales, éticos o legales debemos mitigar?

Este criterio asegura que la experimentación con IA esté alineada con principios de responsabilidad, transparencia e impacto social.

a) Resultado esperado

El resultado de esta etapa es preparar la mentalidad de innovación de la entidad o equipo para llevar adelante proyectos que involucren el uso responsable de IA. Una organización que adopta esta mentalidad:

- Aprende del error sin penalizarlo.
- Itera con base en datos y evidencia.
- Colabora de manera abierta entre disciplinas.
- Evalúa sus soluciones desde el valor público.

Diagrama 3

Cultura organizacional por la experimentación con IA en el sector público

Cultura organizacional por la experimentación con IA en el Sector Público

Exploración

1.1 Descripción de la cultura organizacional por la experimentación con IA

Actividad

Identifique cómo su proyecto conecta con cada uno de estos criterios de mentalidad experimental con IA.

Instrucciones

- Lea cada sección y utilice el ejemplo como referencia para completar la información específica de su contexto.
- Sea claro y específico en sus respuestas, enfocándose en el criterio identificado.
- Valide la información con datos, testimonios o evidencia relevante.
- Involucre a las partes interesadas para garantizar que el criterio esté bien definido y alineado con las necesidades reales de su proyecto.

Resultado esperado

Preparar la mentalidad de innovación de la entidad o equipo para proyectos que implican uso de IA.

Explorar diferentes aproximaciones	**Aprender de las fallas**	**Interacción continua**	**Colaboración multidisciplinaria**	**Orientación a valor público**
Se promueve la diversidad de ideas y soluciones: no existe una única manera de abordar los retos públicos, se exploran distintas aproximaciones y metodologías antes de decidir.	Las fallas son aprendizajes valiosos. Se documentan los errores, se comparten y convierten en insumos para la mejora continua.	La experimentación con IA no es un evento único, es un proceso constante. Mantenga un diálogo abierto con usuarios, equipos técnicos y actores clave en cada iteración.	La innovación requiere sumar capacidades. Involucre expertos de diferentes disciplinas, equipos de políticas públicas, técnicos y ciudadanía en el diseño y prueba de soluciones.	Cada experimento debe conectar con un propósito claro de servicio a la ciudadanía. Evalúe siempre cómo el uso de IA contribuye al valor público y evita impactos negativos.
• ¿Qué alternativas existen para resolver este reto con IA? • ¿Qué metodologías diferentes podemos probar antes de decidir?	• ¿Qué podríamos aprender de este piloto si es que no llegara a funcionar? • ¿Cómo registramos y compartimos los aprendizajes?	• ¿Con qué frecuencia podríamos validar avances con usuarios? • ¿Cómo integramos la retroalimentación de actores clave?	• ¿Quiénes deben participar en este experimento (técnicos, legales, sociales, ciudadanos)? • ¿Qué valor agrega cada disciplina a la solución?	• ¿Qué problema ciudadano resolvemos con esta prueba? • ¿Qué riesgos sociales, éticos o legales debemos mitigar?

Fuente: Elaboración propia.

nivel de madurez o estado de avance del producto IA dentro de la organización. Al combinar estas dos variables se puede ubicar fácilmente el escenario en el que se encuentra el proyecto IA actualmente en la organización, invitando a mejorar y tomar correctivos para blindar el correcto desarrollo del proyecto.

a) Tres estados de madurez institucional

i) Inmaduro. Una entidad en este nivel se caracteriza por tener capacidades limitadas o inexistentes en la gestión y uso estratégico de datos. Su estructura, procesos y recursos no permiten desarrollar o liderar proyectos de IA de forma sostenible, y si se llevan a cabo iniciativas, suelen ser esfuerzos aislados, sin gobernanza ni continuidad.

ii) Madurez media. Una entidad en este nivel ha iniciado un proceso deliberado de fortalecimiento de sus capacidades en datos. Si bien aún presenta retos de interoperabilidad, estandarización y cultura, ya es capaz de desarrollar pilotos de IA, avanzar proyectos controlados y demostrar resultados de valor.

iii) Maduro. Una entidad madura ha consolidado un ecosistema de datos sólido, seguro, ético y sostenible. Ha institucionalizado las capacidades necesarias para liderar iniciativas de IA en todo su ciclo de vida, desde la exploración hasta la operación y el monitoreo continuo en entornos reales.

Fuente: Elaboración propia.

b) ¿Cómo ubicar el escenario correcto?

Esta guía permite al dueño de un proyecto IA identificar el estado actual de su entidad y proyecto. Para esto se diseñaron unas características aproximadas y referenciales que mejor describen una situación real o hipotética del entorno del proyecto con base en estos criterios: alineación estratégica, calidad de datos, talento, gobernanza, modelamiento, métricas, ética, seguridad, tecnología y cultura. Estos criterios son los más utilizados en la literatura sobre medición de madurez organizacional basada en datos.

Lo primero es identificar la fase en la que se encuentra el proyecto. Si se encuentra iniciando la fase será la de exploración. Si se encuentra en pruebas y validaciones será la fase de desarrollo y si ya la solución se encuentra desplegada, la fase será la de entorno real. Una vez ubicada la fase se procede a identificar las características que mejor describen la situación actual de la entidad y tu proyecto.

Definición	La entidad no tiene capacidades básicas en datos y está intentando explorar IA sin estructuras, sin datos confiables ni talento técnico.	La organización empieza a ordenar su trabajo con datos. Puede realizar pilotos controlados con apoyo institucional.	La entidad explora de manera metódica, estratégica y ética. Cada piloto es un proceso formal de aprendizaje organizacional.
Características aproximadas o referenciales			
Alineación estratégica	No existe claridad sobre el problema; el piloto responde a una intuición, no a una necesidad institucional.	El piloto responde a un problema institucional claro.	Los pilotos se priorizan según valor público esperado.
Calidad de datos	Los datos son de baja calidad, incompletos y no preparados para ningún tipo de análisis.	Datos moderadamente depurados; metadatos básicos.	Datos limpios, integrados y documentados.
Talento	No hay perfiles técnicos, se depende de consultores externos sin guía interna.	Hay analistas y científicos de datos junior; se combina con apoyo externo.	Equipo multidisciplinario capaz de diseñar, ejecutar y evaluar pilotos.
Gobernanza	No hay reglas de tratamiento de datos; alto riesgo legal.	Políticas básicas de acceso y privacidad.	Reglas formales, trazabilidad y auditoría.
Modelamiento	Pruebas improvisadas, sin reproducibilidad.	Prototipos reproducibles; pruebas exploratorias con metodología.	Prototipos robustos con experimentación controlada y reproducible.
Métricas	No se miden resultados ni se define el éxito.	Indicadores básicos definidos; bitácora de resultados.	Indicadores técnicos + institucionales.
Ética	Alto riesgo de sesgo y uso indebido de datos personales.	Primer análisis de sesgo y uso legítimo de los datos.	Evaluaciones éticas en etapas tempranas; mitigación de sesgos.
Seguridad	Sin controles de seguridad.	Controles mínimos implementados.	Entornos seguros y separados para pruebas.
Tecnología	Equipos limitados; no hay capacidades para modelar.	Herramientas de análisis disponibles; entorno de pruebas estable.	Infraestructura sólida y escalable.

Alineación estratégica	El desarrollo no está conectado a procesos misionales.	El proyecto forma parte de la estrategia de datos.	El proyecto está plenamente priorizado y supervisado.
Calidad de datos	Datos mal preparados; cada equipo los trata de forma distinta.	Existen procesos formales de limpieza y versionado de datos.	Alto nivel de madurez. Trazabilidad completa.
Talento	Ausencia de roles clave. Alto riesgo técnico.	Equipo funcional (datos + negocio); soporte externo en temas complejos.	Equipo experto en datos, ML, legal, ético y ciberseguridad.
Gobernanza	No existe control del ciclo de vida de los datos.	Roles definidos; reglas de acceso aplicadas.	Procesos completos y maduros de gestión de datos y modelos.
Modelamiento	Modelos hechos sin metodología, sin versionado ni control de calidad.	Uso de metodologías reproducibles; repositorios versionados.	MLOps en marcha; *pipelines* automatizados.
Métricas	Evaluaciones incompletas o inexistentes.	Métricas técnicas claras; validación cruzada.	Monitoreo continuo de métricas técnicas y de valor público.
Ética	No se evalúan efectos ni sesgos del modelo.	Evaluación inicial de impacto; mitigación incipiente de sesgos.	Evaluaciones profundas, mitigación efectiva de sesgos.
Seguridad	Existen riesgos críticos (datos sensibles expuestos).	Protocolos mínimos, pero suficientes.	Endurecimiento del entorno, segregación de ambientes, monitoreo continuo.
Tecnología	No hay infraestructura adecuada para entrenar modelos.	Infraestructura adecuada para entrenamiento y pruebas.	Plataforma escalable que facilita el despliegue posterior.

Alineación estratégica	No existe alineación, la operación es reactiva.	El sistema es parte de un proceso institucional.	La IA contribuye al cumplimiento de metas institucionales.
Calidad de datos	No se monitorea cambios, ni fallos de calidad.	Se monitorean fallas o cambios con acciones correctivas.	Monitoreo automatizado y trazabilidad completa.
Talento	No hay capacidad para operar, corregir o explicar el modelo.	El personal está capacitado para operar y entender el modelo.	Equipo altamente capacitado, capaz de mejorar modelos continuamente.
Gobernanza	No hay auditoría ni control del modelo.	Comité o instancia que supervisa la operación.	Supervisión permanente, auditorías, transparencia pública.
Modelamiento	El modelo se degrada sin supervisión.	Trazabilidad del modelo; ajustes programados.	MLOps completo: recalibración, explicación, *retraining*.
Métricas	No hay medición de servicio ni impacto.	Indicadores de desempeño operativo; satisfacción del usuario.	Impacto cuantificado en eficiencia, transparencia, equidad y satisfacción.
Ética	Alto riesgo de discriminación, opacidad o injusticia.	Revisión periódica de sesgos; transparencia básica.	Mecanismos de explicabilidad, justicia algorítmica y participación ciudadana.
Seguridad	Modelo expuesto a accesos no autorizados.	Controles de producción implementados.	Cumplimiento de estándares avanzados de ciberseguridad.
Tecnología	Infraestructura no preparada para operación continua.	Entorno estable, con sistemas integrados.	Infraestructura distribuida, escalable y resiliente.

Fuente: Elaboración propia.

- **Tener una visión más clara de mejora gradual**, que conecta directamente con la siguiente etapa de la guía: la adopción de principios y prácticas de IA responsable.

En síntesis, esta sección no busca evaluar ni calificar a la entidad, sino orientarla, ayudándole a entender en qué escenario se encuentra hoy y qué pasos puede dar para fortalecer sus capacidades.

4. Principios de IA responsable

Esta parte de la guía ayuda a equipos del sector público a evaluar y madurar proyectos de inteligencia artificial de forma responsable, segura y transparente. Lo hace mediante preguntas de verificación y una lista de evidencias mínimas que, juntas, permiten decidir si un proyecto puede avanzar, debe mitigarse o debe detenerse temporalmente.

La guía organiza el ciclo de vida en cinco etapas —Exploración, Diseño y Riesgos, Construcción y Pruebas, Piloto Controlado, Despliegue y Operación— y cruza cada etapa con los seis principios de IA responsable: Gobernanza, Rendición de Cuentas, Seguridad y Robustez, Privacidad, Explicabilidad y Transparencia, y Equidad y Detección de Sesgos.

La siguiente plantilla orienta al usuario sobre su estado de cumplimiento de estos principios con preguntas concretas de acuerdo con cada una de las fases de desarrollo del producto IA. Permite identificar las evidencias mínimas que debe tener el producto o entidad para cumplir con los principios.

Exploración	que incorpore estos principios? • ¿El caso de uso está en un registro de iniciativas con responsable y patrocinador?	responsable de cada decisión automatizada/ asistida?	continuidad alta? • ¿Se mapeó la criticidad?	de datos? • ¿El uso de datos es proporcional al problema? (minimización)	operador, auditor) y que detalle necesita cada uno.	(género, edad, etnia, discapacidad, territorio, etc.) y población afectada?
Diseño y riesgos	• ¿Hay un proceso de aprobación con criterios de riesgo? • ¿Está definido el RACI (quién decide, ejecuta, audita, informa)? • ¿Se estableció un plan de participación ciudadana y documentación pública?	• ¿La decisión tiene revisión humana significativa cuando impacta derechos? • ¿Existe proceso de quejas y corrección para ciudadanos/funcionarios?	• ¿Se aplican controles de acceso y segregación de ambientes? • ¿Se planificaron pruebas adversariales/red teaming y plan de incidentes? • ¿Se definieron SLOs de seguridad?	• ¿Se realizó o requiere Evaluación de Impacto en Protección de Datos (DPIA)? • ¿Se definieron retención, eliminación y propósitos (no uso secundario)? • ¿Se evalúan técnicas de anonimización?	• ¿La solución elegida permite el nivel de explicabilidad requerido? • ¿Se definieron formatos de explicación (fichas, tooltips, cartas de notificación)? • ¿Se planificó etiquetar contenido generado por IA?	• ¿Se definieron métricas de equidad relevantes (p.ej. TPR gap, tasa de falsos positivos, disparate impact)? • ¿Hay umbrales de aceptación y plan de remediación?
Construcción y pruebas	• ¿Se versionan datos/ modelos/*prompts* y hay trazabilidad de cambios? • ¿Existen criterios de salida (go/no-go) antes de pasar a piloto?	• ¿Se registran bitácoras de inferencia (quién, cuándo, con qué versión)? • ¿Se definen criterios de override (cuándo el humano puede revertir)?	• ¿Se ejecutaron escaneos de dependencias, hardening de endpoints, límites de tasa? • ¿Se validó la robustez (stress, input sanitization, *prompts* maliciosos)?	• ¿Se eliminan PII innecesarias antes del entrenamiento? • ¿Se prueban riesgos de reidentificación o memorization?	• ¿Se generan explicaciones consistentes (SHAP/ LIME/ reglas/razonamientos estructurado) y se validan con usuarios? • ¿Se documentan limitaciones y supuestos en la tarjeta del modelo?	• ¿Los datos de entrenamiento son representativos? • ¿Se midió cobertura por grupo? • ¿Se ejecutan pruebas A/B de sesgo y se reportan métricas desagregadas?
Piloto controlado **Despliegue y operación**	• ¿Hay un registro de modelos en producción y calendario de auditorías? • ¿Se publican tarjetas de datos y de modelo actualizadas?	• ¿Se mide el tiempo de respuesta a apelaciones y tasa de correcciones? • ¿Se reportan incidentes y acciones correctivas?	• ¿Existe monitor de abuso y detección de drift que dispare alertas? • ¿Hay backups y plan de recuperación (DRP) testeado?	• ¿Se audita el cumplimiento de retención y acceso a datos? • ¿Se informa al ciudadano sobre el tratamiento y derechos ARCO (o equivalentes)?	• ¿El 95% de decisiones críticas cuentan con explicación accesible? • ¿Se publican tarjetas y changelogs de versión?	• ¿Se monitorea equidad en producción con alertas por grupo? • ¿Hay un proceso para corrección (re-entrenar, recalibrar, reglas compensatorias)?
Evidencias Mínimas	• Política de IA • Registro de casos. • Acta de aprobación. • RACI.	• Dueño identificado. • Procedimiento de apelación. • Plantillas de bitácora. • KPI de apelaciones.	• Reportes de pentest/ red team • Plan de incidentes. • Runbooks. • DRP probado.	• DPIA • Matrices de datos con base legal y retención. • Registro de accesos. • Pruebas de reidentificación.	• Plantillas de explicación por actor • Tarjetas publicadas. • Pruebas de usabilidad de explicaciones	• Reporte de sesgos. • Métricas por grupo. • Umbrales aprobados. • Plan remediación ejecutable.
Fuentes		1. NIST AI RMF 2. OECD AI Principles	1. NIST AI RMF 2. EU AI Act	1. Guía del EDPB/Art. 29 sobre DPIA (wp248rev.01) 2. ISO/IEC 23894	1. UK Algorithmic Transparency Recording Standard (ATRS) 2. Model Cards (Model Cards for Model Reporting)	1. NIST AI RMF 2. ISO/IEC 23894

Fuente: Elaboración propia.

El resultado observable es un sistema de IA legal y seguro, con decisiones explicables, sesgos controlados y monitoreo continuo en producción (cambios, desempeño y equidad) más canales de reclamos y retroalimentación.

5. Mapeo de datos

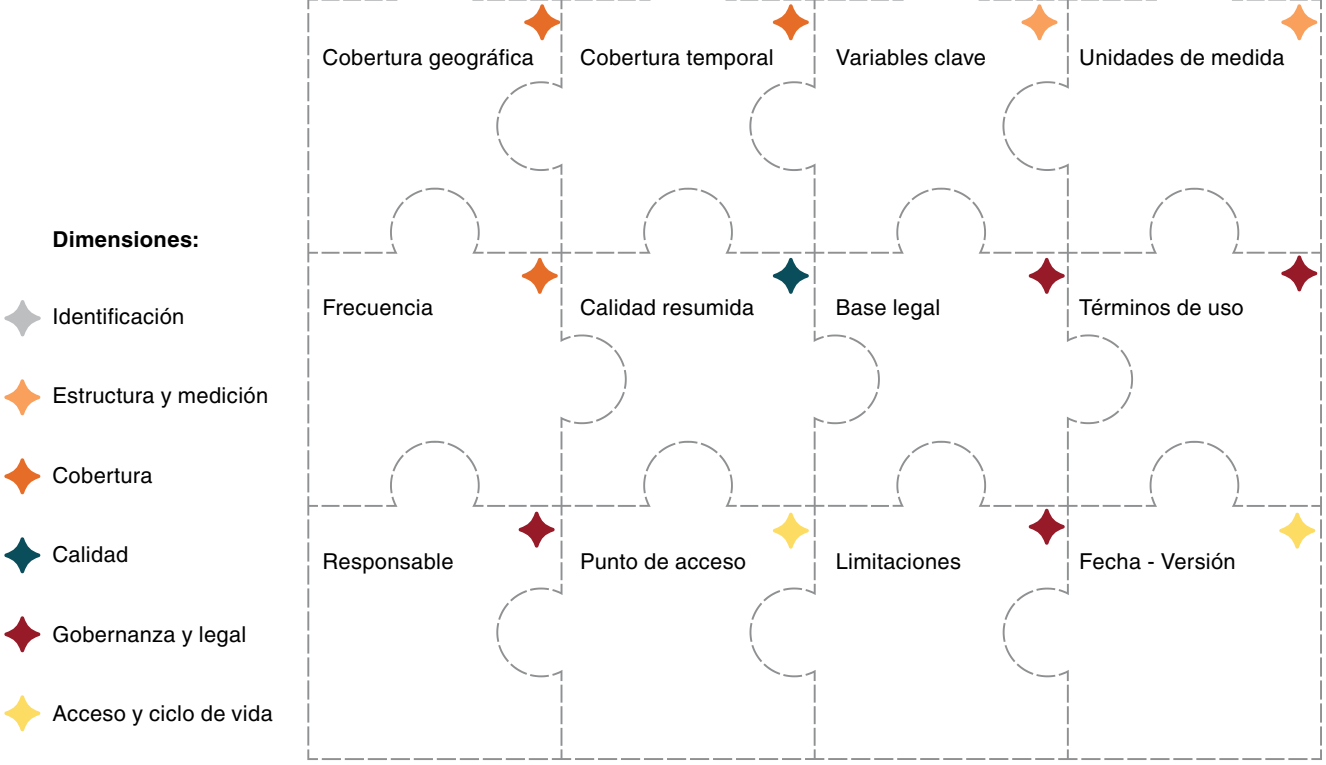
Convierte inventarios dispersos en un catálogo listo para IA: estandariza nombres, fuentes, coberturas, calidad, licencias y responsables para que cada *dataset* sea encontrable, entendible y reutilizable por equipos técnicos, jurídicos y misionales. Facilita la evaluación de sesgos y privacidad desde el inicio, acelera la integración con APIs/pipelines y deja trazabilidad de dónde salen las variables del modelo. La estructura se alinea con buenas prácticas de metadatos como SDMX (series estadísticas), DCAT/DCAT-AP (catálogos de datos públicos), *Datasheets for Datasets y Data Cards* (documentación orientada a riesgo/uso).

Qué campos son críticos: Unidad de observación y Variables clave (definen qué mide el *dataset*), Coberturas temporal/geográfica y Frecuencia (alcance y vigencia), Calidad resumida (idoneidad para el caso), Base legal y Términos de uso (cumplimiento), Responsable y Punto de acceso (gobernanza y operatividad).

a) Dimensiones de la plantilla

- **Identificación**: fuente de datos, Título del *dataset*, Descripción.
- **Estructura y medición**: unidad de observación, Variables clave, Unidades de medida.
- **Cobertura**: cobertura geográfica, Cobertura temporal, Frecuencia.
- **Calidad**: calidad resumida (completitud, exactitud, actualidad, sesgos conocidos).
- **Gobernanza y legal**: responsable, Base legal, Términos de uso, Limitaciones.
- **Acceso y ciclo de vida**: punto de acceso (URL/API), Fecha-Versión.

Con esta ficha, el implementador puede decidir rápidamente si los datos son aptos, legales y suficientes para su caso de IA y qué acciones faltan (mejorar calidad, conseguir permisos, ampliar cobertura, etc.).



Fuente: Elaboración propia.

sí mismo; ejecuta pasos predefinidos.

Ejemplos en el sector público:

- **Trámites masivos**: descargar diariamente planillas de un portal, validar formato y cargar a un sistema interno.
- **Back-office**: leer correos con adjuntos de proveedores, renombrar y archivar los PDFs en carpetas según NIT/fecha.
- **Finanzas**: conciliar pagos copiando datos entre Tesorería y el sistema contable.
- **RR. HH.**: generar certificados automáticos (antigüedad, ingresos) a partir de datos registrados.

- **Modelos de ML tradicionales** (clasificación, regresión, árboles, XGBoost): útiles cuando hay datos estructurados y etiquetas (p. ej., priorización de inspecciones). Se aplican para predecir o analizar datos. Usualmente usados para procesos de optimización. Son simples algoritmos.

Ejemplos en el sector público:

- **Call center**. Demanda de servicios para dimensionar turnos.
- **Alertas tempranas**. Riesgo de deserción escolar.
- **Salud, ambiente, industria**. Riesgo para priorizar inspecciones.

LLM "sencillo" con RAG (*Retrieval-Augmented Generation*), que ni se inventa las respuestas para responder con base en documentos oficiales: suele ser suficiente para preguntas frecuentes, sin necesidad de un agente autónomo. Google resume que un sistema completo combina modelo, herramientas, orquestación y *runtime*, y que el "contexto/grounding" (RAG, *Search*, etc.) es clave para respuestas fiables.

Ejemplos en el sector público:

- **Atención ciudadana**: preguntas sobre requisitos, plazos, costos y pasos de trámites, citando la documentación correspondiente.
- **Contratación**: preguntas frecuentes sobre pliegos y Ley de Compras, con fragmentos exactos del pliego.

- **Triaje y ruteo de casos**: derivación a áreas (salud, social, seguridad) con criterios transparentes y registro auditable.
- **Gestión documental**: extracción/resumen de expedientes, dictámenes y jurisprudencia (RAG/GraphRAG).
- **Compras y contratación**: asistente que verifica requisitos, consolida antecedentes y prepara minutas (Human-In-The-Loop: HITL en decisiones).
- **Inspección/supervisión**: priorización de visitas con ML clásico; el agente prepara expedientes y coordina notificaciones.
- **Respuesta a emergencias**: integración con mapas, radios y protocolos; el agente sugiere acciones y registra bitácora (HITL).
- **Analytics operativo**: consultas en lenguaje natural sobre indicadores (el agente traduce a SQL, recupera y explica hallazgos).

Estos ejemplos fueron consultados de guías de OpenAI, Anthropic y Google.

El modelamiento inicial es una fase crítica en la implementación de proyectos de inteligencia artificial en el sector público, pues permite traducir un problema institucional en un diseño técnico viable, ético y alineado con los objetivos de política pública. Esta etapa sienta las bases para el desarrollo posterior del prototipo, la evaluación del desempeño y la puesta en producción del sistema. A continuación, se presentan los elementos fundamentales que deben abordarse durante esta fase.

c) Definición de objetivos: de la investigación a la producción

El primer paso para el descubrimiento de un modelo de IA en una entidad pública consiste en comprender profundamente el problema que se desea resolver y los objetivos institucionales asociados. Para ello, es esencial analizar de manera exhaustiva la evidencia disponible, incluyendo los hallazgos de investigaciones previas, estudios técnicos y diagnósticos sectoriales.

evaluación permite determinar si dichas metodologías son apropiadas para el problema institucional, si son replicables y qué adaptaciones se requerirían para implementarlas en un entorno real dentro del sector público.

Al realizar este ejercicio, la entidad pública obtiene una comprensión sólida del problema, valida la alineación con sus objetivos estratégicos y sienta las bases para el diseño del primer modelo de IA.

d) Limitaciones de la experimentación

El segundo paso en desarrollo inicial de un modelo de IA dentro del sector público consiste en identificar de manera rigurosa las limitaciones y brechas asociadas al problema y a la evidencia disponible. Todo estudio o análisis previo presenta restricciones que deben ser reconocidas para determinar la viabilidad real de aplicar sus hallazgos en una solución de IA orientada a políticas o servicios públicos.

Este proceso implica revisar posibles sesgos en los datos administrativos o estadísticas utilizadas, identificar limitaciones en el desempeño de modelos desarrollados en otros contextos y evaluar restricciones propias del entorno institucional, como normas, capacidades técnicas, infraestructura o calidad de los registros, que pueden no trasladarse directamente a una implementación en producción dentro de una entidad pública.

Al anticipar estas brechas, la institución puede prever desafíos futuros, diseñar estrategias para mitigarlos y ajustar las expectativas sobre el alcance inicial del modelo de IA.

e) Evaluar la disponibilidad y requerimientos de *data*

El tercer paso en el desarrollo del primer modelo de IA dentro del sector público consiste en evaluar de manera detallada los requisitos y la disponibilidad de los datos necesarios. La investigación previa y las experiencias externas suelen basarse en conjuntos de datos específicos que pueden no estar disponibles, no tener la misma calidad o no ser directamente aplicables en el entorno institucional de una entidad pública.

f) Métricas de desempeño

En el modelamiento inicial se deben definir las métricas con las que se evaluará la efectividad del modelo, asegurando que sean relevantes para el problema público y transparentes para las partes interesadas. Entre las métricas comunes están:

- Exactitud, precisión, recall y F1-score para tareas de clasificación.
- AUC (Área bajo la curva) para evaluar capacidad de discriminación.
- Error cuadrático medio u otras métricas de error según el tipo de modelo.

También debe analizarse:

- Si las métricas capturan adecuadamente los riesgos y el impacto social.
- Si se evalúa desempeño diferencial entre grupos poblacionales.
- La robustez del modelo frente a nuevos datos o escenarios no vistos.

g) Convertir los resultados de la investigación a objetivos misionales

En esta etapa, el equipo debe evaluar cómo los métodos, algoritmos o enfoques analíticos identificados pueden integrarse en los sistemas, procesos o plataformas existentes del sector público. El objetivo no es solo demostrar viabilidad técnica, sino asegurar que la solución tenga un impacto real en términos de eficiencia, transparencia, equidad, calidad del servicio o cumplimiento de mandatos misionales.

Ejemplos prácticos en el sector público:

- **Gestión de subsidios sociales**: Si la investigación identifica un algoritmo capaz de predecir riesgo de filtraciones (personas no elegibles que reciben beneficios), el equipo debe analizar cómo integrarlo en el sistema actual de validación de beneficiarios para mejorar la focalización sin afectar derechos ni equidad.
- **Salud pública**: Si la investigación muestra una técnica que predice brotes epidemiológicos con 10 días de anticipación, la entidad debe evaluar cómo incorporarla en sus sistemas de vigilancia, estableciendo alertas tempranas que permitan respuestas más rápidas y coordinadas.

h) Consideraciones éticas y regulatorias

De acuerdo con la etapa cuatro donde se establecen los principios de IA responsable, esta sección busca que al momento del planteamiento del modelo se tenga presente el cumplimiento de estos principios.

Frente al tema regulatorio, es bueno asegurar desde un inicio el cumplimiento con normativas de privacidad, seguridad y acceso a la información. Del mismo modo, considerar implicaciones sociales, especialmente afectaciones a poblaciones vulnerables.

i) Seleccionar el modelo

Finalmente, con los objetivos, limitaciones y métricas claras, se selecciona el modelo más adecuado. Esta decisión debe considerar:

- La capacidad del modelo para generalizar a nuevos datos.
- Su alineación con las métricas de desempeño priorizadas.
- Su interpretabilidad y explicabilidad, clave en el sector público.
- Su complejidad computacional y compatibilidad con la infraestructura disponible.
- La facilidad para ajustarlo y mantenerlo en ambientes reales de operación.

Introducción

La fase de Desarrollo abarca los niveles intermedios de madurez tecnológica (TRL 4-6). Durante esta etapa, se construyen y prueban soluciones basadas en IA, avanzando desde prototipos hasta modelos validados. El foco está en la experimentación controlada, la validación técnica y la documentación rigurosa para garantizar la reproducibilidad.

7. Diseño de prototipo

8. Testeo

9. Desarrollo de IA hasta operación a gran escala (modelo base [MVP], validación, ajuste, mejora, despliegue)

La etapa 6 de modelamiento inicial también se puede considerarse como una actividad dentro de la fase de desarrollo debido a que en ella se dan los primeros pasos que permiten dimensionar el esfuerzo requerido en desarrollo computacional y analítico. Sin embargo, se deja clasificada en la primera fase dado que tiene una mayor connotación de exploración y descubrimiento.

1. Diseño de prototipado

a) ¿Qué es un prototipo?

Un prototipo es una versión preliminar, simplificada o parcial de una solución final. En el contexto de proyectos de inteligencia artificial, un prototipo permite validar la **viabilidad técnica**, **evaluar la efectividad del modelo** y **detectar posibles problemas** antes de escalar hacia una implementación completa.

realizar compras o despliegues.

Prototipar IA permite validar utilidad, riesgos y costos rápidamente, con usuarios reales y sin comprometer el servicio público ni el presupuesto.

c) Tipos de prototipos específicos para proyectos IA

- **Prototipo Wizard-of-Oz (WoZ).** Simula la "IA" con una persona detrás (el *mago*) para probar tareas, flujos conversacionales o recomendaciones sin construir el modelo. Ideal para asistentes, chatbots, voz o agentes. Permite medir utilidad, expectativas y fallos de comprensión antes de invertir en datos/infraestructura. (Interaction Design Foundation).
- **Prototipo de datos (*Data Cards* / muestreo responsable).** Construye una muestra mínima de datos (entradas/salidas) + su documentación transparente (*Data Cards Playbook*), incluyendo origen, sesgos, permisos y públicos. Sirve para evaluar factibilidad, gobernanza y costo de preparación/etiquetado antes del entrenamiento masivo. (Google PAIR/Data Cards, 2022-2024).
- **Prototipo de modelo/documentación (*Model Cards*).** Artefacto rápido para describir uso previsto, métricas, "slices" poblacionales, límites y riesgos de un modelo. Útil para comités de ética, compras y audiencias no técnicas del Estado. (Mitchell et al., "Model Cards", 2018/2019).
- **Prototipo de *prompts* y cadenas (*LLM prompt chaining*).** Explora, en pequeño, *prompts*, funciones y reglas (instrucciones, extracciones, herramientas) para ver si el LLM resuelve la tarea pública (por ejemplo, clasificar solicitudes ciudadanas) y cómo se comporta ante errores. (Google PAIR – People + AI Guidebook).
- **Prototipo de interacción humano-IA (*HAX*/*Guidelines*).** Boceta estados, explicaciones, manejo de errores, confirmaciones y control humano en la UI antes de código final. Útil para cumplir principios de transparencia y *human-in-the-loop*. (Microsoft Research: Guidelines for Human-AI Interaction + HAX Toolkit).

público para testear con ciudadanía la experiencia completa donde la IA es solo un componente del servicio. (GOV.UK Service Manual/Prototype Kit).

- **Prototipo de gobernanza y compras**. Pilotos de bajo riesgo que validan roles, salvaguardas, DPIA, cláusulas de contratación, auditorías y métricas de impacto público antes de tender o escalar. Útil para comités y oficinas de compras. (G7/OECD “AI in the Public Sector” y “AI Procurement in a Box”).

d) ¿Cómo elegir el prototipo correcto?

- Si se quiere exploración temprana se recomienda usar WoZ.
- Para analizar riesgos y datos, se recomienda *Data Cards, Model Cards y Sandbox*.
- Para validar experiencia de usuario y explicable, se recomienda Interacción HAX/Guidelines y experiencia de prototipado del gobierno británico.
- Para temas de gobernanza y compras públicas, se recomienda usar los lineamientos de la OECD.

e) En resumen

Desarrollar prototipos en el sector público es fundamental porque permite validar soluciones antes de realizar inversiones significativas, reduciendo riesgos y costos. Además, promueven la transparencia al facilitar la revisión de decisiones y garantizar la equidad en los modelos de IA, al mismo tiempo que permiten evaluar el impacto ético y social mediante la identificación de sesgos y posibles efectos en poblaciones vulnerables. Los prototipos también fortalecen la colaboración interinstitucional al alinear equipos técnicos, jurídicos, operativos y de política pública; ayudan a clarificar expectativas para los tomadores de decisión al ofrecer una versión funcional del proyecto, y contribuyen al desarrollo de capacidades institucionales al impulsar prácticas internas sólidas en el uso responsable de la inteligencia artificial.

Las pruebas iniciales permiten medir el desempeño básico del modelo mediante métricas como exactitud, precisión, recall, F1-score y AUC. Estas métricas ayudan a identificar si el modelo se comporta adecuadamente con respecto al problema público que se quiere resolver. Organismos como NIST y la OCDE recomiendan evaluar al menos dos métricas complementarias para evitar interpretaciones sesgadas del rendimiento. Por ejemplo, en problemas de detección de fraude o focalización social, una alta exactitud puede ocultar fallas graves si la clase positiva es muy rara.

Según NIST (2023), el 78% de los sistemas evaluados mostró diferencias significativas entre métricas, lo que destaca la importancia de revisar más de una medida.

b) Pruebas iterativas y refinamiento

El desarrollo de proyectos de IA debe incluir múltiples ciclos de prueba, ajuste y revisión. Cada ciclo permite mejorar el rendimiento del modelo, corregir errores, afinar hiperparámetros y ajustar el tratamiento de datos.

Gobiernos como los de Canadá, Reino Unido y Singapur recomiendan trabajar en ciclos cortos de validación para reducir riesgos y evitar que pequeños errores se amplifiquen al escalar. En cada iteración deben documentarse:

- Cambios realizados
- Problemas detectados
- Mejoras observadas

Este proceso garantiza transparencia y facilita auditorías futuras.

c) Pruebas unitarias

Las pruebas unitarias permiten verificar que cada componente del sistema funciona de forma aislada. Esto incluye:

Estas pruebas asegurarán que los diferentes componentes del sistema funcionen juntos sin problemas. En el sector público, esto es especialmente importante debido a:

- Sistemas heredados (*legacy*).
- Inconsistencias en plataformas entre entidades.
- Dependencia de APIs externas.
- Interoperabilidad con plataformas de datos institucionales.

La OCDE enfatiza que la interoperabilidad es uno de los principales desafíos para la adopción de IA en gobiernos. Asegurar integración fluida reduce tiempos de implementación y fallos operacionales.

e) Pruebas de rendimiento

Las pruebas de rendimiento evalúan:

- Tiempos de respuesta.
- Escalabilidad bajo alta demanda.
- Robustez ante cargas variables.
- Capacidad para procesar grandes volúmenes de datos.

Esto es fundamental en el sector público, donde los sistemas deben funcionar incluso en períodos de alta demanda (por ejemplo, inscripción escolar, reportes tributarios, solicitudes de subsidios). El Banco Mundial estima que hasta un 30% de los fallos en proyectos digitales gubernamentales se deben a problemas de escalabilidad no detectados en etapas de prueba.

f) Pruebas de aceptación del usuario (UAT)

El User Acceptance Testing (UAT) involucra a funcionarios públicos o usuarios finales para validar que el sistema:

- Sea fácil de usar.
- Resuelva realmente el problema.
- Entregue resultados comprensibles.
- Se adapte a los flujos de trabajo reales.

- Mejorar la experiencia del usuario.
- Validar cumplimiento normativo.
- Corregir posibles sesgos detectados.
- Verificar interpretabilidad del modelo.

Los estándares del AI Risk Management Framework del NIST recomiendan realizar una última evaluación de riesgos antes del despliegue.

h) Preparación para el despliegue a gran escala

Antes de implementar oficialmente el sistema de IA, se deben cumplir tres tareas clave:

i) Documentar todo el proceso: decisiones, pruebas, métricas y cambios.

ii) Asegurar confiabilidad: el modelo debe ser robusto, auditable y reproducible.

iii) Planear monitoreo continuo: el Banco Mundial recomienda monitorear modelos de IA al menos cada 3 meses durante el primer año de vida, para revisar:

- Desviación o "drift" del modelo
- Cambios en la calidad de los datos
- Impactos en usuarios reales

Un monitoreo adecuado puede reducir de manera importante los riesgos de fallos operativos.

3. Desarrollo de IA hasta operación a gran escala

Hasta este punto hemos recorrido 8 etapas. Entramos a la etapa 9, donde luego de conducir testeos sobre el prototipo construido, se prepara el terreno para llevar esta solución a un nivel más arriba. La siguiente secuencia de siete pasos describe los elementos clave para lograr desarrollos robustos, confiables y listos para escalar:

Por qué es importante

Permite identificar desde temprano:

- Si el enfoque técnico es viable.
- Si los datos actuales son suficientes.
- Qué tan lejos está el modelo de los objetivos del proyecto.

Recomendaciones:

Comenzar con un modelo simple: “Start simple” (Google ML Engineering Guide).

Evitar optimizaciones tempranas: “First make it work, then make it better” (Andrew Ng).

b) Estándares de validación en IA

Aquí se definen los criterios que el modelo debe cumplir para ser aceptado:

- Métricas de desempeño (precisión, recall, F1-score, velocidad).
- Comportamiento esperado en diferentes escenarios.
- Límites de errores aceptables (falsos positivos y negativos).
- Consideraciones éticas (sesgos, impacto en grupos vulnerables).

Recomendaciones:

Incluir métricas de equidad y asegurar transparencia y trazabilidad del modelo. (Microsoft Responsible AI Standard, Stanford HAI.).

c) Aumento de datos y mejora de la calidad

Una vez revisado el modelo base con los criterios de aceptación, se identifica la necesidad de más y mejores datos.

- Usar “data-centric AI”: mejorar datos antes que cambiar modelos.
- Aplicar *data versioning* para rastrear cambios (DVC, LakeFS).

d) Ajuste y regularización de modelos

Optimización del modelo para mejorar su equilibrio entre precisión y recall.

Acciones

- Ajuste de hiperparámetros.
- Uso de regularización L1/L2 o dropout para evitar sobreajuste.
- Modificar la complejidad del modelo según rendimiento.

Recomendaciones (DeepMind/OpenAI)

- Aplicar *cross-validation* para evitar errores de sobreajuste.
- Usar búsqueda bayesiana en lugar de *grid search* para eficiencia.

e) Técnicas avanzadas e iteración

Etapa para llevar el modelo a otro nivel.

Acciones

- Métodos de ensamble (*Random Forest, boosting, stacking*).
- Funciones de pérdida personalizadas según objetivos del negocio.
- Aprendizaje activo para etiquetar datos sobre los que el modelo tiene mayor incertidumbre.

Recomendaciones (Meta AI/Google Brain)

- Aplicar *ensembling* solo si el costo computacional no afecta escalabilidad.
- Para problemas críticos, usar calibración de probabilidades (Platt/Isotonic).

- Gestión de falsos positivos y falsos negativos.

Recomendaciones (Google Cloud/IBM)

- Configurar alertas automáticas cuando métricas clave caen bajo umbrales definidos.
- Hacer auditorías mensuales para revisar sesgos y desempeño.
- Registrar logs de predicción para análisis forense.

g) Preparación de despliegue a gran escala

Consiste en preparar la infraestructura, procesos y equipos para llevar un modelo a un número alto de usuarios.

Acciones

- Pruebas de carga y estrés.
- Validar escalabilidad de la arquitectura.
- Entrenamiento de usuarios y equipos operativos.
- Preparación de políticas éticas y de uso responsable.
- Construcción de documentación y manuales.

Por qué es clave

Permite que el modelo:

- soporte grandes volúmenes de datos,
- responda rápido,
- sea seguro y confiable,
- cumpla normativas del sector público.

Recomendaciones (NIST/MLOps AWS)

- Definir un “plan de rollback” si el modelo falla.
- Utilizar despliegues progresivos (blue/green, canary).

Introducción

La fase de Entorno Real corresponde a los niveles más altos de madurez tecnológica (TRL 7-9). En esta etapa, los modelos se despliegan, monitorean y mantienen operativos en escenarios reales. El énfasis está en la sostenibilidad, la mejora continua y la detección temprana de cambios en los datos o en los conceptos que puedan afectar el desempeño del modelo.

10. Operación del modelo (monitoreo, reentrenamiento, *feedback*, evaluación)

Este enfoque gradual, basado en los niveles de madurez tecnológica, permite a las instituciones públicas implementar proyectos de inteligencia artificial de manera responsable, efectiva y sostenible. A medida que los proyectos avanzan entre fases, aumenta su robustez técnica y su capacidad para generar valor público tangible con este tipo de tecnologías.

1. Monitoreo a cambios conceptuales y patrones de datos (*Data & Concept Drift*)

Una vez que el modelo está en operación, es fundamental vigilar si las condiciones del entorno cambian. Esto implica detectar variaciones en los datos reales que no estaban presentes en los datos de entrenamiento (por ejemplo, cambios en el comportamiento ciudadano, nuevas normativas, variaciones económicas o sociales). Estos cambios pueden afectar la precisión del modelo. Este paso asegura que el modelo siga siendo útil, justo y confiable en contextos dinámicos del sector público.

Por ello es necesario evaluar continuamente la calidad de los datos que ingresan al sistema, así como la coherencia entre distintos modelos. Acá se busca asegurar que las predicciones se basan en datos fiables, representativos y correctamente armonizados entre sistemas.

4. *Feedback* (retroalimentación)

La operación real del modelo permite recibir retroalimentación de usuarios finales, operadores, ciudadanos o entidades involucradas. Esta información es clave para detectar problemas no previstos, mejorar la experiencia de uso y asegurar que el sistema realmente responde a una necesidad pública. El objetivo central es recoger evidencia del mundo real para mejorar continuamente el modelo y su utilidad pública.

5. Documentación y evaluación

Toda la operación del modelo debe documentarse: cambios, reentrenamientos, métricas, incidencias y decisiones técnicas. También deben realizarse evaluaciones periódicas para medir impacto, equidad, eficiencia y cumplimiento normativo. Esto es esencial en el sector público para garantizar transparencia, rendición de cuentas y trazabilidad. Esta último paso busca mantener un registro claro y auditable del funcionamiento del sistema y de su impacto en la ciudadanía.

Agarwal, S. (2025). Successful AI product creation: A 9-step framework. Wiley. AI Now Institute. (2018–2022). Algorithmic Accountability Reports. AI Now Institute at New York University.
AlgorithmWatch. (2021). AI Ethics Guidelines Global Inventory. AlgorithmWatch.
Amazon Web Services. (n.d.). What is artificial intelligence (AI)? Recuperado el 3 de noviembre de 2025, de https://aws.amazon.com/what-is/artificial-intelligence/
Andreessen Horowitz (a16z). (n.d.). AI glossary. Recuperado el 3 de noviembre de 2025, de https://a16z.com/ai-glossary/
Ada Lovelace Institute. (s.f.). *Our work*. Ada Lovelace Institute. Recuperado el 8 de enero de 2026, de https://www.adalovelaceinstitute.org/our-work/
Carnegie Mellon University, Heinz College. (2023, julio). Artificial intelligence, explained. Recuperado el 3 de noviembre de 2025, de https://www.heinz.cmu.edu/media/2023/July/artificial-intelligence-explained
Center for the Governance of AI (GovAI). (2023). Risks and Governance of Machine Learning in the Public Sector. University of Oxford.
CNET. (n.d.). ChatGPT glossary: 59 AI terms everyone should know. Recuperado el 3 de noviembre de 2025, de https://www.cnet.com/tech/services-and-software/chatgpt-glossary-59-ai-terms-everyone-should-know/
Coursera. (n.d.). AI terms. Recuperado el 3 de noviembre de 2025, de https://www.coursera.org/resources/ai-terms
DAMA International. (2017). DAMA-DMBOK: Data Management Body of Knowledge (2nd ed.). Technics Publications.
European Commission. (2021). AI Readiness Assessment for Public Administrations. European Union Publications.
European Commission, AI Watch (2024). AI Adoption in the Public Sector: Key Influencing Factors and Frameworks. https://ai-watch.ec.europa.eu/news/ai-adoption-public-sector-new-study-key-influencing-factors-and-two-frameworks-competencies-and-2024-11-25_en

Google. (2022). MLOps: Continuous Delivery and Automation Pipelines in Machine Learning (2nd ed.). Google Research.
Google PAIR, *People + AI Guidebook* y *Data Cards Playbook*. pair.withgoogle.com+1
Government of Canada (2021). *Algorithmic Impact Assessment (AIA)*.
Government Digital Services (GDS), 2019. *AI Ethics Guidance for the Public Sector*.
GOV.UK, *Service Manual / Prototype Kit*. GOV.UK+1
IBM. (n.d.). Artificial intelligence (AI). Recuperado el 3 de noviembre de 2025, de https://www.ibm.com/es-es/think/topics/artificial-intelligence
Infocomm Media Development Authority of Singapore. (2022). Model AI Governance Framework (2nd ed.). Singapore Government.
Inter-American Development Bank. (2022). Artificial Intelligence for the Public Good in Latin America and the Caribbean. IDB Publications.
Interaction Design Foundation. Wizard of Oz Prototypes. https://www.interaction-design.org/literature/topics/wizard-of-oz-prototypes?srsltid=AfmBOoqnEddCA8knX2aVjJnpZb_OIcqPVKa6WkQ1vlE9ZKjDhZCCi8bD
International Organization for Standardization. (n.d.). Inteligencia artificial. Recuperado el 3 de noviembre de 2025, de https://www.iso.org/es/inteligencia-artificial
International Organization for Standardization. (n.d.). ISO/IEC 22989: Artificial intelligence—Concepts and terminology (visor). Recuperado el 3 de noviembre de 2025, de https://www.iso.org/obp/ui/en/#iso:std:iso-iec:22989:ed-1:v1:en
McKinsey, *The state of AI in 2025* (05-nov-2025). https://www.mckinsey.com/capabilities/quantumblack/our-insights/the-state-of-ai
Microsoft Azure. (n.d.). What is artificial intelligence (AI)? Recuperado el 3 de noviembre de 2025, de https://azure.microsoft.com/en-us/resources/cloud-computing-dictionary/what-is-artificial-intelligence#self-driving-cars
Microsoft. (2021). The Machine Learning Lifecycle. Microsoft Corporation.
Microsoft (2022). *Responsible AI Standard*.
Microsoft Research, *Guidelines for Human-AI Interaction + HAX Toolkit*. microsoft.com+1
Microsoft Research (2023). Estudios sobre *silent failures* y pruebas unitarias en modelos ML.

National Institute of Standards and Technology. (2021). Zero Trust Architecture (NIST SP 800–207). U.S. Department of Commerce.
National Institute of Standards and Technology. (2023). AI Risk Management Framework (NIST AI RMF 1.0). U.S. Department of Commerce.
NIST (2021). *Four Principles of Explainable AI*.
OECD (2023). The State of Implementation of the OECD AI Principles: Insights from National AI Policies. Destaca la importancia de la experimentación responsable, el aprendizaje continuo y la coordinación entre equipos multidisciplinarios. https://oecd.ai/en
OECD Observatory of Public Sector Innovation (OPSI) (2021). Innovation Playbook for Governments. Contiene herramientas sobre cultura experimental, gestión del riesgo y co-creación en entornos públicos. https://oecd-opsi.org/tools/innovation-playbook.
Organisation for Economic Co-operation and Development. (2019). OECD Principles on Artificial Intelligence. OECD Publishing.
Organisation for Economic Co-operation and Development. (2021). Algorithmic Accountability in the Public Sector. OECD Public Governance Directorate.
Organisation for Economic Co-operation and Development. (2021). Good Practice Principles for Data Ethics in the Public Sector. OECD Public Governance Directorate.
Organisation for Economic Co-operation and Development. (2022). State of Implementation of the OECD AI Principles in the Public Sector. OECD Publishing.
OECD/G7, *Toolkit for AI in the Public Sector* y *AI Procurement in a Box*.
Rohit Madan, Mona Ashok, AI adoption and diffusion in public administration: A systematic literature review and future research agenda, Government Information Quarterly, Volume 40, Issue 1, 2023, 101774, ISSN 0740-624X, https://doi.org/10.1016/j.giq.2022.101774
Singapore PDPC (2020). *Model AI Governance Framework 2.0*.
Stanford Institute for Human-Centered Artificial Intelligence (HAI). (n.d.). AI key terms—Glossary & definition [PDF]. Recuperado el 3 de noviembre de 2025, de https://hai.stanford.edu/assets/files/2023-03/AI-Key-Terms-Glossary-Definition.pdf

of Artificial Intelligence. UNESCO.
United States Government Accountability Office. (2021). Artificial Intelligence: An Accountability Framework for Federal Agencies and Other Entities. GAO-21-519SP.
World Bank. (2021). Data Governance Framework. World Bank Group.
World Bank. (2023). AI for Public Sector Innovation: Building Capabilities for Responsible Adoption.Explica cómo los gobiernos pueden desarrollar capacidades institucionales para experimentar con IA y escalar proyectos exitosos. https://www.worldbank.org/en/topic/digitaldevelopment

cada etapa.

El enfoque combina valor público, datos técnicos y gobernanza, lo que ayuda a decidir cuándo usar IA, cómo compararla con alternativas no basadas en IA y cómo gestionar los riesgos asociados a datos, sesgos, privacidad, seguridad y rendición de cuentas.

En esta guía se integran herramientas prácticas, como listas de verificación, criterios de madurez institucional, prototipado, testeo y monitoreo continuo, alineadas con los principios de IA relativos a la responsabilidad. Está dirigida a equipos directivos, técnicos, jurídicos y operativos, y busca crear un lenguaje común que facilite proyectos de IA útiles, auditables y sostenibles en el sector público.

Versión digital disponible online

https://bit.ly/CEPAL2026-33S

CEPAL

Comisión Económica para América Latina y el Caribe (CEPAL)
Economic Commission for Latin America and the Caribbean (ECLAC)
www.cepal.org