Recomendaciones para una

Inteligencia Artificial Fiable

Jefatura de Gabinete de Ministros
Argentina

Secretaría de Innovación Pública

Índice

Consideraciones preliminares

La irrupción de la Inteligencia Artificial (IA), que se expresa en la creciente importancia de los datos y algoritmos en la vida de las personas, empuja a los Estados a definir estrategias para encauzar el potencial transformador de esta tecnología en la resolución de problemas concretos y a favor del bien común.

Las soluciones tecnológicas basadas en IA permiten mayores niveles de automatización y el salto hacia sistemas descentralizados y predictivos para la toma de decisiones. En el plano productivo, la IA es promisoria por su capacidad de promover la innovación, agregar valor, incrementar la productividad del trabajo, dar origen a nuevos bienes y servicios, potenciar las exportaciones, entre otras posibilidades.

En el ámbito público, la IA ofrece soluciones que permiten hacer más eficiente la gestión del Estado y mejorar el diseño y la implementación de las políticas y la prestación de servicios esenciales en salud, educación, seguridad, transporte, cuidado del medio ambiente, etc. Los gobiernos también pueden utilizar la IA para mejorar la comunicación y el compromiso con los ciudadanos.
En este sentido, el Estado cumple un rol fundamental no sólo promoviendo la investigación y el desarrollo de soluciones de IA que estén diseñadas para atender las necesidades reales de las personas, sino también, garantizando que la IA sea transparente, equitativa y responsable. Esto implica, establecer reglas claras para garantizar que las bondades de cualquier desarrollo tecnológico puedan ser aprovechadas por todos los sectores de la sociedad; para promover la responsabilidad en la recolección y uso de los datos personales, evitar la discriminación algorítmica y gestionar los riesgos del uso de la IA para prevenir perjuicios.

Argentina cuenta con un ecosistema científico y tecnológico dinámico, con probadas capacidades para la innovación, el desarrollo y la producción de soluciones tecnológicas basadas en IA. Es clave generar las condiciones políticas e institucionales para que dichas capacidades se pongan en valor al servicio de una estrategia más amplia que priorice la soberanía tecnológica y permita dar respuesta a los problemas sociales, productivos y medioambientales del país.

1.1 Objetivo

A través del presente documento se procura recopilar y brindar herramientas para quienes llevan adelante proyectos de innovación pública a través de la tecnología,

pero específicamente en aquellos que importen el uso de inteligencia artificial. En este sentido, se recomienda adoptar un enfoque multidisciplinar, concibiendo de forma integral las implicancias del uso, adopción, desarrollo e innovación pública a través de la inteligencia artificial.

El manual se encuentra destinado a brindar un marco para la adopción tecnológica de la inteligencia artificial centrada en el ciudadano y sus derechos, concibiendo su aspecto social y estratégico, asegurando un óptimo funcionamiento de la prestación de servicios y un enfoque ético.

1.2 Alcance

Este manual busca ofrecer herramientas teóricas y prácticas a quienes formen parte del sector público, ya sea liderando proyectos de innovación, desarrollando tecnologías, adoptando tecnologías desarrolladas por otros equipos técnicos/proveedores, formulando las especificaciones técnicas para esas adquisiciones.

Marco conceptual

2

La inteligencia artificial actualmente agrupa un conjunto de tecnologías y lleva por nombre una habilidad que durante mucho tiempo fue considerada exclusiva de las personas: **la inteligencia.**

En el momento en que este conjunto de tecnologías fue bautizado con ese nombre, el concepto de inteligencia era bastante diferente a las ideas y teorías que en la actualidad se discuten sobre lo que hoy entendemos por inteligencia humana.

Así, a mediados del siglo XX, el estudio de la inteligencia se encontraba centrada exclusivamente en las capacidades cognitivas y, dentro de ellas, en las lógico- matemáticas y lingüísticas. En esa época también comenzaba a popularizarse la analogía mente-computadora con la cual profesionales de área de la psicología cognitiva realizaban metáforas computacionales para explicar los avances, teorías y descubrimientos de la mente humana, así como profesionales de las ciencias de computación usaban esa analogía con la mente humana como inspiración para definir la arquitectura de las primeras computadoras.

Los estudios actuales sobre la inteligencia humana ampliaron ese concepto específico de inteligencia y reformularon su comprensión expandiendo a distintas áreas que antes no eran consideradas como pertenecientes a la inteligencia humana. Howard Gardner, al desarrollar las inteligencias múltiples, expone como adicional a la lógico- matemática y la verbal-lingüística, a la inteligencia musical, kinestésica-corporal, visual-espacial, intrapersonal, interpersonal, y natural. Estas teorías representan un marco abierto al que, a medida que avanzan los estudios, se van definiendo y

agregando nuevos tipos de inteligencia, como la inteligencia emocional. Las inteligencias múltiples contribuyen a comprender el alcance de las tecnologías de inteligencia artificial actuales, ya que se puede asociar el ámbito de alguna de esas inteligencias múltiples con los destinos de uso o tipos de tecnología de inteligencia artificial.

A diferencia de los humanos, que poseemos todas esas inteligencias en mayor o menor grado, más desarrolladas unas que otras, la inteligencia artificial hasta hace muy poco tiempo, sólo podía cubrir a la vez un tipo de esas inteligencias. Por ejemplo, en la actualidad existen inteligencias artificiales utilizadas únicamente para el procesamiento natural del lenguaje, sólo realizan esa tarea, son digamos “buenas escritoras”. Pero, por ejemplo, no todas pueden “escucharnos” sólo “leer lo que escribimos” (luego de que nuestra escritura es codificada en binario, el idioma en el que las inteligencias artificiales procesan la información). Aunque ninguna de ellas puede aún hablarnos, mirarnos, inferir lo que pensamos, y de forma simultánea empatizar con nosotros, así como tampoco otras acciones similares que son (hasta el momento) típicamente humanas.

Esta estrechez, es una característica por la cual se denomina la gran mayoría de las inteligencias artificiales que hoy en día usamos, la “inteligencia artificial estrecha” (también llamada débil). Es decir, estas tecnologías se consideran “inteligentes” en un aspecto muy puntual tomando en cuenta el amplio espectro de la cognición humana.

Existe también conceptualmente la inteligencia artificial general (también llamada fuerte), que sería equivalente a la inteligencia humana, pero hasta ahora es un abordaje teórico. No obstante, la evolución tecnológica avanza muy rápidamente, y en la actualidad ya existen tipos de inteligencias

artificiales multimodales. La multimodalidad permite sumar dos o más inteligencias que trabajan con un solo tipo de datos. Por ejemplo, una que funciona con texto y otra con imágenes, y hacerlas trabajar en conjunto ampliando el alcance de las inteligencias artificiales estrechas.

Dichas inteligencias artificiales, que además de recibir entradas del tipo texto reciben imágenes, ellas a su vez pueden contener texto, y también pueden reconocerlo y tomarlo como entrada. Si bien la multimodalidad parece ser un paso en la dirección correcta en el camino hacia la inteligencia artificial general, cualquier pronóstico que se pueda realizar en ese sentido, hoy en día, es meramente especulativo. Para contextualizar esto último, podemos listar conceptualmente distintas similitudes y diferencias entre los sistemas de inteligencia artificial y los humanos.

El marco que describe a los sistemas de inteligencia artificial elaborado por la Organización para la Cooperación y el Desarrollo Económicos (OCDE), muestra que los modelos de inteligencia artificial interactúan con el contexto recibiendo distintos tipos de datos (generados por personas, sensores, curados por expertos, públicos, privados, dinámicos, estáticos, etc.), que son utilizados para construir el modelo de

inteligencia artificial, el cual, una vez entrenado, procesa ese tipo de datos para brindar distintas respuestas de salida (reconocimiento, detección de eventos, pronóstico, y otras acciones) con distintos destinos de uso que pueden ser lenguaje humano, visión artificial, automatización, optimización robótica, etc. Todas estas acciones afectan al contexto, es decir pueden alterar el entorno en el cual vivimos.

Conceptualmente se puede trazar una analogía con esta manera de describir a los sistemas de inteligencia artificial con la forma en que los seres humanos interactúan con sus contextos. Percibimos con los sentidos la información del entorno, representamos mentalmente esa información y allí podemos procesarla realizando distintas operaciones mentales, luego podemos actuar de diferentes maneras y como consecuencia de esas operaciones mentales podemos, hablar, escribir, reconocer personas, crear, etc.

En este nivel de descripción las similitudes son generales y están siendo contrastadas con las “cajas negras” de la inteligencia artificial[1], es decir, con aquellos modelos poco transparentes e incapaces de explicar sus resultados. No obstante, si se compara observando dentro de dichas cajas, existen diferencias sustanciales entre las máquinas y los humanos que hacen que el camino hacia una inteligencia artificial general no ocurra, al menos, en el corto plazo. Una de las diferencias principales es la consciencia. Antonio Damasio, al abordar el tema de la consciencia describe tres estadios, uno es la consciencia la que permite tener la capacidad de percibir lo que sucede en el interior de nuestro cuerpo, que es diferente a la capacidad de percibir lo que sucede en el exterior y el entorno, el segundo estadio, sobre estos dos se construye el tercer estadio de la conciencia, denominada autobiográfica, que permite recordar el pasado y proyectar o imaginar el futuro.

La autodeterminación es la facultad de una persona para decidir por sí misma algo, y ésta es también una capacidad humana, que permite actuar con libertad y elegir acciones con intención y propósito, al mismo tiempo que comprendemos las consecuencias de dichas acciones y la responsabilidad que tenemos sobre ellas. Sirve para construir nuestro autoconcepto. Es decir, comprender la imagen que tenemos de nosotros mismos, por ejemplo, con las habilidades y competencias que poseemos para hacer ciertas tareas de un modo efectivo; a la vez que también empuja a cubrir la necesidad de integrar grupos de pertenencia en los que participamos por afinidad con otras personas. Estas características humanas refieren a las necesidades psicológicas básicas de autonomía, competencia y afinidad, definidas en la teoría de la Autodeterminación de Deci y Ryan.

---

1 Se trata de algoritmos de aprendizaje automático o redes neuronales profundas, entre otros, que no revelan cómo procesan la información o toman decisiones. Es decir, modelos cuyo funcionamiento interno es desconocido o no transparente para los observadores externos. Frente a las cajas negras, los observadores externos sólo pueden ingresar datos de entrada y recibir resultados de salida, sin tener una comprensión clara de los pasos intermedios o los factores que influyen en las decisiones tomadas. Aunque las cajas negras pueden ser altamente efectivas para resolver problemas complejos y lograr resultados precisos, plantean desafíos en términos de explicabilidad y ética.

Todos estos aspectos, junto con la experiencia acumulada que brinda el conocimiento del mundo, nuestro cuerpo sensible al entorno y las emociones que modulan los pensamientos, conforman la experiencia subjetiva humana de la cual (al menos por ahora), las inteligencias artificiales no gozan.

No obstante, a través de la estadística, la matemática, grandes volúmenes de datos, la infraestructura informática y distintas interfaces que pueden brindar un cierto grado de acción en el entorno en que vivimos las personas, estas tecnologías son un reflejo de nuestra propia humanidad, un reflejo parcial pero reflejo al fin, construido con las virtudes y defectos propios. Estos conceptos fueron abordados por la filósofa Shannon Vallor a través de la teoría del espejo, estableciendo que este reflejo debe ser observado y optimizado, no sólo a través del desarrollo y evolución de las tecnologías sino que fundamentalmente apuntando a mejorar nosotros mismos como personas.

Recomendaciones y principios para la implementación de proyectos de IA.

El desarrollo e implementación de IA puede, sin embargo, generan desafíos, los cuales demandan que su adopción se proyecte siguiendo una serie de principios éticos de forma tal de mantener la tutela a derechos fundamentales, respetar valores democráticos, prevenir o disminuir los riesgos, fomentar la innovación y el diseño centrado en las personas.

Para explicar de manera ordenada cómo juegan estos principios, se los enmarcará en una línea temporal que contempla el ciclo de vida de la inteligencia artificial.

3.1 ¿Cómo es recomendable concebir a la inteligencia artificial?

El momento de partida, previo al ciclo, refiere a la concepción de la inteligencia artificial. Es decir, cómo se debe concebir, cómo se la comprende antes de trabajar con ella. Este punto resulta relevante dada la tendencia humana a antropomorfizar la tecnología. En este sentido, un aspecto recomendable viene dado por **diferenciar claramente los conceptos de responsabilidad y ejecución**.

Cuando se contratan servicios tecnológicos, lo que se transfiere al proveedor es la ejecución de distintas tareas pero no la responsabilidad de su efectiva concreción, con la inteligencia artificial sucede lo mismo. Cuando se utilizan algoritmos de inteligencia artificial, al igual que antes, se está trasladando la ejecución, pero no la responsabilidad. Es decir, la inteligencia artificial únicamente lleva a cabo una ejecución sin intención propia y de manera reactiva a una solicitud humana, quien ha decidido programarla, entrenarla e implementarla con un destino de uso específico con el fin de que ejecute distintas acciones.

En consecuencia, **surge que un algoritmo no posee autodeterminación y/o agencia para tomar decisiones libremente** (aunque muchas veces en el lenguaje coloquial se utiliza el concepto de "decisión" para describir una clasificación ejecutada por un algoritmo luego de un entrenamiento), **y por ende no se le pueden atribuir responsabilidades de las acciones que se ejecutan a través de dicho algoritmo en cuestión.**

Dicho con otras palabras, para que una persona humana pueda ser jurídicamente responsable sobre las decisiones que tome para realizar una o más acciones, debe existir discernimiento (plenas facultades mentales humanas), intención (pulsión o deseo humano) y libertad (para actuar de manera calculada y premeditada). Por lo tanto, para evitar caer en antropomorfismos que podrían dificultar eventuales regulaciones y/o atribuciones equivocas, resulta importante establecer la concepción de las inteligencias artificiales como artificios, es decir, como tecnología, una cosa, un medio artificial para lograr objetivos humanos pero que no deben confundirse con una persona humana. **Es decir, el algoritmo puede ejecutar, pero la decisión debe necesariamente recaer sobre la persona y por lo tanto, también la responsabilidad.**

Antecedentes internacionales.

Desde el momento mismo de la concepción, también resulta relevante abordar ciertos principios que todos los actores involucrados deben cumplir, los cuales debieran ser tomados como principios de diseño, desarrollo, implementación y uso de la inteligencia artificial. En este sentido la Organización de Naciones Unidas (ONU) a través de la Organización de las Naciones Unidas para la Educación, la Ciencia y la Cultura (UNESCO) emitió la Recomendación sobre la Ética de la Inteligencia Artificial, a la que adhirieron todos los países miembros en la Asamblea General de noviembre de 2021, entre los cuales se encuentra Argentina. Dicha recomendación, contiene un conjunto de principios que se transcriben de manera resumida a continuación.

> **Proporcionalidad e inocuidad.** Debería reconocerse que las tecnologías de la IA no garantizan necesariamente, por sí mismas, la prosperidad de los seres humanos ni del medio ambiente y los ecosistemas. En caso de que pueda producirse cualquier daño para los seres humanos, debería garantizarse la aplicación de procedimientos de evaluación de riesgos y la adopción de medidas para impedir que ese daño se produzca.
>
> **Seguridad y protección.** Los daños no deseados (riesgos de seguridad) y las vulnerabilidades a los ataques (riesgos de protección) deberían ser evitados y deberían tenerse en cuenta, prevenirse y eliminarse a lo largo del ciclo de vida de los sistemas de IA para garantizar la seguridad y la protección de los seres humanos, del medio ambiente y de los ecosistemas.
>
> **Equidad y no discriminación.** Los actores de la IA deberían promover la diversidad y la inclusión, garantizar la justicia social, salvaguardar la equidad y luchar contra todo tipo de discriminación, de conformidad con el derecho internacional. Los actores de la IA deberían hacer todo lo razonablemente posible por reducir al mínimo y evitar reforzar o perpetuar aplicaciones y resultados discriminatorios o sesgados a lo largo del ciclo de vida de los sistemas de IA, a fin de garantizar la equidad de dichos sistemas.
>
> **Sostenibilidad.** Debería llevarse a cabo con pleno conocimiento de las repercusiones de dichas tecnologías en la sostenibilidad la evaluación continua de los efectos humanos, sociales, culturales, económicos y ambientales de las tecnologías de la IA.
>
> **Derecho a la intimidad y protección de datos.** Es importante que los datos para los sistemas de IA se recopilen, utilicen, compartan, archiven y supriman de forma consistente con el derecho internacional y acorde a los valores y estos principios enunciados, respetando al mismo tiempo los marcos jurídicos nacionales, regionales e internacionales pertinentes.
>
> **Supervisión y decisión humanas.** Puede ocurrir que, en algunas ocasiones, los seres humanos decidan depender de los sistemas de IA por razones de eficacia, pero la decisión de ceder el control en contextos limitados seguirá recayendo en los seres humanos, ya que estos pueden recurrir a los sistemas de IA en la adopción de decisiones y en la ejecución de tareas, pero un sistema de IA nunca podrá reemplazar la responsabilidad final de los seres humanos y su obligación de rendir cuentas.
>
> **Transparencia y explicabilidad.** La transparencia y la explicabilidad de los sistemas de IA suelen ser condiciones previas fundamentales para garantizar el respeto, la protección y la promoción de los derechos humanos, las libertades fundamentales y los principios éticos. Las personas deberían tener la oportunidad de solicitar explicaciones e información al responsable de la IA o a las instituciones del sector público correspondientes. Dichos responsables deberían informar a los usuarios cuando un producto o servicio se proporcione directamente o con la ayuda de sistemas de IA de manera adecuada y oportuna.
>
> **Responsabilidad y rendición de cuentas.** Deberían elaborarse mecanismos adecuados de supervisión, evaluación del impacto, auditoría y diligencia debida, incluso en lo que se refiere a la protección de los denunciantes de irregularidades, para garantizar la rendición de cuentas respecto de los sistemas de IA y de su impacto a lo largo de su ciclo de vida.

> **Sensibilización y educación.** La sensibilización y la comprensión del público respecto de las tecnologías de IA y el valor de los datos deberían promoverse mediante una educación abierta y accesible, la participación cívica, las competencias digitales y la capacitación en materia de ética del uso de la IA, la alfabetización mediática e informacional y la capacitación dirigida conjuntamente por los gobiernos, las organizaciones intergubernamentales, la sociedad civil, las universidades, los medios de comunicación, los dirigentes comunitarios y el sector privado, y teniendo en cuenta la diversidad lingüística, social y cultural existente, a fin de garantizar una participación pública efectiva.
>
> **Gobernanza y colaboración adaptativas de múltiples partes interesadas.** La participación de las diferentes partes interesadas a lo largo del ciclo de vida de los sistemas de IA es necesaria para garantizar enfoques inclusivos en la gobernanza de la IA. Entre estas se encuentran, los gobiernos, las organizaciones intergubernamentales, la comunidad técnica, la sociedad civil, los investigadores y los círculos universitarios, los medios de comunicación, los responsables de la educación, los encargados de formular políticas, las empresas del sector privado, las instituciones de derechos humanos y los organismos de fomento de la igualdad, los órganos de vigilancia de la lucha contra la discriminación y los grupos de jóvenes y niños, entre otros.

Si bien la citada Recomendación es actualmente la de mayor adhesión por parte de gobiernos, se han desarrollado y gestado otros encuentros con distintos actores del ecosistema de la IA a los efectos de abordar y consensuar principios comunes. Así en enero de 2017, se llevó a cabo la Conferencia de Asilomar organizada por el Instituto "Future of Life"[1] con el objetivo de visibilizar la visión de la academia e industria sobre las oportunidades y amenazas que crea la IA. En ese marco, los participantes realizaron diversos aportes en función de los cuales se compilaron una lista de 23 principios sobre cómo se debe administrar la IA, basados en tres ejes: (i) problemas de investigación, (ii) ética y valores y (iii) problemas a largo plazo.

Entre los principios que se enuncian en la primera categoría, se destaca el Principio 4, relativo a fomentar una cultura de cooperación, confianza y transparencia entre los investigadores y desarrolladores de IA y el Principio 5, dirigido a promover la cooperación de los equipos que desarrollan sistemas de IA para evitar tomar atajos en los estándares de seguridad.

En lo relativo a la ética y valores comprometidos, se resaltan los siguientes principios:

> **6) Seguridad:** los sistemas de IA deben ser seguros y protegidos durante toda su vida operativa, y de manera verificable cuando sea aplicable y factible.
>
> **7) Transparencia de fallas:** si un sistema de IA causa daño, debería ser posible determinar por qué.
>
> **9) Responsabilidad:** los diseñadores y constructores de sistemas avanzados de IA son partes interesadas en las implicaciones morales de su uso, mal uso y acciones, con la responsabilidad y la oportunidad de dar forma a esas implicaciones.
>
> **10) Alineación de valores:** los sistemas de IA altamente autónomos deben diseñarse de modo que se pueda garantizar que sus objetivos y comportamientos se alineen con los valores humanos a lo largo de su operación.

Respecto a las dificultades que la IA puede representar a largo plazo, en la Conferencia se referenciaron aquellos ligados a:

> **11) Valores humanos:** los sistemas de IA deben diseñarse y operarse de modo que sean compatibles con los ideales de dignidad humana, derechos, libertades y diversidad cultural.
>
> **12) Privacidad personal:** las personas deben tener derecho a acceder, administrar y controlar los datos que generan, dado el poder de los sistemas de inteligencia artificial para analizar y utilizar esos datos.
>
> **13) Libertad y privacidad:** la aplicación de la IA a los datos personales no debe restringir injustificadamente la libertad real o percibida de las personas.

14) Beneficio compartido: las tecnologías de IA deberían beneficiar y empoderar a tantas personas como sea posible.

15) Prosperidad compartida: la prosperidad económica creada por la IA debe compartirse ampliamente para beneficiar a toda la humanidad.

16) Control humano: los humanos deben elegir cómo y si delegar decisiones a los sistemas de IA para lograr los objetivos elegidos por los humanos.

En mayo de 2019, los 36 países miembros de la OCDE, junto con Argentina, Brasil, Colombia, Costa Rica, Perú y Rumanía, suscribieron los Principios de la OCDE sobre IA, en el marco de la Reunión del Consejo de Ministros de la Organización, con el objetivo de guiar a los gobiernos, organizaciones e individuos para que en el diseño y la gestión de los sistemas de IA, se prioricen los intereses de las personas, y se garantice la responsabilidad por su correcto funcionamiento.

Elaborados a partir de las orientaciones proporcionadas por gobiernos, instituciones académicas, representantes del sector privado, la sociedad civil, organismos internacionales, la comunidad tecnológica y sindicatos, se consensuaron los siguientes cinco principios, basados en valores para el desarrollo responsable de los sistemas de IA:

Crecimiento inclusivo, desarrollo sostenible y bienestar. Las partes interesadas deben comprometerse de forma proactiva en la gestión responsable de la inteligencia artificial confiable que busque resultados beneficiosos para las personas y el planeta, como el aumento de las capacidades humanas y la mejora de la creatividad, el fomento de la inclusión de las poblaciones subrepresentadas, la reducción de las desigualdades económicas, sociales, de género y todo tipo, y la protección de los entornos naturales, reforzando así el crecimiento integrador, el desarrollo sostenible y el bienestar.

Valores y equidad centrados en el ser humano. Los actores de la IA deben respetar el Estado de Derecho, los derechos humanos y los valores democráticos durante todo el ciclo de vida del sistema de IA. Estos valores incluyen la libertad, la dignidad y la autonomía, la privacidad y la protección de datos, la no discriminación y la igualdad, la diversidad, la equidad, la justicia social y los derechos laborales reconocidos internacionalmente. Para ello, los actores de la IA deben aplicar mecanismos y salvaguardias, como la capacidad de determinación humana, que sean adecuados al contexto y coherentes con el estado del arte.

Transparencia y explicabilidad. Los actores de la IA deben comprometerse con la transparencia y la divulgación responsable en relación a los sistemas de la IA. Para este fin, deben proporcionar información significativa, adecuada al contexto y coherente con el estado del arte:

i) para fomentar una comprensión general de los sistemas de IA,

ii) para que las partes interesadas sean conscientes de sus interacciones con los sistemas de IA, incluso en el lugar de trabajo,

iii) para permitir que los afectados por un sistema de inteligencia artificial entiendan el resultado, y ;

iv) para permitir que aquellos afectados negativamente por un sistema de IA desafíen su resultado basado en información clara y fácil de entender sobre los factores, y la lógica que sirvió de base para la predicción, recomendación o decisión.

Robustez, seguridad y protección. Los sistemas de IA deben ser robustos, seguros y protegidos a lo largo de todo su ciclo de vida para que, en condiciones de uso normal, uso previsible o uso indebido, u otras condiciones adversas, funcionen adecuadamente y no planteen riesgos de seguridad irrazonables. Con este fin, los actores de la IA deben garantizar la trazabilidad, incluso en relación con los conjuntos de datos, los procesos y las decisiones tomadas durante el ciclo de vida del sistema de IA, para permitir el análisis de los resultados del sistema de IA y las respuestas a las preguntas, de forma adecuada al contexto y

coherente con el estado del arte. Los actores de la IA deben, en función de sus roles, el contexto y su capacidad de actuación, aplicar un enfoque sistemático de gestión de riesgos a cada fase del ciclo de vida del sistema de IA de forma continua para abordar los riesgos relacionados, incluida la privacidad, la seguridad digital, la seguridad y la parcialidad.

Responsabilidad.
Los actores de la IA deben ser responsables del correcto funcionamiento de los sistemas de IA y del respeto de los principios anteriores, en función de sus roles, del contexto y en consonancia con el estado del arte.

Asimismo, en el marco de la Reunión Ministerial del G20 sobre Comercio y Economía Digital que tuvo lugar en el mes de junio de 2019, y bajo la premisa de que las tecnologías digitales innovadoras siguen brindando inmensas oportunidades para el desarrollo económico y productivo de las naciones, y al mismo tiempo creando desafíos, el G20 adoptó los Principios de IA centrados en el ser humano, los cuales se replicaron fundamentalmente en los Principios de IA de la OCDE mencionados precedentemente.

Expuestos los principios que se recomienda incorporar a todas las fases del diseño e implementación del proyecto de IA y algunas de las instancias internacionales en las que se viene abordando la temática, se desarrollarán los puntos principales a abordar en cada etapa y la información con la que se debería contar.

3.2 ¿Qué es recomendable hacer antes de comenzar con el ciclo de la IA?

Antes de pensar en el diseño, desarrollo, implementación y/o en el uso de la inteligencia artificial, es recomendable trabajar sobre algunas cuestiones relevantes y que deben ser definidas antes de abordar la resolución de problemas con este tipo de tecnologías.

3.2.1 Conformar un equipo humano diverso y multidisciplinario

La diversidad de conocimientos y de perspectivas en estos equipos es fundamental para abordar los desafíos éticos, comprender las implicaciones sociales, prioirzar soluciones centradas en el usuario, evitar sesgos y discriminación, y fomentar la innovación.

Contar con un equipo humano con diversidad en las perspectivas, conocimientos y experiencias variadas en diferentes áreas, puede ayudar a lograr una comprensión más profunda de los usuarios y sus contextos, y por lo tanto, a abordar los desafíos de la IA desde diferentes puntos de vista. Puede conducir a soluciones más completas y creativas, más intuitivas y adaptadas a las necesidades reales de las personas. La diversidad en los equipos también puede ayudar a identificar y abordar sesgos

inherentes en los datos, algoritmos y decisiones automatizadas, contribuyendo a mitigar la discriminación y garantizar que los sistemas de IA sean diseñados y/o implementados de manera responsable, justa y equitativa.

Además, siempre es recomendable generar canales de comunicación con actores externos al gobierno, que puedan ser consultados y escuchados aunque no se involucren directamente en el desarrollo, implementación o ejecución del proyecto. Por ejemplo, actores de la sociedad civil, de las universidades, la academia, el empresariado, especialistas en ética y en las disciplinas involucradas, entre otros.

3.2.2 ¿Cuál es el nivel de concientización existente en la organización?

Es importante partir comprendiendo cuál es el nivel de conocimiento de las personas que integran la organización sobre el tema de inteligencia artificial. Pero no solamente respecto de conocimiento técnico del tema sino también sobre diferentes aspectos éticos relacionados con el modelo de adopción que se utilizará, destino de uso y control humano, gestión de los riesgos, así como también las mejores prácticas para la innovación pública.

En consecuencia, es recomendable comenzar realizando concientización a través de distintos medios, tales como campañas de comunicación, charlas y capacitaciones, describiendo la adhesión a los principios, indicando las acciones de capacitación y reubicación de puestos de trabajo en función del tipo de modelo de adopción a utilizar en cada caso, así como la existencia de los humanos como agentes de control y/o interpretación de resultados y toma de decisiones. Estos aspectos, y otros similares que se difundan en el mismo sentido, contribuirán a bajar la resistencia de adopción de este tipo de tecnologías aumentando las probabilidades de éxito, sostenibilidad e innovación.

3.2.3 ¿Es excluyente el uso de inteligencia artificial para el problema que se quiere resolver?

Dado que el uso de inteligencia artificial conlleva distintos riesgos, y que resulta recomendable no solucionar problemas comenzando con el único propósito de "utilizar la inteligencia artificial", es importante que se realice la exploración de diferentes tipos de tecnología antes de concluir que la inteligencia artificial ofrece la mejor solución a la problemática que se quiere resolver. En ciertas ocasiones, pueden emplearse otras soluciones de software más simples, menos riesgosas e igualmente eficientes para dar respuesta al mismo desafío.

3.2.4 ¿Cuál es el alcance de los modelos básicos para la adopción de IA?

Se pueden definir básicamente dos tipos de modelos sobre los que se puede optar para adoptar inteligencia artificial. Uno de ellos es el de automatización. Básicamente

consiste en sustituir trabajo humano por hardware, software y/o algoritmos para realizar ciertas tareas, operaciones o procesos repetitivos, secuenciales, de distinto grado de complejidad, pero que responden a problemas debidamente tipificados. En este tipo de modelo se pueden reducir los tiempos de procesamiento de distintos pedidos y responder a ellos automáticamente, siempre y cuando estos se encuentren tipificados y mantengan un mismo tipo de datos (que es con el que trabajará la inteligencia artificial).

Es importante **definir que este modelo de adopción,** como cualquier otro, no se basa simplemente en la incorporación, configuración y puesta en marcha de una nueva tecnología, debe contemplar también otros aspectos organizacionales, entre los cuales se encuentra uno fundamental y decisivo: la capacitación y reubicación de las personas cuyas tareas serán automatizadas, estos aspectos deben estar abordados en una planificación anterior al inicio de la automatización de procesos.

Asimismo, independientemente del grado de automatización alcanzado, s**iempre resulta indispensable contar con la intervención humana para verificar y controlar la correcta ejecución de los procesos automatizados;** para ofrecer una vía personal ante la demanda de aquellos personas que no posean los medios tecnológicos para realizar una o más solicitudes de manera automática; para evaluar si existe degradación del algoritmo; y para observar nuevos tipos de solicitudes que no estén contempladas en la automatización.

El otro modelo de adopción implica una participación humana mucho más allá del control, se trata de un modelo en donde máquinas y humanos colaboran para resolver problemas, se lo conoce como modelo humano-máquina, "inteligencia aumentada", o también "human in the loop". Todas ellas son expresiones de soluciones tecnológicas conceptualmente similares que aprovechan las capacidades y características únicas de los humanos en varios momentos del ciclo de vida de la IA. En este modelo las tecnologías de inteligencia artificial contribuyen con una parte importante del trabajo que resulta muy costosa para los humanos, como por ejemplo el pensamiento estadístico basado en grandes volúmenes de datos, pero los resultados de estos análisis son presentados a humanos que aportan el trabajo difícil para las máquinas, y complementan el análisis realizado por la máquina tomando decisiones, o volviendo a solicitar nuevos análisis e informes para poder tomar esas decisiones de una manera mejor informada.

Comprender las diferencias de cada modelo de adopción y las acciones que corresponden realizar en cada caso resulta clave para comprender los posibles riesgos que existan en el destino de uso de esta tecnología.

3.2.5 ¿Cuál es el destino de uso de la IA, y cómo se realizará el control humano?

La versatilidad de las tecnologías de inteligencia artificial permiten su implementación en una variedad muy amplia respecto de los destinos de uso. Pero cada uno de los destinos de uso conlleva diferentes niveles de riesgos, lo que implica a su vez varios niveles de tratamiento de riesgo y de control humano (por ejemplo, control de auditabilidad y trazabilidad). En algunas áreas esto resulta muy importante dado que ciertos destinos de uso, por ejemplo, en la ciberseguridad, pueden representar potenciales riesgos si su uso no es controlable, auditable y trazable. Dichos riesgos impactan negativamente en la transparencia y consecuente rendición de cuentas.

Otra vez, la tecnología ejecuta distintas acciones que le son ordenadas por una persona humana con intención, discernimiento y libertad de acción. Es por esto que, para el eventual caso de que estas intenciones proporcionadas por uno o más humanos no estén alineadas con el bien común y los derechos humanos, deben existir instrumentos de control diseñados para identificar la responsabilidad y rendición de cuentas.

3.2.6 ¿Qué es un análisis de premortem?

Una manera interesante de identificar eventuales riesgos en un proyecto de inteligencia artificial es utilizar la técnica de premortem. Al igual que en la prospectiva, la idea es imaginar un futuro, pero en lugar de imaginar el futuro deseado se imagina uno en donde luego de implementar ese proyecto, los resultados fueron distintos a los esperados. Es decir, imaginar que el proyecto fue un fracaso. Una vez que las personas se sitúan en ese futuro indeseado se analiza por qué o dónde falló el proyecto.

En este análisis participa todo el equipo diverso y multidisciplinario que diseñará el proyecto de inteligencia artificial, y cada persona intentará identificar las causas por las cuales el proyecto fracasó. Luego, cada participante comunica sus hallazgos y entre todos clasifican las causas según su probabilidad de ocurrencia y su impacto. Seleccionan aquellas con probabilidad más alta y/o impacto más negativo, las identifican con un nombre, para luego gestionar el riesgo que cada una representa. El tratamiento de riesgos puede optar entre aceptar, mitigar, eliminar o transferir estos riesgos identificados.

Esta técnica permite, de una manera sencilla y rápida, identificar potenciales causas de fracaso, riesgos del proyecto y cómo poder tratarlos durante la fase de diseño.

3.2.7 ¿Y ahora qué...?

Luego de responder estas preguntas y trabajar sobre las recomendaciones antes mencionadas, pero antes de iniciar el ciclo de vida de la IA, resulta interesante identificar a los actores que participarán en la adopción de esta tecnología y comprender el aporte que cada uno de ellos realiza dentro del proceso de innovación pública [2].

[2] Para facilitar la identificación y compresión, se recomienda el uso de la Guía para el diseño y la adopción tecnológica de proyectos de innovación pública, que facilita la identificación, y comprensión de la participación e interdependencia de los actores involucrados.

3.3 Aspectos éticos a considerar dentro del ciclo de la IA

Dado que los aspectos éticos son propios de las personas, a lo largo de cada una de las etapas del ciclo de vida de la IA se debe asegurar que las personas que integren el equipo diverso y multidisciplinario a cargo del diseño (que es la primera actividad del ciclo de vida), conozcan y comprendan los aspectos éticos básicos necesarios involucrados.

Etapa N°1: Diseño y modelado de datos

Esta es la primera etapa del ciclo de vida de la IA. Se comienza con el diseño de los datos y los modelos involucrados. Es importante que desde esta primera etapa se incluyan como criterios de diseño aspectos éticos que facilitarán el cumplimiento de los principios definidos y aumentarán en consecuencia las probabilidades de éxito del proyecto.

3.3.1. Punto de partida común para el equipo diverso y multidisciplinario

Dado que cada una de las personas integrantes de un equipo diverso y multidisciplinario posee conocimientos variados con experiencias diferentes, resulta recomendable acordar de manera clara el propósito del proyecto. En consecuencia, todas y cada una de las personas integrantes deberán conocer, comprender, acordar, y comprometerse a llevar a cabo los siguientes aspectos mínimos:

> **a.** Los principios de diseño, desarrollo, implementación y uso ético de la inteligencia artificial, definidos por la UNESCO.
>
> **b.** El impacto en la sociedad en general y las necesidades a cubrir en los destinatarios en particular.
>
> **c.** Los potenciales riesgos evaluados por nivel de impacto y probabilidad de ocurrencia, y los tratamientos definidos para cada uno de ellos.
>
> **d.** Los mecanismos de transparencia y rendición de cuentas a utilizar para la trazabilidad y auditoría (ya sea de lo ejecutado por las máquinas y/o lo decidido por las personas).
>
> **e.** El rol, el alcance de las actividades y distribución de las responsabilidades de cada persona integrante del equipo.
>
> **f.** La definición y asignación formal de la persona responsable de asegurar la sostenibilidad del proyecto a lo largo del tiempo.
>
> **g.** El relevamiento y comprensión respecto de los diversos perfiles de personas destinatarias ya sea toda la ciudadanía o parte de ella (contribuyentes, empleados públicos, beneficiarios de seguridad social, estudiantes, pacientes, etc.). Esto incluye, los aspectos que eventualmente podrían dar lugar a distintos sesgos. Se recomienda asimismo que cada uno de estos perfiles estén representados por al menos una persona.
>
> **h.** El relevamiento y comprensión de los alcances, implicancias e impacto de la normativa involucrada.
>
> **i.** La documentación, registro y socialización de la experiencia para promover buenas prácticas y lecciones aprendidas necesarias para el aprendizaje organizacional y la innovación pública.

3.3.2 Consideraciones éticas respecto al diseño de los datos

No se debe subestimar el tratamiento que corresponde otorgar a los datos involucrados en el proyecto. Estos deben ser tratados por profesionales sobre la base de las buenas prácticas de la ciencia de datos. La calidad de los datos que se utilicen determinará no solamente la calidad del modelo entrenado sino que también contribuirá con el éxito del proyecto. Los datos son la materia prima para construir el modelo entrenado de inteligencia artificial que se utilizará para que, al ingresar diferentes entradas, se obtenga una respuesta correcta.

En este sentido, se deben considerar los distintos aspectos que se detallan a continuación para poder realizar un diseño ético de datos.

a. La clasificación de los datos según su confidencialidad. Es recomendable que exista un acuerdo respecto de dicha clasificación y que la misma sea elaborada sobre la base de normas internacionales relativas a la seguridad de la información. A modo de ejemplo se esboza una clasificación general consensuada a nivel internacional.

i. Datos confidenciales. Refiere a aquellos datos o información sensible que pueden referirse a cuestiones de inteligencia, defensa, seguridad, y otros similares.

ii. Datos personales. Refiere a aquellos datos o información de las personas que específicamente han sido definidos como tales por la normativa vigente.

iii. Datos internos. Refiere a aquellos datos o información de gestión interna, que no resultan ser ni confidenciales ni personales pero que no catalogan como información pública.

iv. Datos públicos. Refiere a aquellos datos o información de dominio público que generalmente se encuentran disponibles, ya sea como conjuntos de datos abiertos, y/o como contenidos en sitios web.

b. Las fuentes de datos que se utilizarán para diseñar y construir el set de datos correspondiente al entrenamiento del modelo.

i. Datos disponibles en internet. Es el caso menos costoso, no obstante, se debe tener en cuenta que existe un alto grado de probabilidad de que los mismos sean inexactos, posean sesgos de diferente tipo, puedan estar sujetos a propiedad intelectual, entre varios aspectos que no sólo degradan la calidad de los datos, sino que también impiden crear datos de entrenamiento de manera ética.

ii. Datos existentes en la organización. En este caso, se requiere dimensionar los costos asociados: previo a su uso se debe tener en cuenta la clasificación según su confidencialidad, derechos de uso, consentimiento de los titulares, posibilidad de anonimizar dichos datos y otros aspectos que establezca la normativa vigente.

iii. Datos solicitados a terceras partes. En este caso, también se deben dimensionar los costos asociados, ya que no están disponibles en internet, y para conseguirlos y utilizarlos se debe tener en cuenta la clasificación según su confidencialidad, derechos de uso, consentimiento del titular, la trazabilidad de todo el proceso de obtención y creación de datos para entrenamiento y otros aspectos que establezca la normativa vigente.

c. La calidad de los datos. En todos los casos se deberá asegurar la calidad de los datos. Por ejemplo, evitando la existencia de sesgos, verificar que sean precisos o que reflejen la realidad que pretenden representar, entre otras. Este tratamiento debe ser llevado a cabo por profesionales de las ciencias de datos, quienes deberán evaluar continuamente los distintos conjuntos de datos con el fin de asegurar que el entrenamiento de los modelos de IA se realice según los principios de la UNESCO antes citados.

3.3.3 Consideraciones éticas respecto del diseño de los modelos

Los modelos deben ser diseñados de manera tal de que no introduzcan sesgos propios de su concepción. Por ejemplo, a través de una definición que omita aspectos del contexto que privilegien o perjudiquen a unas personas sobre otras, o bien utilizando algoritmos que funcionan mejor con ciertas variables o características que con otras, que podrían generar eventuales imprecisiones o sesgos en los resultados.

En línea con los principios de la UNESCO, los modelos deben ser transparentes y explicables. Es decir, la ejecución que llevó a su resultado debe poder ser comprendida por personas que operan dichos sistemas, para que éstas, a su vez,

puedan tomar decisiones con esos resultados, y además, para poder explicarle a las personas afectadas por la decisión tomada o a terceros cómo se llegó a dicho resultado de forma clara.

Etapa N°2: Verificación/Validación

En una segunda etapa, dentro del ciclo de vida de IA, resulta importante realizar las correspondientes verificaciones y validaciones de los diseños realizados en la primera etapa. Para ello, se debe tener en cuenta el diseño del equipo, de los datos y de los modelos involucrados. Estas verificaciones y validaciones se realizan teniendo en cuenta tanto los principios definidos por la UNESCO, como la interacción de las personas destinatarias con los prototipos diseñados (primeras soluciones conceptuales del o de los modelos entrenados), en condiciones similares a las que tendrá su implementación definitiva.

3.3.4 ¿Cómo se validan los conocimientos éticos específicos necesarios para el proyecto de IA?

Asimismo, luego de las capacitaciones (o concientizaciones) realizadas, para que los integrantes puedan conocer, comprender, acordar, y comprometerse a llevar a cabo aspectos éticos mínimos necesarios, las mismas puedan ser volcadas por escrito y firmadas en un acta de compromiso ético del proyecto de IA.

3.3.5 Ética de datos

Los conjuntos de datos armados específicamente para entrenar modelos de inteligencia artificial deben ser validados de manera previa a la implementación en campo. Las personas del equipo que sean profesionales de las ciencias de datos deberán ser las encargadas de evaluar la calidad de datos que se utilizarán para entrenar los modelos de IA.

Se deberá establecer una clasificación de riesgo (por ejemplo, de tres niveles tipo semáforo, o con valores del uno al cinco) respecto de cuánto se ajustan a los principios de la UNESCO antes citados.

3.3.6 ¿Cómo se validan los aspectos relativos al diseño ético de los modelos de IA?

Las pruebas con prototipos deben ser llevadas a cabo por profesionales con conocimientos de metodologías ágiles y es recomendable que el equipo diverso y multidisciplinario encargado del proyecto esté presente en la realización de las mismas. En esta instancia, se validarán también los modelos entrenados en condiciones similares a las que tendrán en su implementación. Para realizar dichas pruebas se utilizarán uno o más prototipos de los modelos entrenados, con una interfaz mínima de usuario pero de similar aspecto al definitivo.

Para probar los modelos se invitará al menos a una persona de cada perfil definido para que el equipo de trabajo pueda observar cómo se utiliza el modelo y aprovechar esa interacción para verificar distintos aspectos éticos del diseño. Por ejemplo, que no existan sesgos, que la persona encargada de tomar la decisión pueda comprender el resultado de la ejecución del modelo (en el caso del modelo de adopción humano- máquina), que se pueda explicar de manera sencilla a las personas afectadas, validando que éstas comprendan clara y acabadamente el resultado del modelo y la consecuente decisión humana.

Es decir, en esta prueba con prototipos, se validarán distintos aspectos éticos tales como; la congruencia entre los resultados y las expectativas del diseño; la ausencia de sesgos; la explicabilidad del modelo, así como otros aspectos éticos del diseño que sean susceptibles de mejoras. Se deberá establecer una clasificación de riesgo (por ejemplo, de tres niveles tipo semáforo, o con valores del uno al cinco) respecto de cuánto se ajustan a los principios de la UNESCO antes citados.

3.3.7 ¿Cómo se registran las verificaciones/validaciones?

Todas las acciones y decisiones que se tomen dentro de un proyecto de IA, incluidas las relativas a las verificaciones y validaciones de los aspectos éticos realizados en la etapa del diseño, deben ser registradas. Este punto resulta crítico para poder cumplir con los principios relativos a la transparencia y rendición de cuentas correspondientes a las acciones y decisiones involucradas en cada proyecto de IA. Se deberá utilizar un medio de registro formal que permita realizar la trazabilidad y auditorías de todas y cada una de las acciones de verificación y validación.

Etapa N°3: Implementación

En esta etapa entran en juego los profesionales de infraestructura que forman parte del equipo diverso y multidisciplinario del proyecto IA. En este caso existen opciones de implementación que pueden estar basadas en la contratación de servicios de nube, en el despliegue de infraestructura propia o en una solución que contemple ambas opciones.

Cualquiera sea el caso, se deberá asegurar que la implementación permita: establecer un grado adecuado de seguridad de la información; realizar trazabilidad sobre las acciones y decisiones ocurridas en el proyecto identificando a las personas que las llevaron a cabo; realizar auditorías (este punto es especialmente importante cuando se contratan servicios de nube) y ofrecer al usuario facilidades de accesibilidad a las tecnologías de información y comunicaciones (TIC).

3.3.8 ¿Cómo establecer un grado adecuado de seguridad de la información?

Resulta importante que se lleven a cabo las mejores prácticas relativas a la seguridad de la información. Para ello, los responsables de la seguridad de la información que forman el equipo de trabajo diverso y multidisciplinario, deberán tener en cuenta los siguientes aspectos:

a. El relevamiento, conocimiento y comprensión del alcance de los estándares y normativas internacionales, y mejores prácticas en materia de seguridad de la información.

b. El relevamiento, conocimiento y comprensión de la normativa vigente en materia de seguridad de la información.

c. La utilización de aplicaciones accesorias encargadas de gestionar los registros (logueos, eventos, etc.) de los sistemas involucrados de manera tal de facilitar el tratamiento de eventuales incidentes de seguridad; automatizar la creación de informes de auditoría; y mejorar la transparencia a través del control de las personas que acceden a los sistemas, de las aplicaciones y de los equipos.

d. La realización de diferentes tests para hallar vulnerabilidades de seguridad que pudieran ocasionar eventuales incidentes no deseados.

e. La participación del área o de la autoridad con responsabilidad primaria en materia de seguridad de la información, que entiende en todos los aspectos relativos a la ciberseguridad y a la protección de las infraestructuras críticas de información, así como también a la generación de capacidades de prevención, detección, defensa, respuesta y recupero ante incidentes de seguridad informática. Esto es particularmente importante en el caso de que la institución adoptante del desarrollo basado en IA no disponga de un área específica de seguridad de la información.

3.3.9 ¿Qué aspectos se deben tener en cuenta a la hora de establecer la trazabilidad?

Los sistemas involucrados en el despliegue de infraestructura para la implementación del proyecto de IA, así como los procedimientos definidos para la gestión de los mismos, deben poseer los medios adecuados de registro de todas las acciones realizadas en el sistema (altas, bajas, modificaciones en la configuración) para todas las jerarquías y todos los perfiles de usuarios (Administrador, operador, usuarios, etc.), de manera tal de poder identificar fehacientemente a todas las personas que llevaron a cabo las distintas acciones y decisiones en el proyecto.

En el caso de despliegue a través de servicios de nube, ya sea total o parcial, corresponde comprender, de manera previa a la contratación, la trazabilidad que ofrecen los prestadores de servicios de nube para poder comprender si el alcance de la trazabilidad ofrecida permite instrumentar los principios éticos correspondientes a dicha materia.

3.3.10 ¿Qué aspectos se deben tener en cuenta para que los sistemas sean auditables?

Para garantizar el cumplimiento de los principios éticos es necesario auditar el modelo y la trazabilidad es la mejor herramienta para lograr este objetivo. Es clave poder identificar y comprender el registro de acciones, decisiones y/o cualquier otro evento que afecte a los sistemas involucrados del proyecto de IA.

En el caso de despliegue de soluciones on premise (dentro de la infraestructura de la organización), resulta importante, asegurar además del control del acceso a los sistemas, el control del acceso físico en donde se aloja la infraestructura involucrada.

En el caso de despliegue a través de servicios de nube, resulta importante comprender, de manera previa a la contratación, las facilidades de auditoría que ofrecen los prestadores de servicios de nube, para poder comprender si el alcance ofrecido permite instrumentar los principios éticos correspondientes a dicha materia.

3.3.11 ¿Qué aspectos se deben tener en cuenta para que los sistemas posean accesibilidad TIC?

Es necesario que se lleven a cabo las mejores prácticas propias de la accesibilidad TIC, ya sean a través de páginas web o aplicaciones móviles. Para ello, las personas profesionales encargadas de la accesibilidad TIC, que forman el equipo de trabajo diverso y multidisciplinario, deberán tener en cuenta los siguientes aspectos:

a. El relevo, conocimiento y comprensión del alcance de las normativas internacionales y mejores prácticas en materia de accesibilidad TIC.

b. El relevo, conocimiento y comprensión del alcance de la normativa nacional en materia de accesibilidad TIC.

c. La evaluación del sitio web. En el caso particular de la accesibilidad web, se recomienda utilizar aplicaciones específicas disponibles encargadas de evaluar la accesibilidad de sitios web que utilizarán los usuarios para acceder a los sistemas involucrados de manera tal de asegurar un nivel mínimo de accesibilidad (nivel A). Asimismo, se recomienda requerir la asistencia de la autoridad de aplicación de la Ley 26.653 de Accesibilidad Web.

Etapa N°4: Operación y mantenimiento

Los proyectos de innovación tecnológica no terminan con la implementación; la operación y mantenimiento constituye la etapa final del ciclo de vida de la IA. Un problema frecuente es que estas dos acciones, a pesar de su importancia, suelen no ser consideradas en el diseño de los proyectos. Estas tareas son las operaciones y el mantenimiento tanto de la infraestructura en donde se despliega la solución tecnológica basada en IA, así como también del propio modelo, dado que, por ejemplo, muchas veces los modelos se degradan y dejan de responder de manera correcta. Dichas acciones permiten que exista disponibilidad, continuidad, y sostenibilidad del servicio prestado a través de la solución de IA.

3.3.12 ¿Cómo podría realizar un monitoreo y qué se debe monitorear considerando el uso ético de la IA?

El monitoreo es una acción que se realiza en esta etapa para asegurarse de que todo funcione según lo esperado. Se pueden monitorear distintas variables que se elegirán según el propósito que se persiga. Por ejemplo, si lo que se busca es comprender si el modelo responde tal cual se validó en las pruebas con prototipos, se puede monitorear su desempeño a través de la medición de distintos parámetros de forma automática, y de manera manual, es decir, llevada a cabo por personas que inspeccionan y realizan valoraciones del comportamiento del modelo.

El monitoreo de valoraciones manuales, permite que las personas involucradas comprendan las salidas generadas por el modelo en función de las entradas provistas por los usuarios. Por lo tanto,

no sólo habilita la verificación del desempeño del modelo en términos de la calidad de la respuesta otorgada, sino también, respecto de eventuales sesgos que pueden haber sido omitidos o pasados por alto en el proceso de diseño y prueba. Asimismo, se pueden detectar otro tipo de resultados indeseables que, de no ser monitoreados, podrían tener distintos grados de impacto negativo o perjudicial en las personas.

Con este tipo de valoraciones también es posible comprender el grado de aplicabilidad que puede ofrecer el operador y el nivel de transparencia que puede brindar al usuario final.

3.3.13 ¿Qué aspectos generales se deben considerar respecto de la existencia de incidentes éticos?

Los incidentes éticos pueden ser causados por distintos motivos. Por ejemplo, pueden ser causados por un error humano involuntario en alguna de las etapas del ciclo de vida que ocasione un mal funcionamiento en una o más tecnologías involucradas, un uso intencional e indebido de una o más personas dentro de la organización, un uso indebido de los usuarios finales, un ataque a la seguridad de la organización (interno y/o externo), entre otros.

Si se receptaron los principios y recomendaciones incluidas en el presente documento se poseen las bases mínimas para poder brindar un correcto tratamiento a un eventual incidente ético cualquiera fuera su causa.

Siendo que la ocurrencia de incidentes no puede eliminarse, la correcta y completa documentación de los mismos será un insumo fundamental para poder tomar cuenta de los detalles y condiciones en que ocurrieron. Posteriormente, dichos registros

serán útiles para confeccionar los informes de rendición de cuenta necesarios y cumplir con los principios definidos por la UNESCO.

El tratamiento de incidentes permite aprender de ellos para evitar que se repitan, poniendo en evidencia aquellos aspectos que fallaron para poder corregirlos.

3.3.14 ¿Qué recaudos desde el punto de vista ético es recomendable considerar para el control de los usuarios internos?

Como cualquier otro sistema informático se deben realizar los controles mínimos necesarios de autenticación y autorización de usuarios internos independientemente del rol que posean (administrador, operador, usuario, etc.), se deberá evitar la existencia de usuarios genéricos tales como “mantenimiento”, “monitoreo”, etc., dado que no permiten identificar a la persona que los usa.

Los usuarios internos que no hayan formado parte del equipo diverso y multidisciplinario involucrado en el diseño del proyecto deben tener el mismo tratamiento que dichos integrantes. Es decir, todos y cada uno de los usuarios internos deben comprender de manera clara el propósito del proyecto y registrar formalmente su compromiso ético, ya sea en la administración, operación o simple uso de las tecnologías involucradas dentro del proyecto de IA.

Todos los cambios en las configuraciones, reemplazos, actualizaciones, mejoras, o cualquier acción efectuada en las tecnologías involucradas dentro del proyecto de IA deben ser planificadas, registradas y autorizadas formalmente por la persona responsable del proyecto (y/o del impacto de los servicios que se brindan a través de las tecnologías de IA) quien a su vez será quien rinda cuentas a las autoridades, al comité de ética (en caso de existir) y a distintos organismos de control y auditoría.

Ninguno de los cambios en las configuraciones, reemplazos, actualizaciones, mejoras, o cualquier acción efectuada en las tecnologías involucradas dentro del proyecto de IA debe realizarse de manera individual, privada, unilateral, discrecionalmente, y/o sin quedar formalmente registrado.

3.4 ¿Qué aspectos éticos se deben considerar fuera del ciclo de la IA?

El orden cronológico establecido en el presente documento, estableció los distintos aspectos éticos a considerar en diferentes momentos. El momento de la concepción de la IA, el momento previo al inicio del ciclo de la IA, y el momento en que transcurre el ciclo de la IA. Ahora toca trabajar sobre los aspectos éticos en el momento posterior al ciclo.

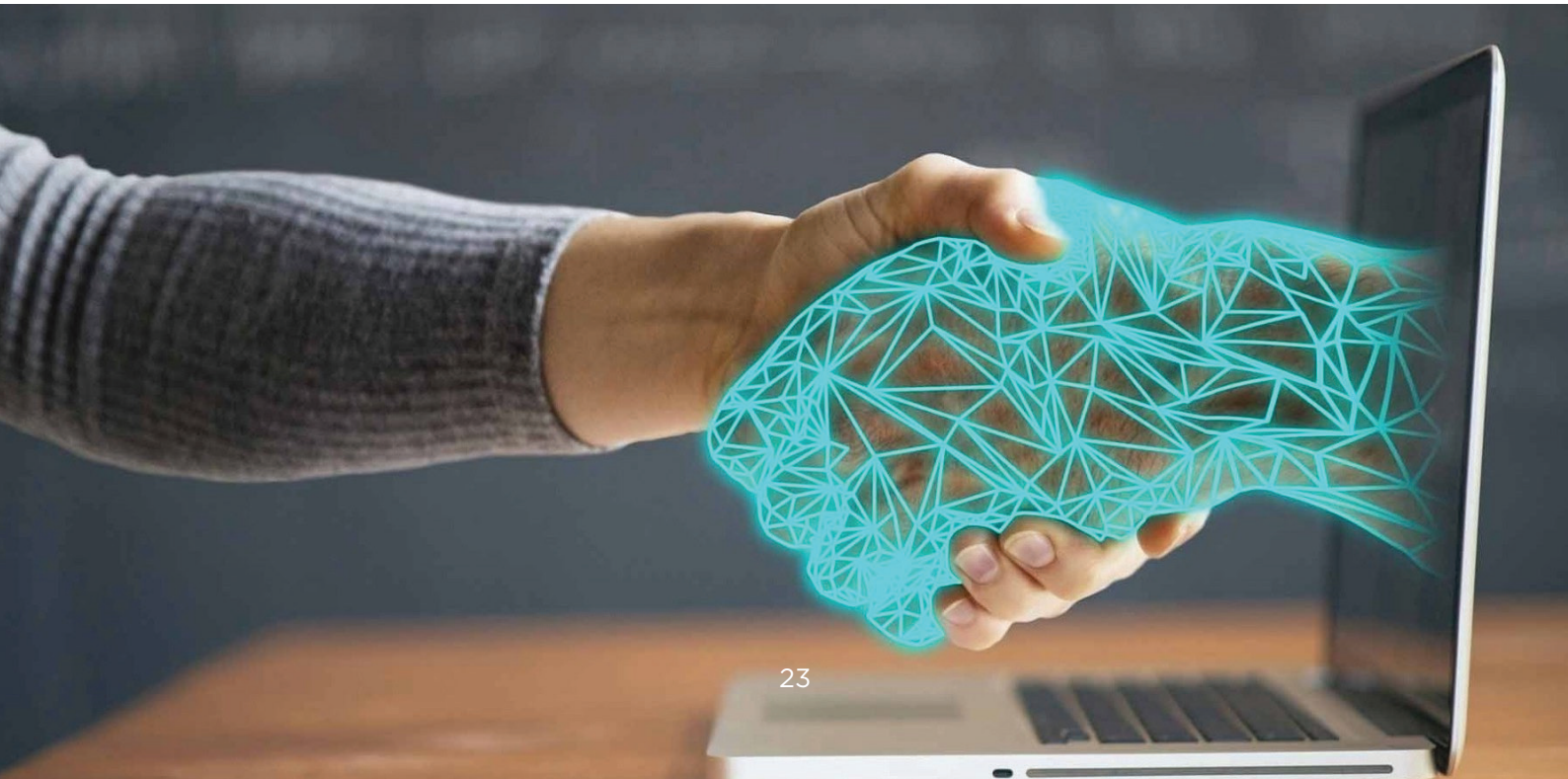
Por supuesto, la operación y el mantenimiento de los sistemas siguen vigentes para asegurar su disponibilidad y sostenibilidad, aunque en este momento algunas

cuestiones pueden cambiar: quizá el equipo diverso y multidisciplinario se disolvió; cambiaron las prioridades, los responsables, cambiaron los enfoques, etc.. No obstante, mientras el (o los) servicios que se prestan mediante las tecnologías de IA sigan vigentes, corresponde realizar las tareas de operación y mantenimiento, aunque las responsabilidades no se limitan a ellas.

Si bien en la etapa de diseño se relevaron diferentes riesgos y se planificaron distintos tratamientos para evitar su ocurrencia o impacto negativo, esto no impide su ocurrencia y potenciales perjuicios que se puedan derivar de ellos.

Las personas designadas formalmente como responsables deben actuar de manera inmediata y personalmente para comprender el alcance del perjuicio y arbitrar por los medios posibles las acciones necesarias para subsanar y/o revertir el perjuicio causado. Los medios necesarios para poder llevar a cabo dicha subsanación deben definirse previamente como procedimientos formales e institucionalizados. Las acciones de responsabilidad y rendición de cuentas involucradas deberán quedar registradas debidamente, y ser definidas como casos testigos para ser estudiados y discutidos formando parte de las lecciones aprendidas, necesarias para el aprendizaje organizacional y la mejora continua de procesos: ambos aspectos que favorecen a la innovación pública.

Sin excepción, para todos los casos en que existan servicios brindados a través de tecnologías de IA, se deberá establecer una vía humana (con atención presencial) para atender a aquellas personas que por su perfil y/o situación contextual no tengan acceso a los dispositivos y servicios tecnológicos básicos universales mínimos necesarios para poder ser usuarios de dichos servicios, o prefieran la atención de una persona humana.



Glosario

Adopción tecnológica: Es un requisito necesario para la innovación que ocurre tanto cuando las organizaciones son usuarios finales de una o más tecnologías y/o cuando son quienes ponen en marcha de una o un conjunto de tecnologías que contratadas mediante un proyecto de adopción tecnológica y estas resultan ser un medio efectivo para brindar servicios y/o instrumentar políticas públicas dado que son adoptadas por las personas destinatarias.

Algoritmos: Documentos de la OCDE los definen como conjuntos secuenciales exactos de comandos que se ejecutan sobre una entrada diseñada para generar una salida en un formato claramente definido. Los algoritmos se pueden representar en lenguaje sencillo, diagramas, códigos informáticos y otros lenguajes.

Aprendizaje automático (Machine learning): Naciones Unidas la define como una rama de la inteligencia artificial (IA) centrada en la creación de aplicaciones que aprenden de los datos y mejoran su precisión con el tiempo sin estar programadas para hacerlo. Documentos de la OCDE lo definen como un subconjunto de inteligencia artificial en el que las máquinas aprovechan los enfoques estadísticos para aprender de los datos históricos y hacer predicciones en situaciones nuevas.

Aprendizaje profundo (Deep learning): Documentos de la OCDE refieren a modelos de aprendizaje inspirados en las neuronas biológicas, no obstante, las redes neuronales no necesariamente aprenden igual que los humanos. Dichas redes organizan la computación a través de grandes colecciones de unidades computacionales simples. El término "profundo" se refiere al número de capas en la red. Hasta hace poco, la falta de poder de cómputo y datos de entrenamiento significaba que solo se podían explorar redes pequeñas. Varias décadas de investigación sobre mejoras de algoritmos, combinadas con unidades de procesamiento gráfico desarrolladas originalmente para videojuegos, finalmente permitieron entrenar grandes redes utilizando cantidades masivas de datos. Esto ha llevado a sistemas que funcionan mucho mejor que los enfoques anteriores en tareas como subtítulos de imágenes, reconocimiento facial, reconocimiento de voz y traducción automática de lenguaje natural.

Automatización (a través de IA): Sistemas de IA concebidos para automatizar tareas tipificadas, monótonas, masivas y repetitivas. La automatización representa una manera de adopción de la IA que debe ir acompañada de un proceso de recapacitación de las personas desplazadas por dicha automatización para su reubicación dentro de la organización.

Ciclos iterativos de prueba y error: Metodología de desarrollo orgánico que permite a los diseñadores y desarrolladores obtener retroalimentación en tiempo real sobre su trabajo y hacer ajustes rápidos y efectivos. También facilita la detección temprana de problemas y errores, para realizar correcciones antes de que se conviertan en problemas mayores. Esta metodología se utiliza ampliamente en el desarrollo de productos y servicios innovadores en diversos campos, incluyendo la tecnología, el diseño, la ingeniería, entre otros. Uno de estos ciclos está representado por la secuencia iterativa de crear-medir-aprender.

Ciencia de datos: Disciplina que mediante la combinación de modelos matemáticos y estadísticos, la programación computacional y las técnicas de visualización de datos, permite apoyar los procesos de toma de decisiones, por ejemplo, para diseñar proyectos de innovación pública, a partir del procesamiento de grandes volúmenes de datos.

Construcción de valor: Capacidad de la solución para proporcionar un beneficio significativo y medible objetivamente para las personas destinatarias. El servicio o política pública debe generar valor a las personas destinatarias, a través de por ejemplo, su capacidad de cubrir las necesidades, dificultades y frustraciones de esas personas creando o mejorando sus experiencias como usuarios de las tecnologías que se utilizaron como medio para instrumentar dichos servicios y/o políticas.

Creatividad: Desde una perspectiva individual refiere a la capacidad o habilidad de la persona para realizar aportes que son a la vez nuevos y valiosos. También se la puede entender como un proceso (compuesto por diferentes etapas), como productos (que deben poseer valor y novedad), como contextos (los cuales cultivados para favorecerla). También refiere a prácticas o acciones que una

persona realiza aprovechando su experiencia acumulada y conocimientos para interactuar con su contexto social y material, permitiéndole de esta manera, llevar a cabo dichos aportes que deberán ser nuevos y valiosos en los contextos para los cuales fueron creados.

Datos sesgados: Presencia de desequilibrios o distorsiones en los datos de entrenamiento, por ejemplo, utilizados para desarrollar un modelo de inteligencia artificial. Pueden deberse a una variedad de factores, como la falta de diversidad en los datos, la inclusión de datos incorrectos o incompletos, la exclusión de ciertas categorías de datos o la selección de datos que reflejan una realidad parcial o limitada. Por ejemplo, cuando un modelo de aprendizaje automático está entrenado con datos que no representan completamente la población a la que se aplica, puede dar lugar a predicciones incorrectas o sesgadas en el mundo real.

Desarrollo orgánico: Refiere a un proceso de creación y evolución de productos y/o servicios que se basa en ciclos iterativos de prueba y error. El proceso implica una continua retroalimentación y adaptación en función de los resultados obtenidos en cada ciclo, lo que permite una evolución natural y fluida del producto o servicio. Este enfoque es fundamental cuando se trabaja con metodologías ágiles, ya que permite una mayor flexibilidad y adaptabilidad en el proceso de diseño y creación de soluciones tecnológicas potencialmente innovadoras.

Diseño centrado en las personas: Ejercicio de las actividades de diseño que se enfoca en las necesidades, deseos, dificultades y frustraciones de las personas que utilizarán el producto o servicio diseñado (no en los aspectos técnicos o tecnológicos). Involucra la exploración del comportamiento de las personas destinatarias y promueve una iteración desde el inicio del diseño y a lo largo de este para la retroalimentación por parte de dichas personas. Prioriza a las personas a través de la identificación de oportunidades de mejora en sus experiencias, proponiendo soluciones que sean intuitivas, útiles, efectivas y fáciles de adoptar.

Explicabilidad: Documentos de la OCDE la definen como aquel aspecto que permite que las personas afectadas por el resultado de un sistema de IA entiendan cómo se llegó a él. Esto implica proporcionar información fácil de entender a las personas afectadas por el resultado de un sistema de IA que les permita cuestionar el resultado en particular, en la medida de lo posible, los factores y la lógica que condujeron a un resultado.

IA centrada en el ser humano (Inteligencia aumentada): Sistemas de IA concebidos para amplificar y aumentar las capacidades humanas y el control de los humanos sobre las máquinas, no para reemplazarlos. Son sistemas que priorizan los intereses y los derechos de las personas por sobre la automatización. También representa una manera de adoptar IA en donde una o más tecnologías no reemplazan a las personas involucradas sino que estas trabajan de modo colaborativo comúnmente conocido como modalidad humano-máquina.

Innovación: Acción y efecto producido al crear algo nuevo o alterar/modificar algo existente, dando lugar a otra cosa sustancialmente diferente que aporta valor en un determinado contexto dado que dicha novedad es adoptada por las personas que lo integran mejorando o transformando algún aspecto de su quehacer.

Innovación Pública: Procesos, productos o servicios, que entregan valor, y resultan ser nuevos o mejorados para responder a desafíos colectivos y mejoran la satisfacción ciudadana, incrementan la productividad de la administración estatal, la apertura democrática de sus instituciones, la producción de servicios y políticas públicas, entre otras.

Inteligencia Artificial: No existe una definición universalmente aceptada de IA. En noviembre de 2018, el Grupo de Expertos en IA de la OCDE (AIGO) estableció un subgrupo para desarrollar una descripción de un sistema de IA. Este grupo lo define como un sistema basado en máquinas que es capaz de influir en el entorno produciendo un resultado (predicciones, recomendaciones o decisiones) para un conjunto determinado de objetivos. Utiliza datos e insumos basados en máquinas y/o humanos para (i) percibir entornos reales y/o virtuales; (ii) abstraer estas percepciones en modelos a través del análisis de manera automatizada (por ejemplo, con aprendizaje automático), o manual-

mente; y (iii) usar la inferencia del modelo para formular opciones para los resultados. Los sistemas de IA están diseñados para operar con diferentes niveles de autonomía. Asimismo, Naciones Unidas define a la inteligencia artificial como la capacidad de una computadora o un sistema robótico habilitado por computadora para procesar información y producir resultados de manera similar al proceso de pensamiento de los seres humanos en el aprendizaje, la toma de decisiones y la resolución de problemas.

Inteligencia Artificial estrecha: Según la OCDE, la inteligencia artificial "estrecha", "débil" o "aplicada" está diseñada para realizar una tarea específica de razonamiento o de resolución de problemas dentro de un dominio limitado. Si bien estas tareas pueden estar impulsadas por algoritmos altamente complejos y redes neuronales, siguen siendo singulares y orientadas a objetivos puntuales. La IA estrecha no tiene la capacidad de adaptarse a situaciones nuevas sin una reprogramación previa.

Inteligencia Artificial general: También conocida como inteligencia artificial "fuerte" o inteligencia artificial "generalizada", es un área de investigación y desarrollo en constante evolución (OCDE). Se refiere a sistemas de inteligencia artificial que tendrían la capacidad de aprender, generalizar, inducir y abstraer el conocimiento a través de diferentes funciones cognitivas. Tendrían una fuerte memoria asociativa y serían capaces de juzgar y tomar decisiones. Podrían resolver problemas multifacéticos, aprender a través de la lectura o la experiencia, crear conceptos, percibir el mundo ya sí mismo, inventar y ser creativo, reaccionar ante lo inesperado en entornos complejos y anticiparse. Sólo existe como concepto teórico, su advenimiento es incierto.

Proyecto de adopción tecnológica: Manera en que se organiza y alcanza, a través de distintas acciones, la adopción de una o más tecnologías con el propósito de realizar algún tipo de innovación, como por ejemplo, innovación pública.

Red neuronal artificial: Documentos de la OCDE la definen como una técnica sofisticada de modelado estadístico. Esta técnica va acompañada de un poder computacional creciente y la disponibilidad de conjuntos de datos masivos ("big data"). Las redes neuronales involucran la interconexión repetida de miles o millones de transformaciones simples en una máquina estadística más grande que puede aprender relaciones sofisticadas entre entradas y salidas. En otras palabras, las redes neuronales modifican su propio código para encontrar y optimizar enlaces entre entradas y salidas. Finalmente, el aprendizaje profundo es una frase que se refiere a redes neuronales particularmente grandes; no hay un umbral definido en cuanto a cuándo una red neuronal se vuelve "profunda".

Sesgos: En documentos de la OCDE que describen términos claves de base común para discusiones del G20, definen cuatro tipos de sesgos que pueden ocurrir en los sistemas de inteligencia artificial.

> **Sesgo de percepción:** Se produce cuando los datos recopilados representan en exceso o en defecto a una determinada población y hacen que el sistema funcione mejor (o peor) para esa población en comparación con otras.
>
> **Sesgo técnico:** Ocurre cuando la propia tecnología introduce sesgos o imprecisiones debido, por ejemplo, a algoritmos que funcionan mejor con ciertas variables o características del sistema de IA que se introducen con diferentes variables o características.
>
> **Sesgo de modelado:** Se produce cuando el diseño manual de un modelo por parte de expertos no tiene en cuenta algunos aspectos del entorno, ya sea consciente o inconscientemente.
>
> **Sesgo de activación:** Se produce cuando las salidas del sistema de IA se utilizan en el entorno de forma sesgada.

Anexo

Concepción

Humano
(Decide)

IA
(Ejecuta)

Pre ciclo

Concientización

¿IA?

Modelo de adopción

Destino de uso y control humano

Premortem

Ciclo de vida de IA

Diseño Datos Modelo	Verificación Validación	Implementación (On premise / as a service)	Operación Mantenimiento
Equipo diverso y multidisciplinario	Verificación validación con todos los actores	Seguridad de la información	Monitoreo
Sensibilidad Privacidad Sesgos Consentimiento (Datos)	Sensibilidad Privacidad Sesgos Consentimiento (Datos)	Trazabilidad	Registro y tratamiento de incidentes Éticos
		Auditoría	
Sesgos Explicabilidad (Modelo)	Sesgos Explicabilidad (Modelo)	Accesibilidad	Control de usuarios internos

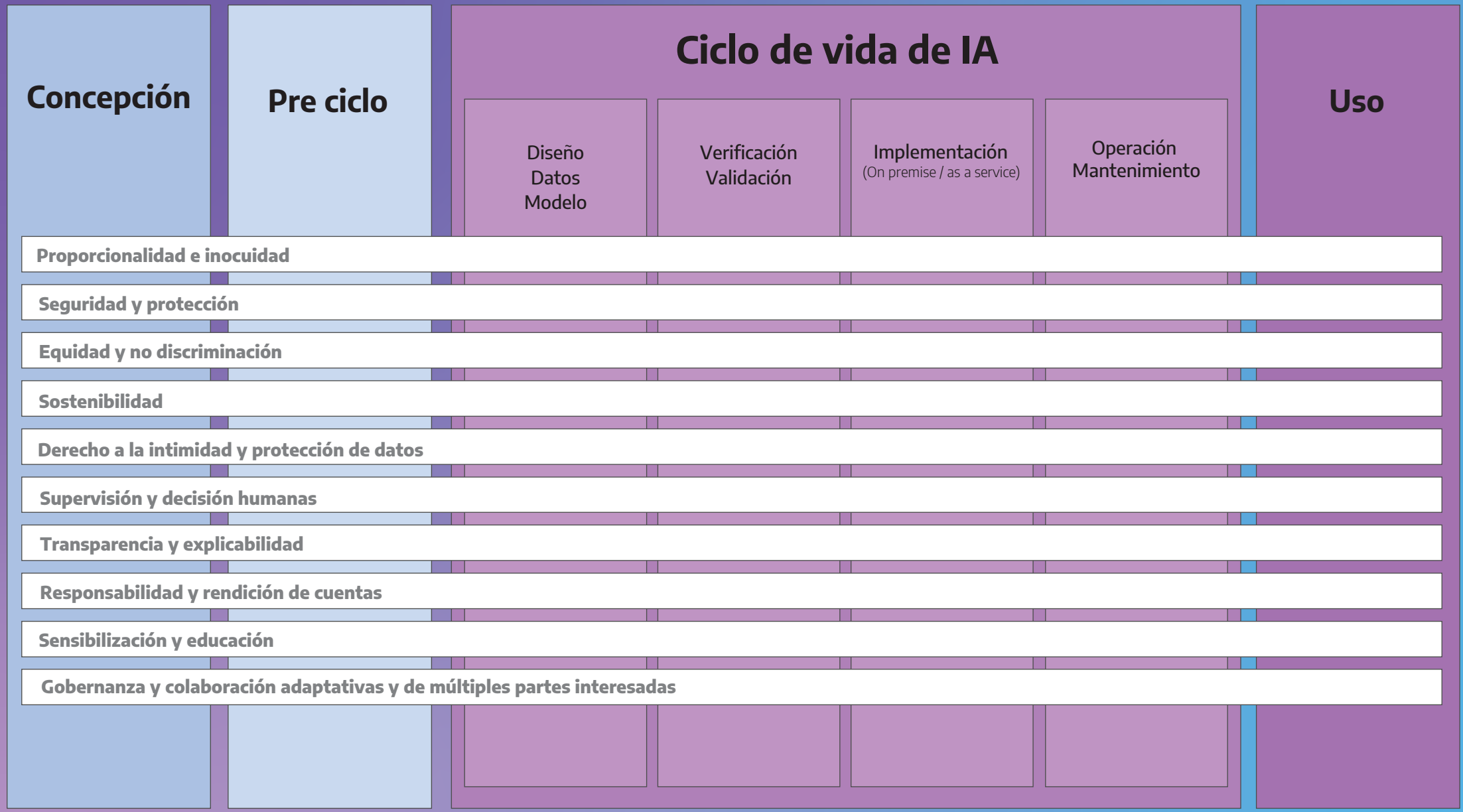
Mapa de adopción

Uso

Reversibilidad

Responsabilidad del daño

Vía alternativa no tecnológica IMPACTO


Concepción
Pre ciclo
Ciclo de vida de IA
Diseño
Datos
Modelo
Verificación
Validación
Implementación
(On premise / as a service)
Operación
Mantenimiento
Uso
Proporcionalidad e inocuidad
Seguridad y protección
Equidad y no discriminación
Sostenibilidad
Derecho a la intimidad y protección de datos
Supervisión y decisión humanas
Transparencia y explicabilidad
Responsabilidad y rendición de cuentas
Sensibilización y educación
Gobernanza y colaboración adaptativas y de múltiples partes interesadas