MEXICO | AGOSTO 2023

Open Loop México:

Prototipo de Políticas Públicas sobre Transparencia y Explicabilidad de Sistemas de IA

CLAUDIA MAY DEL POZO
NORBERTO DE ANDRADE
DANIELA ROJAS ARROYO

Acerca de Open Loop

Open Loop es un programa mundial que conecta a responsables políticos con empresas tecnológicas para ayudar a desarrollar políticas eficaces y basadas en pruebas en torno a la IA y otras tecnologías emergentes. El programa, apoyado por Meta (antes Facebook), se basa en la colaboración y las contribuciones de un consorcio de reguladores, gobiernos, empresas tecnológicas, académicos y representantes de la sociedad civil.
A través de métodos experimentales de gobernanza, los miembros de Open Loop co-crean prototipos de políticas (en inglés: policy prototypes) y ponen a prueba enfoques nuevos y diferentes de leyes y reglamentos antes de que se decreten, mejorando así la calidad de los procesos de elaboración de normas en el ámbito de la política tecnológica. Este informe presenta las conclusiones y recomendaciones del programa de creación de prototipos de políticas de Open Loop sobre la transparencia y explicabilidad de los sistemas de Inteligencia Artificial que se llevó a cabo en México entre febrero y agosto de 2021.

Citar este informe

Del Pozo, C., Nuno Gomes de Andrade, N., & Rojas Arroyo, D. "Prototipo de Políticas Públicas sobre Transparencia y Explicabilidad de Sistemas de Inteligencia Artificial" (2023), en: https://openloop.org/reports/2023/10/Prototipo-de-Politicas-Publicas-sobre-Transparencia-y-Explicabilidad-de-Sistemas-de-IA.pdf

Agradecimientos

Este programa prototipo de política fue co-diseñado y liderado por Open Loop, un programa global de gobernanza experimental apoyado por Meta, el Eon Resilience Lab de C Minds, y el Banco Interamericano de Desarrollo (BID), a través de su iniciativa fAIr LAC, con el apoyo del Instituto Nacional de Transparencia, Acceso a la Información y Protección de Datos Personales (INAI) de México.

Queremos agradecer a las siguientes empresas por su colaboración y compromiso, además de su activa participación, sin la cual este informe no hubiera sido posible:

Ai360	ai360 Analítica Inmobiliaria
Fincomún	Fincomún.
helKi	helKi
Hitch	hitch.
Inndot	inndot PIENSA SOLUCIONES
LUZi	LUZI
Nauphilus	nauphilus
Nowports	nowports
OS City	OS CITY
Rhisco	rhisco TO COMPLY & COMPETE

Un agradecimiento especial a las siguientes personas expertas por su invaluable tiempo, ideas y aportaciones para el desarrollo e implementación de este informe y del prototipo de política pública (quienes aparecen en orden alfabético):

- **Carla Vázquez Wallach**
  Fundadora y Directora General de Legal + Innovation en México y Miembro del Brain Hive de C Minds
- **César Said Rosales**
  Director de Proyectos de la iniciativa fAIr LAC del Laboratorio del Banco Interamericano de Desarrollo (BID Lab)
- **Constanza Gómez-Mont**
  Presidenta y Fundadora de C Minds
- **Cristian Guerrero**
  Ex consultor técnico de C Minds
- **Cristina Pombo**
  Consultora del Sector Social del BID
- **Daniel Castaño**
  Fundador y socio de Mokzy, profesor de la Universidad Externado de Colombia e investigador y consultor especializado en IA, ética digital y regulación
- **Edson Prestes**
  Investigador y profesor de ética de la IA en el Instituto de Informática de la Universidad Federal de Rio Grande do Sul (Brasil), miembro sénior de la Sociedad de Robótica y Automatización del Instituto de Ingenieros Eléctricos y Electrónicos (IEEE RAS) y la Asociación para las Reglas, miembro de la Asociación para la Regulación de la Inteligencia Artificial, y miembro del *Brain Hive* de C Minds
- **Guillermo Larrea**
  Abogado corporativo con enfoque en América Latina en Jones Day
- **Jesús Sánchez**
  Subdirector de Investigación del Instituto Nacional de Transparencia, Acceso a la Información Pública y Protección de Datos (INAI)
- **Jonathan Mendoza**
  Secretario de Protección de Datos Personales del INAI
- **Laura Galindo Romero**
  Responsable de Políticas de Inteligencia Artificial de Meta
- **Natalia González**
  Ex coordinadora de la iniciativa fAIr LAC del BID
- **Paula Vargas**
  Directora de Privacidad y Políticas Públicas de Meta
- **Rafael Ramírez de Alba**
  Profesor del Departamento de Entorno Económico de la Escuela de Negocios IPADE y miembro del *Brain Hive* de C Minds
- **Ricardo Baeza-Yates**
  Director de Investigación del Instituto de Inteligencia Artificial Experiencial de Northeastern University en el Silicon Valley y parte del Brain Hive de C Minds.
- **Tetsuro Narita**
  Especialista Sénior de la Unidad de Inversiones del BID Lab
- **Verena Kontschieder**
  Ex Directora del Programa de Política de Inteligencia Artificial de Meta
- **Victoria Martín del Campo**
  Ex filósofa de la tecnología en C Minds

Asimismo, agradecemos a nuestros socios de difusión Wizeline y MC Luhan, claves para la etapa de convocatoria del programa.

Índice

Índice

Instituto Nacional de Transparencia, Acceso a la Información y Protección de Datos Personales (INAI)

En esta era digital nos enfrentamos a un inminente avance tecnológico que no cesa. Formamos parte de una generación que utiliza la tecnología en sus actividades diarias. En este contexto, es de suma importancia tener presentes las ventajas y desventajas del mundo virtual, por lo que es indispensable detenernos un minuto a reflexionar sobre las garantías que debemos exigir, tanto a los desarrolladores como a las autoridades, y el papel del usuario en el uso de las herramientas que nos facilitan nuestra vida.

Desde el punto de vista de la privacidad, la confianza es indispensable y se vuelve clave para seguir el ritmo acelerado de la innovación, por lo que es de suma importancia considerar los aspectos éticos del diseño que demuestren un uso responsable en el tratamiento de datos personales.

Como menciona Yuval Noah Harari, historiador y escritor israelí, "La primera regulación que sugeriría es hacer obligatorio que la IA revele que es una IA. Si estoy teniendo una conversación con alguien y no puedo saber si es un humano o una IA, se acabó la democracia. Este texto ha sido generado por un humano".

La ética debe ir de la mano con la innovación para generar mecanismos de regulación y políticas públicas. Es crucial implementar un marco global para el uso ético de los datos a lo largo de todas las fases del ciclo de vida, desde su generación, utilización, hasta su eliminación. Este marco debe garantizar que el tratamiento de los datos se base en principios éticos, generando confianza al titular de los datos. Por lo tanto, debemos asegurar un enfoque ético es esencial para garantizar que los usuarios sigan estando en el centro de los procesos de toma de decisiones.

Como señaló el antiguo Supervisor Europeo de Protección de Datos, Giovanni Butarelli: "La innovación humana siempre ha sido el resultado de las actividades de grupos sociales específicos y contextos específicos, que generalmente reflejan las normas sociales de la época. Sin embargo, las decisiones de diseño tecnológico no deben dictar nuestras interacciones sociales y la estructura de nuestras comunidades, sino que deben respaldar nuestros valores y derechos fundamentales".

El papel del INAI en este proyecto consistió en orientar a las empresas participantes sobre las mejores prácticas para garantizar la protección de datos personales, para lo cual sugerimos acciones de privacidad desde el diseño y por defecto, así como el apego estricto a los principios para el tratamiento de datos personales.

El INAI celebra el haber participado en este tipo de iniciativas que nos permiten acercarnos e ir de la mano con desarrolladores y empresas que implementan nuevas tecnologías y que se preocupan, al mismo tiempo, por garantizar el respeto a los derechos humanos de sus usuarios. Open Loop representó una oportunidad única para el INAI, la cual nos permitió resaltar la importancia de la protección de datos personales y el respeto a la privacidad de los interesados. Como autoridades regulatorias debemos fomentar la colaboración público-privada que será un instrumento clave y un

canal de comunicación efectivo para enfrentar los retos futuros que presentan las tecnologías emergentes, permitiéndonos fortalecer e implementar diversos mecanismos de prevención centrados en el ser humano.

Jonathan Mendoza Iserte
Secretario de Protección de Datos Personales
Instituto Nacional de Transparencia, Acceso a la Información Pública y Protección de Datos (INAI)

C Minds

Los sistemas de inteligencia artificial (IA) desempeñan un papel cada vez más destacado en nuestra vida cotidiana, por lo que es de suma importancia que les demos prioridad a un uso responsable y centrado en los derechos. En este sentido, el equipo de C Minds está comprometido con la creación de estrategias que minimicen los riesgos potenciales y maximicen los impactos sociales positivos de los sistemas de IA y otras tecnologías emergentes. Creemos que un enfoque colaborativo es esencial para lograr este objetivo, ya que al tender puentes entre diferentes sectores y perspectivas, esto puede conducir a una mayor comprensión de los matices asociados a la aplicación práctica de los principios. Esto, a su vez, conduce a estrategias más holísticas, inclusivas y pragmáticas para el desarrollo y uso de sistemas de IA.

En 2019, C Minds fue coautor de la Estrategia Nacional de IA de México, que se centró en el uso responsable y ético de la tecnología, posicionando a México entre los 10 países que han creado una estrategia relativa a la IA. Desde entonces, hemos continuado con nuestra misión de mejorar la calidad de vida en México y otros países de América Latina y el Caribe a través del uso responsable de las nuevas tecnologías. Lo hemos logrado desarrollando proyectos pioneros de ética de IA en la región, generando recomendaciones de políticas públicas y creando lineamientos y marcos enfocados en el desarrollo y uso responsable.

El proyecto presentado en este informe es emocionante, sobre todo teniendo en cuenta su carácter único en la región y su modelo de gobernanza multisectorial. La creación de prototipos de políticas públicas es una metodología dinámica y pionera que ha surgido como un enfoque prometedor para mitigar algunos de los retos presentes en el desarrollo actual de políticas públicas, especialmente los relacionados con la tecnología. También subraya la importancia de aprender de la retroalimentación y de involucrar a las partes interesadas desde una fase temprana. Entre otras cosas, esto permite desarrollar políticas públicas contextualizadas y soluciones pragmáticas a los retos -en nuestro caso, los del campo de la ética de la IA. Este proyecto se basa en una filosofía de inclusión que reúne a todos los sectores para garantizar perspectivas complementarias; esto es el resultado de una orgullosa colaboración entre Meta y el Eon Resilience Lab de C Minds, junto con el Banco Interamericano de Desarrollo (BID) bajo su iniciativa de fAIr LAC, el Instituto Nacional de Transparencia, Acceso a la Información Pública y Protección de Datos (INAI) de México, empresas participantes y personas expertas en la materia.

Con este documento, esperamos contribuir a la creación de marcos regulatorios para el desarrollo y uso de sistemas de IA centrados en el ser humano en México. Esperamos que las recomendaciones presentadas inspiren a las instituciones reguladoras no sólo de México, sino de toda la región de América Latina y el Caribe, a continuar promoviendo el uso y desarrollo responsable de las tecnologías emergentes, a crear procesos más incluyentes, y beneficios tecnológicos más equitativos para todos y todas.

Constanza Gómez Mont
Fundadora y Presidenta
C Minds

Meta

La transparencia y la explicabilidad son fundamentales para el desarrollo responsable de la Inteligencia Artificial (IA). La OCDE identifica la transparencia y la explicabilidad como uno de sus principios fundamentales, afirmando en su Recomendación sobre IA que "los actores de la IA deben comprometerse con la transparencia y la divulgación responsable de los sistemas de IA.... [proporcionando] información significativa [que sea] adecuada al contexto y coherente con el estado de la técnica". Sin duda, para Meta estos son aspectos centrales dentro de nuestros cinco pilares de IA responsable.

Nuestro equipo interdisciplinario de IA Responsable (RAI, por sus siglas en inglés) ha colaborado estrechamente con el mundo académico, la sociedad civil, gobiernos y otros socios de la industria en varios proyectos de transparencia y explicabilidad. Por ejemplo, hemos adoptado el uso de tarjetas de sistemas de IA para explicar el funcionamiento de la IA de nuestros productos de forma comprensible para los usuarios. En nuestro primer proyecto piloto explicamos el proceso de categorización del *feed* de Instagram, ofreciendo una explicación simplificada, paso a paso, de las operaciones del sistema. Además, ofrecemos un ejercicio interactivo en el que los usuarios pueden experimentar con perfiles hipotéticos para predecir el aspecto de sus *feeds*. La intención es proporcionar a los usuarios la información técnica necesaria para entender nuestros productos y tomar decisiones informadas sobre sus experiencias.

Otro ejemplo es la herramienta "¿Por qué veo esto?" (WAIST, por sus siglas en inglés), que se actualizó a principios de 2023. Al personalizar las experiencias de los usuarios con nuestros productos, utilizamos la IA para presentarles los contenidos y anuncios que más se ajustan a sus intereses. La actualizada herramienta WAIST permite a los usuarios comprender mejor este proceso de personalización, tanto para *News Feed*, como para los anuncios. Las actualizaciones incluyen información resumida sobre cómo la actividad del usuario, tanto dentro como fuera de nuestras tecnologías, influye en los modelos de aprendizaje automático que utilizamos para dar forma y ofrecer anuncios. También hemos incluido nuevos ejemplos e ilustraciones que explican cómo nuestros modelos de aprendizaje automático conectan diversos temas para presentar anuncios relevantes a los usuarios.

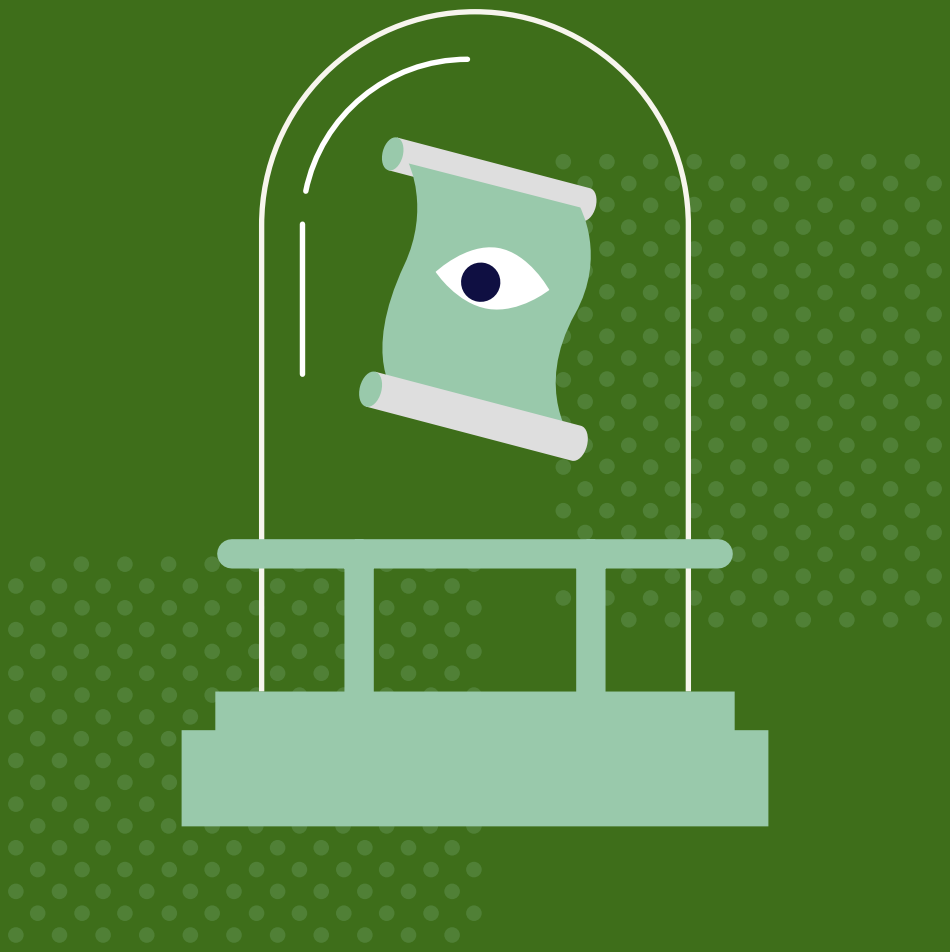
Además de nuestros esfuerzos internos, la colaboración externa desempeña un papel vital en nuestras iniciativas de uso responsable de IA.

Plantear la elaboración de políticas de forma experimental y basada en pruebas permite a los políticos y reguladores evaluar sistemáticamente el impacto de sus propuestas en personas y empresas. Este enfoque les ayuda a comprender mejor la repercusión de sus propuestas en el mundo real antes de que se conviertan en leyes y regulaciones concretas. Nuestro objetivo con Open Loop es compartir las enseñanzas con los responsables políticos y las partes interesadas de todo el mundo, animándoles a adoptar iniciativas de prototipo similares y a adoptar un enfoque innovador y colaborativo en el desarrollo de políticas públicas.

El proyecto de Open Loop descrito en este informe gira en torno a la creación de un marco normativo y un manual práctico que describan los principios de transparencia y explicabilidad. A modo de prototipo de política pública, estos principios fueron posteriormente puestos a prueba por empresas mexicanas que utilizan sistemas automatizados de apoyo en la toma de decisiones para ofrecer bienes y/o servicios. El objetivo principal de esta iniciativa era explorar cómo estas empresas podrían incorporar eficazmente la transparencia y la explicabilidad, permitiendo a los usuarios acceder a información esencial sobre sus interacciones con los sistemas de IA. Al fomentar una comprensión más profunda de las capacidades, limitaciones y procesos que conducen a resultados específicos, los usuarios estarían mejor equipados para navegar por sus experiencias. El diseño del prototipo de política pública sobre transparencia y explicabilidad de los sistemas automatizados de apoyo en la toma de decisiones se guió por las mejores prácticas internacionales, teniendo en cuenta las consideraciones contextuales y de alcance. Open Loop México enfocó sus esfuerzos principalmente en empresas nacientes y de madurez temprana, buscando entender las circunstancias que enfrentan las organizaciones con recursos técnicos, económicos y humanos limitados. Este énfasis es crucial, ya que dichas entidades constituyen una porción significativa del panorama empresarial en economías en desarrollo como México (representando el 98% de la composición).

Quisiera mencionar que esto no habría sido posible sin el espíritu de colaboración del Instituto Nacional de Transparencia, Acceso a la Información y Protección de Datos Personales (INAI) de México, la iniciativa fAIr LAC del Banco Interamericano de Desarrollo (BID) y las incansables y profesionales colegas del Eon Resilience Lab de C Minds. Agradecemos también, por supuesto, la participación de las empresas mexicanas que se sumaron al proyecto y el apoyo de un excelente grupo de personas expertas.

Paula Vargas
Directora de Políticas de Privacidad y Compromiso en LATAM
Meta



Resumen ejecutivo

Resumen ejecutivo

¿Qué es Open Loop?

Open Loop es un programa mundial que pone en contacto a desarrolladores de políticas y empresas tecnológicas para ayudar a elaborar políticas eficaces y basadas en pruebas en torno a la IA y otras tecnologías emergentes. El programa, que cuenta con el apoyo de Meta, se basa en la colaboración y las contribuciones de un consorcio compuesto por reguladores, gobiernos, empresas tecnológicas, académicos y representantes de la sociedad civil. A través de métodos experimentales de gobernanza, los miembros de Open Loop co-crean prototipos de políticas y prueban enfoques nuevos y diferentes de las leyes y normativas antes de que se promulguen, mejorando así la calidad de los procesos normativos en el ámbito de la política tecnológica.

¿Qué es Open Loop México?

En el caso de México, el **"Prototipo de política pública sobre transparencia y explicabilidad de los sistemas de Inteligencia Artificial"** (en adelante se hará referencia a los sistemas de IA como sistemas de A/TDA; ya que también nos referiremos a los sistemas de Toma de Decisiones Automatizadas (TDA), manteniendo la neutralidad tecnológica ante posibles desarrollos tecnológicos futuros)[1] fue realizado por Meta y el Eon Resilience Lab de C Minds, en colaboración con el Banco Interamericano de Desarrollo (BID), a través de su iniciativa fAIr LAC y con el apoyo del Instituto Nacional de Transparencia, Acceso a la Información y Protección de Datos Personales Protección de Datos Personales (INAI) de México, así como con la industria y personas expertas en el tema. El objetivo de este programa fue diseñar un marco de gobernanza y un manual práctico (*playbook*) que esboza los principios de transparencia y explicabilidad (TyE). Estos documentos (prototipo de política) fueron probados por empresas mexicanas que utilizan sistemas de A/TDA para suministrar bienes o servicios. El objetivo general de la política era fortalecer el uso de la IA responsable en México, centrándose en la TyE.

Este ejercicio tenía como objetivo garantizar que las personas sepan cuándo están interactuando con un sistema de A/TDA y comprender sus limitaciones y capacidades, así como la forma en que logra resultados específicos.

¿Por qué enfocarse en transparencia y explicabilidad?

Cualquier ser humano que interactúe con un sistema de A/TDA debe poder saber cómo y por qué se producen ciertos resultados, conclusiones o predicciones y así comprender el razonamiento lógico detrás de una decisión o recomendación dadas por este tipo de sistemas.

El principio de transparencia se refiere a la capacidad de las personas para comprender y describir el funcionamiento interno de un sistema. A su vez, la explicabilidad salvaguarda el derecho a conocer la mecánica interna de un sistema de A/TDA y comprenderlo en términos humanos.

El prototipo de política pública sobre TyE de los sistemas de A/TDA (de ahora en adelante, "prototipo") se diseñó basado en el contexto y el alcance de las mejores prácticas internacionales en torno a los principios de TyE. La idea era ponerlo a prueba en relación con las siguientes preguntas:

- **Claridad:** ¿hasta qué punto entienden las empresas participantes los requisitos establecidos en el prototipo?
- **Eficacia:** ¿hasta qué punto contribuye el prototipo a alcanzar el objetivo político general?
- **Viabilidad:** ¿hasta qué punto los beneficios compensan los costos de alcanzar los objetivos del prototipo de política pública?

¿Cómo se llevó a cabo Open Loop?

El programa se llevó a cabo con 10 empresas mexicanas que realizaron una serie de actividades para implementar el conjunto de normas y prácticas contenidas en la propuesta de marco normativo y el manual creados por los socios del proyecto con el apoyo de un grupo de personas expertas en la materia. Las empresas entonces proporcionaron información sobre la claridad, viabilidad y eficacia de estos documentos.

Mediante una serie de actividades, las empresas se ajustaron y probaron el alcance del prototipo desde un punto de vista de cumplimiento. Los comentarios y sugerencias aportados por las empresas permitieron ajustar el contenido del prototipo para mejorarlo, en términos de claridad, viabilidad y eficacia, además de promover la comprensión de lo que significa observar los principios de TyE.

Metodología

El programa Open Loop México se llevó a cabo entre febrero y agosto de 2021 y se estructuró en tres fases:

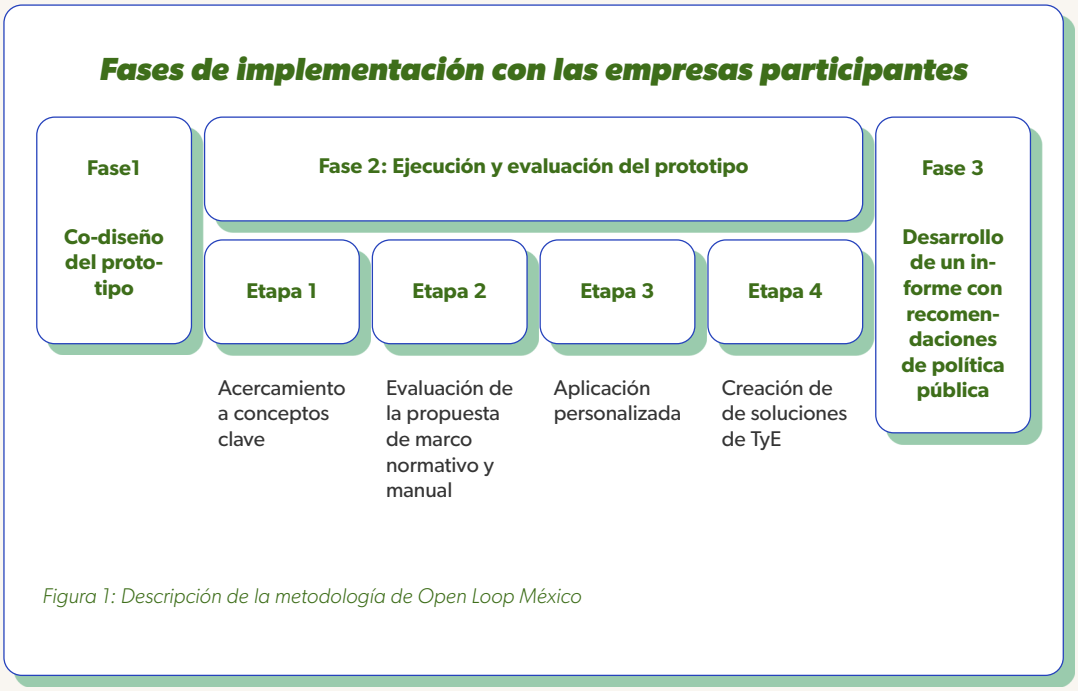


Figura 1: Descripción de la metodología de Open Loop México

Resultados

Open Loop México demostró que la metodología propuesta para evaluar que la idoneidad del prototipo era la adecuada para probar el marco normativo de gobernanza y el manual con las empresas participantes. El programa permitió recibir retroalimentación sobre algunos de los éxitos y retos vividos en la aplicación de una posible normativa sobre pruebas y ensayos de los sistemas de A/TDA en el marco de la evaluación de los tres sistemas establecidos. Se obtuvieron los siguientes resultados:

- **Claridad:** El prototipo era claro para las empresas participantes. En términos de TyE, la comprensión del tema por parte de las empresas pasó de 5,3 sobre 10 antes de leer los documentos a 8 sobre 10 después de leerlos, siendo 0 "ningún conocimiento" y 10 "expertise". Los objetivos y acciones concretas necesarias para cumplir los requisitos establecidos en el marco también quedaron claros, sobre todo con el apoyo del manual. Este último impactó de forma significativa en la efectividad de la implementación, permitiendo así a las las empresas fundamenten sus argumentos en el ámbito de las normas prescriptivas de la propuesta. En general, se les dio a los documentos una puntuación de 7,4 sobre 10 en cuanto a claridad, siendo 10 “extremadamente claro”.

- **Eficacia:** La mayoría de las empresas diseñaron y/o publicaron mensajes de explicabilidad como parte de la experiencia del usuario con los productos o servicios. Algunas optaron por incluir notificaciones, mensajes o videos para que el usuario entendiera cómo funcionaba el sistema de A/TDA. Aunque el prototipo demostró que engloba los elementos esenciales para que las empresas desarrollen soluciones de TyE para sus productos o servicios, aún hay oportunidad de mejorar el entendimiento de los potenciales riesgos relacionados con los sistemas de IA. Esto resalta la necesidad de realizar campañas exhaustivas de concienciación sobre las repercusiones de los sistemas de A/TDA antes de considerar la implementación de un marco de gobernanza obligatorio. Sin una concienciación adecuada, existe el riesgo de incumplimiento significativo de la normativa.

- **Viabilidad:** En general, las empresas manifestaron que experimentaban cierto grado de dificultad para cumplir con el marco normativo establecido en el prototipo debido a la falta de tiempo y de personal suficientemente capacitado en cuestiones técnicas. También señalaron que la viabilidad de implementar mecanismos de TyE variaría en función de la complejidad y el impacto (nivel de riesgo) del modelo utilizado por las empresas.

El programa Open Loop México dio origen a un marco normativo que puede servir de insumo para que las instituciones reguladoras desarrollen políticas públicas en materia de TyE de sistemas de A/TDA, con la ventaja de haber sido probadas y mejoradas con recomendaciones de las empresas que lo implementaron. El contenido del marco y del manual se fortaleció de la siguiente manera:

- Se integró una redacción más clara; se establecieron definiciones individuales para favorecer la transparencia, la explicabilidad y la interpretabilidad global y local. Se agregaron casos hipotéticos para mejorar la comprensibilidad del prototipo.

- Se incluyó una nueva etapa que consideraba la importancia y la aplicación de los procesos de medición de impacto para determinar los riesgos asociados a los sistemas de A/TDA.

Aunque la información recabada y las recomendaciones obtenidas no pueden generalizarse a todo tipo de empresas (sería importante llevar a cabo un ejercicio complementario similar con empresas grandes y multinacionales), Open Loop México se enfocó principalmente en empresas nacientes y en etapas tempranas para conocer la situación de las organizaciones con menos recursos

técnicos, económicos y humanos, que representan la gran parte de la composición empresarial de economías en vías de desarrollo (las micro, pequeñas y medianas empresas (MIPYMES) representan el 99.8% de las empresas en México) (INEGI, 2019).

Recomendaciones

Con base en los resultados del programa Open Loop México, incluyendo la información recibida de las empresas participantes y del grupo de personas expertas, las siguientes recomendaciones de política pública sobre TyE para sistemas de A/TDA deben ser consideradas por los y las formuladoras de políticas:

1. **Promover activamente la Inteligencia Artificial como una prioridad nacional, con un enfoque en la operacionalización de los principios confiables de la IA.**
2. **Desempeñar un papel proactivo en la gobernanza de la IA en México.**
3. **Desarrollar capacidades para tener una IA confiable, especialmente en materia de transparencia y explicabilidad (TyE) en organismos gubernamentales no técnicos.**
4. **Aumentar la capacidad técnica de la IA confiable en México.**
5. **Invertir en la investigación y el desarrollo de la IA confiable.**
6. **Fortalecer la capacidad de desarrollo y adopción de la IA responsable en la mano de obra mexicana.**
7. **Incrementar la conciencia cívica sobre la IA en México.**

1 Promover activamente la Inteligencia Artificial como una prioridad nacional, con un enfoque en la operacionalización de los principios confiables de la IA.

- Las y los formuladores de políticas públicas podrían utilizar los recursos, las mejores prácticas internacionales y herramientas para crear una estrategia nacional de IA[2].
- Esta estrategia podría esbozar los objetivos políticos de la IA alineados a los Principios de la IA de la OCDE y la UNESCO, así como con las políticas que podrían ser necesarias para alcanzar dichos objetivos. La estrategia también podría incluir medidas específicas para promover la transparencia y la explicabilidad de los sistemas de IA[3].
- Este ejercicio debería ser un esfuerzo multilateral liderado por organismos gubernamentales nacionales. Asimismo, en la creación de la estrategia también podrían participar el sector privado, la academia y la sociedad civil a través de ejercicios innovadores.

2 Desempeñar un papel proactivo en la gobernanza de la IA en México

- Las y los formuladores de políticas públicas en México podrían desempeñar un papel proactivo en la gobernanza del desarrollo y el uso de la IA en el país: i) organizando y promoviendo ejercicios gubernamentales experimentales para identificar y abordar las oportunidades y los retos de la IA, como prototipos de políticas públicas y *sandboxes* regulatorios (antes de que se establezcan políticas/regulaciones), así como hackatones y concursos para comprender mejor las oportunidades y los retos en este ámbito; ii) desarrollar un marco normativo claro y conciso para la IA, basado en las necesidades locales y buenas prácticas internacionales y; iii) promover la colaboración intersectorial para garantizar que el marco normativo de IA sea comprensivo y refleje los puntos de vista de todas las partes interesadas.

3 Desarrollar capacidades para tener una IA confiable, especialmente en materia de transparencia y explicabilidad (TyE) en organismos gubernamentales no técnicos.

- Organizar e implementar sesiones y talleres de fortalecimiento de capacidades sobre las oportunidades y riesgos de la IA, con enfoque en la TyE. Las y los formuladores de políticas públicas podrían colaborar con organizaciones de la sociedad civil y la academia para organizar y llevar a cabo estas sesiones y talleres, así como impartir cursos masivos en línea (MOOCs, por sus siglas en inglés) para nivelar los conocimientos de los y las formuladores de políticas públicas, así como las y los funcionarios públicos, sobre los riesgos y oportunidades de la IA, especialmente en relación con la TyE. El fortalecimiento de sus capacidades y conocimientos les permitiría participar mejor en las conversaciones sobre el tema.

- Crear espacios regulares de diálogo con funcionarios y funcionarias gubernamentales, desarrolladores de IA y otras partes interesadas para debatir sobre temas relacionados con la TyE. Por ejemplo, las y los formuladores de políticas públicas podrían formar un grupo de trabajo integrado por funcionarios y funcionarias públicos, desarrolladores de IA y otras partes interesadas para debatir cuestiones relacionadas con TyE. Estas conversaciones podrían ayudar a crear consenso sobre las mejores prácticas de diseño, desarrollo, despliegue y uso de la IA, así como para identificar las áreas en las que se necesita más orientación.

4 Aumentar la capacidad técnica de la IA confiable en México.

- Los y las formuladores de políticas públicas podrían considerar el desarrollo de un conjunto de normas/protocolos técnicos para sistemas de IA en consulta con desarrolladores de IA, empresas y otras partes interesadas en el ecosistema mexicano de IA, basándose en buenas prácticas internacionales para garantizar que incluyan prácticas "*human-in-the-loop*" (es cuando los sistemas de IA y aprendizaje automático se construyen con interferencia humana en diferentes etapas del ciclo) cuando sea relevante y estén alineados con un enfoque de la IA centrado en el ser humano.

- Explorar el desarrollo de un marco normativo de gestión de riesgos, basado en el contexto mexicano, el cual sea altamente coherente e interoperable con las mejores prácticas internacionales y los esfuerzos de estandarización[4] del diseño, desarrollo y despliegue de sistemas de IA confiables, para reducir el potencial de impactos negativos inesperados. Estos son especialmente relevantes si las empresas optan por cumplir con los principios de TyE. Esto podría ser liderado por las instituciones reguladoras en colaboración con el ecosistema mexicano de IA.

- Además de crear recursos locales, también se debe de considerar la posibilidad de reunir los recursos internacionales existentes de países, empresas y organizaciones multilaterales en una página web gubernamental que se actualice periódicamente.

5 Invertir en la investigación y el desarrollo de la IA confiable.

- Las y los formuladores de políticas públicas podrían establecer incentivos financieros y no financieros para promover proyectos de investigación sobre TyE de la IA a través de organismos gubernamentales, universidades públicas y privadas, en colaboración con la industria y la sociedad civil para garantizar un enfoque práctico. También podrían considerarse las colaboraciones transfronterizas. En particular estos actores podrían invertir en investigación sobre:
  - i) técnicas para hacer que los sistemas de IA sean más transparentes y explicables, lo que podría incluir la investigación sobre métodos para visualizar el proceso de toma de decisiones de los sistemas de IA, así como la investigación sobre métodos para explicar los fundamentos de las decisiones de la IA; e
  - ii) investigación y herramientas para identificar y mitigar sesgos en los sistemas de IA, así como marcos normativos generales de gestión de riesgos.
- Además de fomentar y financiar la investigación, el gobierno, la academia y demás partes interesadas en la IA podrían crear espacios para compartir las principales enseñanzas, recomendaciones y herramientas resultantes de las actividades de investigación.

6 Fortalecer la capacidad de desarrollo y adopción de la IA responsable en la mano de obra mexicana.

- Promover la inclusión de cursos y módulos sobre consideraciones éticas en el desarrollo y adopción de sistemas de IA en carreras técnicas vinculadas a la ciencia de datos, informática e inteligencia artificial, entre otras. Esto podría aplicarse en espacios de educación formal como universidades y otras instituciones de aprendizaje o cursos, incluidos los de aprendizaje permanente.
- Las carreras enfocadas en ciencias sociales y humanidades, en espacios educativos formales e informales, también podrían ofrecer cursos y módulos de introducción a los sistemas de IA para crear una fuerza laboral más diversa que pueda centrarse en la IA responsable desde diferentes perspectivas. Esto podría ser promovido por Organismos de Certificación, así como las y los formuladores de políticas públicas a través de organismos como el Instituto Nacional de Transparencia, Acceso a la Información y Protección de Datos Personales (INAI), en colaboración con la industria, la sociedad civil y la academia, creando programas de capacitación sobre la importancia y cómo construir sistemas de IA transparentes y explicables, especialmente para desarrolladores.

7

Incrementar la conciencia cívica sobre la IA en México.

- Las y los formuladores de políticas públicas podrían lanzar una campaña de concientización pública sobre los riesgos y oportunidades de los sistemas de IA, resaltando la importancia de la TyE en los servicios y productos de IA. Esta campaña podría ayudar a impulsar las prácticas de TyE, como una ventaja competitiva para las empresas y como una solicitud de las personas consumidoras a sus proveedores de productos y servicios.

- La y los formuladores de políticas públicas y los organismos locales de juventud y educación, podrían seguir promoviendo programas de alfabetización digital en escuelas y universidades como cursos de aprendizaje permanente enfocados en la IA (una vez que hayan comprendido los fundamentos digitales) en colaboración con la sociedad civil y la academia.

- Apoyar el desarrollo y despliegue de recursos de alfabetización digital y de IA en español, y trabajar junto con el gobierno y actores alrededor de la IA para promover la alfabetización de la AI.

1

Introducción

La IA tiene el potencial de transformar considerablemente a la sociedad, mejorar el bienestar social, individual y el bien común, y propiciar tanto el progreso como la innovación. Estos sistemas ofrecen una diversidad de oportunidades para las empresas de América Latina. Según el informe de Everis y MIT Tech Review sobre el uso de la IA en México[5], el 47% de las empresas en México tienen un proyecto de IA y el 38% ven beneficios en su uso pero aún no trabajan con ella. Esto refleja un gran interés en el uso de estos sistemas, lo que probablemente se traducirá en un creciente desarrollo y adopción en los próximos años.

Esta adopción tecnológica, especialmente en lo que respecta a los sistemas de IA, ha generado nuevos desafíos para la protección de los derechos y libertades de las personas. Cada vez más este tipo de tecnologías son integradas en nuestro día a día, ya que los sistemas de IA están tomando -o apoyando- decisiones que tienen un impacto en nuestras vidas. En consecuencia, es necesario garantizar el desarrollo y el uso responsables de modelos de IA que protejan los derechos y libertades de los usuarios. Para lograr esto:

> ***"Los sistemas de IA deben estar centrados en el ser humano y utilizarse en beneficio de la humanidad y el bien común; esto con el propósito de mejorar el bienestar y la libertad de las personas"[6].***

En concordancia, el tema de la gobernanza de la IA pretende fomentar un debate informado sobre las implicaciones éticas, normativas y políticas que se derivan del desarrollo y el uso de la IA. Esto, a partir del diagnóstico de los retos y oportunidades de esta tecnología, así como los caminos a seguir con este desarrollo a futuro. El uso responsable de la IA y de los datos que entrenan estos sistemas ha estado en el centro de debates alrededor del mundo, lo cual ha dado como resultado el desarrollo de diversas propuestas y guías basadas en principios éticos. Estas propuestas han sido creadas por instituciones como la Organización para la Cooperación y el Desarrollo Económicos (OCDE)[7], el Parlamento Europeo[8], el Instituto de Ingenieros Eléctricos y Electrónicos (IEEE)[9], la Organización de las Naciones Unidas para la Educación, la Ciencia y la Cultura (UNESCO)[10] y la Comisión Europea[11], entre otros. Los principios de la OCDE con respecto a la IA, adoptados en mayo del 2019, ofrecen, en particular, cinco principios basados en valores[12]: 1) crecimiento inclusivo, desarrollo sostenible y bienestar; 2) valores centrados en el ser humano y la equidad; 3) transparencia y explicabilidad (TyE); 4) robustez, seguridad y protección; y 5) responsabilidad y rendición de cuentas.

Con base en estas diferentes propuestas, se han desarrollado varias propuestas legales y herramientas para operacionalizar los principios. Estos ejemplos demuestran que tanto las directrices éticas como las regulaciones de IA son un pilar fundamental. Sin duda son un elemento clave para generar confianza en las personas usuarias, garantizar el derecho de la persona a comprender una decisión que le concierne y promover la rendición de cuentas para todas las partes interesadas en el desarrollo de sistemas de IA. Este principio se centra en promover que las personas usuarias sean conscientes de su interacción con los sistemas de IA, las capacidades y limitaciones del sistema, y cómo se logran los resultados.

De cara al futuro, será necesario desarrollar mecanismos de aprendizaje colaborativos y dinámicos entre las y los reguladores, las empresas, la academia, la sociedad civil y el ecosistema de la innovación para facilitar la creación de marcos de gobernanza de la IA centrados en TyE que sean pragmáticos, inclusivos y operativos.

Para llevarlo a cabo, y debido a la compleja naturaleza de la tarea, los prototipos de las **políticas públicas resultan especialmente**

interesantes, pues ofrecen un terreno de pruebas seguro para evaluar la idoneidad y las posibles repercusiones de la política pública antes de su aplicación. Con esto en mente, se desarrolló un prototipo de política pública centrada en las TyE a través del programa Open Loop de Meta y el Eon Resilience Lab de C Minds, en colaboración con el Banco Interamericano de Desarrollo (BID) a través de su iniciativa fAIr LAC, sin olvidar, por supuesto, del apoyo del Instituto Nacional de Transparencia, Acceso a la Información y Protección de Datos Personales (INAI). Este fue el primer prototipo de política pública en América Latina y el Caribe.

El programa Open Loop en México fue diseñado para traducir los hallazgos en ideas prácticas, dialogar sobre el uso responsable de las tecnologías (específicamente TyE en sistemas de IA/Toma de Decisiones Automatizada (A/TDA)), y proporcionar recomendaciones a las instituciones reguladoras en México, que puedan a su vez inspirar a otras en otros países de la región. El siguiente informe recoge las y recomendaciones de este ejercicio.

2

Open Loop México y los prototipos de políticas públicas

La política pública se define tradicionalmente como una acción concreta por parte del gobierno con objetivos del interés público que surgen a raíz de decisiones basadas en un proceso diagnóstico de un problema. Una política pública puede convertirse en una ley o en un reglamento que rige un problema en concreto[13]. Sin embargo, dado que los procesos tradicionales de diseño de políticas públicas, en los que suele participar únicamente el gobierno, tienden a estar rezagados en cuestión de innovación tecnológica, la innovación regulatoria ha dado lugar a prototipos de políticas públicas que representan un esfuerzo multipartido por crear políticas adaptables, centradas en el ser humano, inclusivas y sostenibles[14].

¿Qué es Open Loop?

Es un programa global de gobernanza experimental apoyado por Meta. Es donde la innovación reguladora y tecnológica se encuentran a través del desarrollo de políticas públicas basadas en la evidencia en torno a las tecnologías emergentes, con especial énfasis en la IA. Su principal objetivo es generar la información necesaria para crear marcos normativos con una mejor comprensión de la interacción entre la tecnología y las políticas públicas, todo esto basado en la cooperación entre reguladores, gobiernos, empresas tecnológicas, académicos y sociedad civil, y aplicados con metodología experimental de gobernanza para co-crear prototipos de políticas públicas relacionadas con la tecnología para mejorar su desarrollo y aplicación. El programa Open Loop se ha desplegado por todo el mundo en varias ocasiones, cada vez, con un subtema diferente relacionado con el diseño, el desarrollo y el uso responsables de la IA. En noviembre de 2022 se habían llevado a cabo o estaban en curso siete programas (www.openloop.org).

¿Qué es un prototipo de política pública?

Tradicionalmente, el concepto de **prototipo** se vincula estrechamente con la industria, como un proceso experimental en el que se evalúa y aprende de una muestra, un modelo o una versión preliminar de algo (como un producto) antes de que salga al mercado. En *design thinking*[15] un prototipo es la expresión visible, tangible o funcional de una idea que se prueba con interesados externos en una fase temprana del desarrollo para así aprender de ella y reiterar la idea original.

Un **prototipo de políticas** puede definirse como una metodología para probar la eficacia de una política aplicándola primero en un entorno controlado. La creación de prototipos políticos adopta un enfoque centrado en la persona usuaria para desarrollar leyes y políticas[16] que permitan a las y los investigadores probar la claridad, viabilidad y eficacia de potenciales normas o políticas con una serie de prácticas y actividades antes de su aplicación.

¿Por qué diseñar prototipos de políticas públicas?

La idea de desarrollar prototipos surge de la necesidad de crear políticas más eficaces basadas en evidencia, evitando así los costes sociales y económicos de políticas no idóneas. Los prototipos de políticas públicas permiten ver y experimentar una política[17] e invita a que los actores principales involucrados a que participen activamente en el proceso de diseño de una política concreta[18]. De esta forma se pueden conocer los posibles efectos, fortalezas, debilidades y limitaciones de marcos normativos, propuestas de ley, códigos de conducta, entre otros, antes de aplicarlos de forma definitiva y oficial.

De forma general, esta metodología ofrece a las personas tomadoras de decisiones, la posibilidad de aprender, conocer y redireccionar las intervenciones políticas en una fase temprana del proceso al crear un espacio de experimentación de prueba y error para identificar las problemáticas en la aplicación de la política, lo que se traduce en un ahorro de recursos[19]. Puede ser especialmente útil en áreas en las que el ritmo de desarrollo tecnológico y de innovación es bastante acelerado, donde el impacto tanto de su evaluación, como de su regulación, son inciertos y difíciles de prever.

3

Open Loop México

Las y los socios implicadas en este proyecto propusieron y exploraron la idoneidad (claridad, eficiencia y viabilidad[20]) de un marco normativo destinado a reforzar la TyE de los sistemas de A/Decisión Automatizada (A/TDA) de empresas en México. El alcance de aplicación se amplió de la IA a todos los sistemas de TDA para garantizar la neutralidad tecnológica, teniendo en cuenta el posible desarrollo futuro de nuevas tecnologías. Las y los socios del programa trabajaron con 10 empresas mexicanas que utilizaban IA para sus productos o servicios. Este ejercicio requirió que estas empresas implementaran el marco normativo, así como un manual desarrollado para guiar la adopción (de ahora en adelante, el "prototipo"), ajustando y probando las prácticas en sus propios sistemas y operaciones para cumplir con los requisitos del programa. Para ello, las empresas siguieron un plan de trabajo detallado con misiones y actividades que debían cumplirse aproximadamente cada dos semanas, esto con orientación y apoyo técnico continuos.

Contexto de la IA responsable en México

Como se señaló en la introducción, los posibles marcos normativos incluyen consideraciones de TyE, pues se trata de un principio clave en las guías éticas globales de IA. Sin embargo, estos debates se han llevado a cabo en Europa, Estados Unidos y algunos países asiáticos. Dado que la adopción de estos sistemas es más reciente en América Latina, hay menos conciencia en la región sobre lo que significa utilizar la IA de forma responsable y el diálogo en torno a posibles marcos normativos para el uso y desarrollo responsables de la IA aún es emergente. Colombia fue el único país que contaba con un marco normativo para la IA en marzo de 2023. A pesar de los niveles desiguales de adopción global, es importante para México y el resto de la región (así como para los países de ingresos bajos a medios en conjunto) contribuir a las conversaciones internacionales para que las perspectivas, retos y oportunidades locales sean considerados en el desarrollo de buenas prácticas y normas internacionales.

Dicho esto, México ha avanzado en temas de IA en los últimos años. En 2018 se creó una Estrategia Nacional de Inteligencia Artificial[21], la cual fue desarrollada por C Minds, Oxford Insights y la Embajada Británica en México. Posteriormente fue adoptada por la Coordinación Nacional de Estrategia Digital, posicionando a México como uno de los 10 primeros países del mundo en contar con una estrategia de IA. Este documento abordaba las ventajas, oportunidades y retos que el país enfrentaba en materia de IA, así como recomendaciones a corto y mediano plazo para actores del ecosistema. A pesar de estos esfuerzos iniciales de gobernanza de la IA, el país debe seguir reforzando y creando políticas públicas en torno a la IA. De hecho, no hay suficientes estrategias ni herramientas de mitigación de los posibles impactos sociales negativos de los sistemas de A/TDA. De acuerdo con la Encuesta Nacional de Inteligencia Artificial realizada por IA2030Mx en 2019[22], en México el 45% de las personas estaban un poco preocupadas por las implicaciones éticas o los posibles impactos sociales negativos vinculados al desarrollo de la IA, tales como el sesgo y la privacidad de los datos.

En cuanto a la TyE en el contexto mexicano, algunas leyes y reglamentos tales como la Ley Federal de Protección al Consumidor obliga a difundir la información o publicidad relativa de los bienes y servicios de forma veraz y comprobables. Aunque el reglamento no revela cómo debe hacerse cuando se trata de sistemas de IA, sí sugiere una exigencia general de transparencia en los bienes y servicios.

Dada la creciente repercusión de los sistemas de A/TDA en nuestras vidas, es cada vez más importante para México y la región llevar a cabo experimentos que contribuyan a la confiabilidad de los sistemas de A/TDA.

¿Qué son la transparencia y la explicabilidad en el contexto de los sistemas de A/TDA?

Dado que los sistemas de A/TDA impactan cada vez más en nuestra cotidianidad y oportunidades al usarse en sectores como finanzas, educación y salud, por mencionar algunos, es importante que entendamos cómo y por qué se están llegando a las decisiones que nos afectan con el fin de preservar nuestra capacidad humana de determinismo a sistemas inteligentes[23]. Esto puede suponer un reto técnico, sobre cuando estas decisiones son tomadas por sistemas de aprendizaje profundo. Ya que estos sistemas funcionan como cajas negras, lo que significa que su funcionamiento interno es opaco (incomprensible) y dificulta entender el razonamiento detrás de la toma de decisiones.

Lo anterior es importante a la hora de tomar decisiones que no discriminen a personas o grupos poblacionales debido a errores y sesgos no deseados en los datos de entrenamiento o del algoritmo[24]. Además, de permitir la creación de empresas sustentables a largo plazo, ya que la alineación con prácticas responsables reduce la posibilidad de crisis relacionadas con la publicidad negativa.

Si bien la **transparencia** puede llegar a confundirse con el compartir los secretos de los algoritmos de las empresas, no es así. Este término tiene varios significados. En el contexto de este prototipo implica revelar el uso de un sistema de A/TDA (es decir, cuando se genera una predicción, recomendación o decisión, o cuando el usuario está interactuando directamente con un agente impulsado por IA, como un chatbot). El nivel de divulgación debe ser proporcional al impacto potencial del sistema en los derechos y libertades de las personas usuarias. Las y los usuarios deben comprender cómo el sistema se desarrolla, entrena, opera e implementa dependiendo de la aplicación.

Según el Centro Berkman Klein[25], para poder cumplir con el principio de transparencia los sistemas de A/TDA deben diseñarse e implementarse de forma que sea posible la supervisión de sus operaciones. El principio de transparencia debe aplicarse a todo el ciclo de vida (desarrollo e implementación) de los sistemas de A/TDA, incluyendo la selección de datos de entrenamiento, los algoritmos y el propio modelo.

De acuerdo con el Grupo de Alto Nivel sobre IA creado por la Comisión Europea, la **explicabilidad** se refiere a la capacidad de las personas afectadas por los resultados de un sistema entiendan por qué se ha llegado a un resultado concreto. Para ello es necesario conocer qué atributos o variables influyeron en la decisión final. En este sentido, la explicabilidad está estrechamente relacionada con la transparencia, ya que los resultados y subprocesos deben ser comprensibles y trazables[26].

Los principios de la IA de la OCDE[27] establecen que para lograr TyE *"la información relevante para el contexto y coherente con el estado de la técnica debe divulgarse de la siguiente manera:*

- *fomentar la comprensión general de los sistemas de IA para que las partes interesadas sean conscientes de sus interacciones con los sistemas de IA, incluso en el lugar de trabajo;*
- *permitir que las partes afectadas por un sistema de IA puedan comprender el resultado; y*
- *permitir a las partes afectadas por un sistema de IA impugnar sus resultados basándose en información sencilla y comprensible sobre la lógica y los factores que sirvieron de base para la predicción, recomendación o decisión".*

Es importante no confundir estos términos con la interpretabilidad. Aunque son principios complementarios, tienen significados diferentes.

Para efectos de este prototipo de política pública los conceptos se definieron del siguiente modo (para más información, véanse los anexos A y B):

Transparencia

La transparencia beneficia sobre todo a las personas usuarias. Les permite decidir si confiar en un sistema de IA dándoles información sobre el mismo. La cantidad de información varía, pero en general se acepta que los fabricantes de productos deben dar suficiente información a las personas para que entiendan el sistema de IA.

Otra forma de entender la transparencia es como divulgación de información por parte de las y los fabricantes de productos. Las personas que interactúan con un sistema de IA no tendrán la misma información que las personas que lo crearon. La transparencia ayuda a equilibrar estas dos partes.

Para ser considerados transparentes, las y los fabricantes de productos pueden considerar si son claros, abiertos y honestos sobre cómo se construye, cómo opera y cómo funciona un sistema de IA. Probablemente no sea suficiente con generar esta información y dejarla a la vista, pues para que alguien sepa si debe confiar en el sistema debe entender esa información. Por eso la transparencia suele exigir también que las aclaraciones estén en un formato inteligible.

La transparencia es también un mecanismo que permite la rendición de cuentas al hacer posible que los reguladores examinen los resultados.

Interpretabilidad

Este punto garantiza que los fabricantes de productos puedan predecir, de forma coherente, el modo en el que un modelo de IA toma decisiones. Esto ayuda a garantizar que actuará según lo previsto y a promover un sistema confiable de IA. No se trata tanto de entender el "por qué" de un sistema, sino de poder predecir lo que hará un modelo determinado. Cuando existe un alto nivel de consistencia en la predicción, se dice que un modelo es interpretable[28].

Sin embargo, investigadores y personas expertas suelen coincidir en que el aprendizaje profundo y los modelos de "caja negra" pueden dificultar la interpretación. Estos sistemas de IA pueden ser difíciles de entender para los humanos, complicando así una predicción precisa de lo que harán.

Un método para abordar el reto de la interpretabilidad es a través de "*Model Cards*" que ilustran el funcionamiento interno de un sistema de IA. Dichas fichas son interpretables si el modelo se ha analizado utilizando marcos de interpretabilidad como Captum.

Explicabilidad[29]

La explicabilidad de la IA beneficia a la persona usuaria del producto. Ayuda a decidir cuándo confiar en un producto basado en IA, asegurándose de que comprenden sus interacciones con un sistema de este tipo. El objetivo de la explicabilidad de la IA no es garantizar que cada persona tenga profundos conocimientos técnicos sobre un sistema y su funcionamiento, sino que trata más bien de ayudar a entender cómo un sistema de IA influye en su experiencia y cuánto control tienen sobre la interacción con un sistema determinado.

Puede dar a alguien la opción de cambiar o personalizar su experiencia, o informar lo que un sistema no puede hacer.

Tabla 1. Basado en el glosario de TTC Labs, disponible en: https://www.ttclabs.net/glossary

Objetivos para el programa Open Loop México

El programa de Open Loop México tenía los siguientes objetivos:

- proporcionar recomendaciones de política pública a reguladores sobre tecnologías como los sistemas de A/TDA basadas en pruebas compartidas por los desarrolladores de A/TDA;
- brindar una oportunidad para que reguladores y empresas se adelanten a las cuestiones emergentes sobre el desarrollo y uso ético de la IA en Latinoamérica;
- desarrollar mecanismos de colaboración entre reguladores e innovadores en temas de TyE para un aprendizaje ágil y dinámico sobre el tema;
- contribuir a la conversación internacional sobre la ética de la IA a través del ejercicio práctico.
- fortalecer el conocimiento sobre este tema, sus oportunidades y desafíos, y promover la creación de marcos normativos de TyE para los sistemas de A/TDA;
- facilitar una mejor comprensión del diseño, desarrollo y despliegue de los sistemas de A/TDA;
- aumentar la concienciación sobre la importancia de salvaguardar los derechos y libertades de las personas en el desarrollo y aplicación de los sistemas de A/TDA;
- aclarar y aplicar las mejores prácticas internacionales, para que el ecosistema mexicano aumente la confianza en el desarrollo y uso de sistemas autónomos e inteligentes; y
- proporcionar a las empresas un conjunto adecuado de directrices, herramientas y prácticas para garantizar sistemas de A/TDA más transparentes y explicables.

Actores clave

La siguiente tabla proporciona más información sobre los actores involucrados en Open Loop México:

Socios de Open Loop México

Meta

Open Loop es un programa global que pone en contacto a los responsables políticos y a las empresas tecnológicas para ayudar a desarrollar políticas eficaces y basadas en pruebas en torno a la IA y otras tecnologías emergentes.El programa, iniciado y apoyado por Meta (antes Facebook), se basa en la colaboración y las contribuciones de un consorcio compuesto por reguladores, gobiernos, empresas tecnológicas, académicos y representantes de la sociedad civil. A través de métodos de gobernanza experimentales, los miembros de Open Loop co-crean prototipos de políticas y prueban enfoques nuevos y diferentes de leyes y reglamentos antes de que se adopten, mejorando la calidad de los procesos de elaboración de normas en el ámbito de la política tecnológica.

Rol: proveedor de metodología basada en el programa global, co-diseñador del marco normativo y del manual (prototipo), y co-líder de la iniciativa Open Loop México.

C Minds y el Eon Resilience Lab de C Minds

C Minds es una organización liderada por mujeres que promueve la exploración, el desarrollo y uso responsable de tecnologías de frontera para el beneficio de América Latina, En 2019, co-publicó la Estrategia Nacional de IA de México, posicionando al país entre los 10 primeros del mundo en tener una.

El Eon Resilience Lab es el área de C Minds comprometida con preparar a individuos para el futuro, dados los cambios acelerados generados por las nuevas tecnologías y generar estrategias de inclusión digital al colaborar con instituciones públicas y privadas.

Rol: Co-diseñador del marco, co-líder de la iniciativa, coordinador e implementador principal del programa.

Banco Interamericano de Desarrollo (BID) - Sector social

El sector Social (SCL) del Grupo BID está conformado por un equipo multidisciplinario que actúa bajo la convicción de que la inversión en las personas permite mejorar sus vidas y superar los desafíos del desarrollo en América Latina y el Caribe. Formula soluciones de política pública para reducir la pobreza y mejorar la prestación de servicios de educación, trabajo, protección social y salud.

Rol: Acompañamiento y aportación de insumos y conocimiento, así como identificación de oportunidades.

Socios de Open Loop México

BID Lab	Es el laboratorio de innovación del Grupo BID, donde se promueve el desarrollo a través del sector privado al apoyar y probar nuevas soluciones para enfrentar la problemática de inclusión económica y social en América Latina y el Caribe. **Rol:** Acompañamiento y aportación de insumos y conocimiento, así como identificación de oportunidades.
Iniciativa fAIr LAC del BID	Es una alianza (liderada por el BID) entre los sectores público y privado, la sociedad civil y la academia, para incidir tanto en la política pública como en el ecosistema emprendedor en la promoción del uso responsable y ético de la IA. Actualmente se han establecido 4 hubs (ecosistema habilitador para el desarrollo e implementación de la iniciativa fAIr LAC): Jalisco (México), Costa Rica, Colombia y Uruguay.
Instituto Nacional de Transparencia, Acceso a la Información y Protección de Datos Personales (INAI)	Es el organismo constitucional autónomo que tiene como objetivo garantizar el cumplimiento con el derecho del acceso a la información pública y el de la protección de datos personales. Para el primero, garantiza que cualquier autoridad en el ámbito federal, órganos autónomos, partidos políticos, fideicomisos, fondos públicos y sindicatos; o cualquier persona física, moral que reciba y ejerza recursos públicos o realice actos de autoridad entregue la información pública que se solicite. Para el segundo, garantiza el uso adecuado de los datos personales, así como el ejercicio y tutela de los derechos de acceso, rectificación, cancelación y oposición que toda persona tiene con respecto a su información. **Rol:** Acompañamiento, aportación de insumos y receptor de las recomendaciones de política pública.

Tabla 2. Socios de Open Loop México

El programa también incluyó a un grupo de personas expertas en distintos temas de IA que apoyaron el diseño y la ejecución del programa, revisando y complementando los documentos probados. También prestaron acompañamiento directo a las empresas participantes. La siguiente lista incluye a las personas expertas que integraron este grupo:

- **Carla Vázquez Wallach**, Fundadora y Directora General de Legal + Innovation en México y
- **Daniel Castaño**, Fundador y socio de Mokzy, profesor de la Universidad Externado de Colombia e investigador y consultor especializado en IA, ética digital y regulación
- **Edson Prestes**, Investigador y profesor de ética de la IA en el Instituto de Informática de la Universidad Federal de Rio Grande do Sul (Brasil), miembro sénior de la Sociedad de Robótica y Automatización del Instituto de Ingenieros Eléctricos y Electrónicos (IEEE RAS) y la Asociación para las Reglas, miembro de la Asociación para la Regulación de la Inteligencia Artificial, y miembro del *Brain Hive* de C Minds
- **Guillermo Larrea**, Abogado corporativo con enfoque en América Latina en Jones Day
- **Rafael Ramírez de Alba**, Profesor del Departamento de Entorno Económico de la Escuela de Negocios en el IPADE.
- **Ricardo Baeza-Yates**, Director de investigación del Instituto de Inteligencia Artificial Experiencial de Northeastern University en el Silicon Valley.

El prototipo de prueba se puso a prueba con las siguientes empresas de IA, todas ellas con operaciones en México:

Empresas participantes

Nombre	Sector	Etapa	Modelo de negocios	Descripción
ai360 * Analítica Inmobiliaria	*Inmobiliaria	Escalamiento	B2B y B2G	Es una plataforma que estima los precios de la vivienda de forma más rápida y objetiva, comparando varias propiedades simultáneamente. El modelo evalúa los atributos de la propiedad mediante deep learning y reconocimiento de imágenes.
Fincomún.*	*Finanzas	Empresa	B2C y B2B2C	Una corporación financiera que ofrece créditos y préstamos a comunidades no atendidas por instituciones de crédito tradicionales, usualmente por ser consideradas de alto riesgo, utilizando modelos basados en el aprendizaje automático para los procesos de aprobación, crédito o préstamo.
helKi	Educación	Etapa temprana	B2C	Aplicación que acompaña y orienta de forma profesional y personalizada a los padres, madres o personal cuidador en los retos diarios de la crianza, por ejemplo prediciendo situaciones de riesgo en el desarrollo, a través de una asistente virtual conversacional.
hitch.	Recursos Humanos	Etapa temprana	B2B	Plataforma que optimiza y facilita la toma de decisiones en el proceso de selección de talento por los equipos de Recursos Humanos al proponer entrevistas realizadas por un sistema inteligente. Dentro de la tecnología de IA utilizada está aprendizaje automatizado y reconocimiento de imágen.
inndot PIENSA SOLUCIONES	Comunicación	Escalamiento	B2B	Plataforma de monitoreo y gestión de redes sociales y medios digitales para empresas y gobiernos, donde se proponen respuestas automatizadas a consultas.

Continúa en la siguiente página...

Nombre	Sector	Etapa	Modelo de negocios	Descripción
LUZI	Salud	Consolidación	B2B2C	Dispositivo y sistema de control del embarazo que permite medir el riesgo obstétrico (triage), el riesgo fetal a través de una semaforización y realiza un pre diagnóstico diario de la mujer embarazada para disminuir la recurrencia en muerte materna-fetal.
nauphilus *	*Finanzas y Recursos Humanos	Empresa	B2B y B2C	Plataforma que evalúa entrevistas digitales con herramientas biométricas y preguntas especializadas para ayudar a las empresas en su evaluación de riesgos, ya sea para ofrecer servicios y productos financieros o para procesos de reclutamiento.
nowports	Logística	Consolidación	B2B	Plataforma que agiliza las tarifas de transporte, tiempo de llegada y de retención de los contenedores de carga, entre otros sistemas que ayudan a la optimización de los procesos de logística.
OS City *	*GovTech	Escalamiento	B2G	Plataforma que permite que los gobiernos ofrezcan sus procesos administrativos, licencias, permisos y servicios de forma digital y segura, al usar IA y blockchain.
rhisco TO COMPLY & COMPETE	RegTech	Empresa	B2B	Herramienta que facilita la identificación de obligaciones para el cumplimiento normativo a través de la consulta y análisis de documentos regulatorios usando el reconocimiento de lenguaje natural.

Tabla 3. Empresas participantes en Open Loop México

**Debido a los impactos económicos de COVID-19, cuatro de las empresas participantes no pudieron continuar con el programa.*

Metodología

El proyecto inició con la creación de un marco normativo sobre TyE y un manual práctico de apoyo, creados conjuntamente por Meta, Eon Resilience Lab de C Minds y el BID. Los documentos fueron puestos a prueba por 10 empresas mexicanas con un proceso de acompañamiento especializado y personalizado, el cual incluyó espacios de exploración, iteración constante, desarrollo tecnológico, consulta con personas expertas y capacitaciones personalizadas.

El propósito del programa nunca fue evaluar los productos, servicios o modelos de negocio de estas empresas, sino analizar la aplicabilidad del marco normativo y fortalecerlo con base en las experiencias de las empresas participantes, así como producir recomendaciones claras de política pública para instituciones reguladoras de México.

La Figura 1 ilustra las etapas clave en el desarrollo del prototipo.

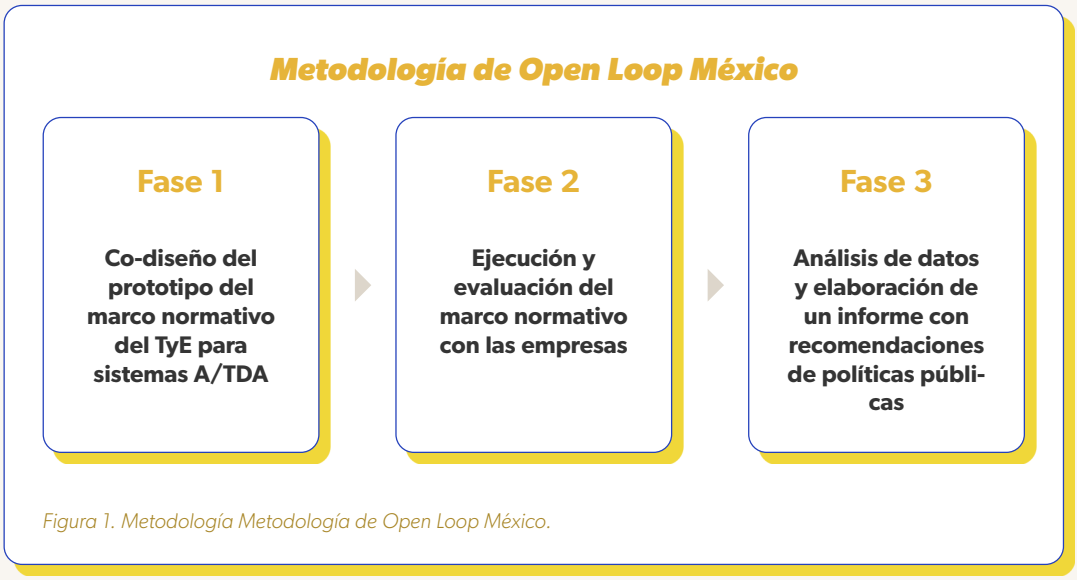


Figura 1. Metodología Metodología de Open Loop México.

Es importante señalar que la propuesta de marco no es una propuesta legislativa ni reemplaza cualquier legislación vigente, sino como un instrumento de exploración para generar la información necesaria para generar posibles futuras políticas públicas basadas en evidencia al respecto a nivel nacional y que sumen a la conversación internacional. Se trata únicamente de un punto de partida, una plataforma de experimentación, y no una conclusión.

El marco normativo y el prototipo del manual

La diferencia entre México y otros países en los que se había implementado el programa de Open Loop, es que en el caso mexicano no se contaba con un marco normativo existente o propuesto sobre uso responsables de tecnología que pudiera probarse en el programa. Por lo tanto, para llevar a cabo el prototipo de política pública, se desarrollaron dos documentos: una propuesta de marco normativo y un manual para la implementación de TyE basado en buenas prácticas internacionales.

La **propuesta de marco normativo** se desarrolló defendiendo los valores de autonomía humana, la determinación y el respeto de los derechos humanos, con el objetivo de promover la adopción de los pilares de TyE en los sistemas de A/TDA de las empresas para reducir los posibles impactos

adversos sobre los derechos y libertades de las y los usuarios y las personas. Se formuló y estructuró como cualquier otro marco normativo, pero, como se mencionó anteriormente, sin vinculación jurídica. Su único propósito era obtener observaciones sobre el contenido y formato de las empresas participantes en el programa.

Los socios del programa también desarrollaron un **manual** complementario que sirvió de guía para cumplir con los requisitos del marco normativo. Este manual incluía explicaciones detalladas de cada artículo del marco con definiciones más amplias y ejemplos prácticos, así como herramientas para llevar a cabo el proceso de refuerzo de TyE y recursos adicionales para los lectores que desearan profundizar en el uso responsable de los sistemas de A/TDA.

Implementación del prototipo

Tanto el marco normativo como el manual fueron puestos a prueba por las empresas participantes mediante una serie de actividades. Cada actividad examinaba partes de los documentos y duraba entre una a dos semanas. Mediante cuestionarios realizados a través de la plataforma de etnografía móvil llamada *dscout*[30], las empresas comentaron su proceso y su experiencia en cada actividad. Esta etapa comenzó a inicios de febrero del 2021 y finalizó en agosto del mismo año.

El ciclo de implantación del marco normativo y del manual por parte de las empresas se dividió en cuatro etapas clave:

- ***Etapa 1. Aproximación a los conceptos clave***

  En esta primera etapa las empresas participantes profundizaron en el conocimiento de la TyE y otros principios éticos de la IA, también recibieron una breve introducción a las políticas públicas.

- ***Etapa 2. Evaluación del marco normativo y de la propuesta del manual***

  El objetivo de esta etapa era determinar cuál era el grado de comprensión de los documentos (marco y manual), la viabilidad de su aplicación en los productos y servicios de cada empresa, y cómo cada empresa elegiría crear y enfocar su solución de TyE. Este último ejercicio fue dirigido por un marco de escenarios de explicabilidad que ayudó a las empresas a definir su solución de TyE mediante la selección del público objetivo, considerando los factores contextuales que impulsaron el desarrollo de la solución, eligiendo el propósito de la explicación y, escogiendo el contenido y la profundidad de la información a compartir. El recuadro 1 presenta las distintas opciones utilizadas para generar estos escenarios.

Recuadro 1. Escenarios de explicabilidad

La siguiente tabla ayudó a las empresas a seleccionar sus escenarios de explicabilidad, facilitando la creación de una solución de TyE a la medida.

Elementos	Posibles opciones[31]
Público objetivo: ¿a quién va dirigida esta explicación?	• Regulador (auditor externo / institución reguladora). • Socio comercial (otra empresa, cliente o proveedor). • Consumidor (persona usuaria del producto o servicio). • Sociedad (el público en general). • La propia empresa (por ejemplo: trabajadores o personal).
Contexto: ¿hubo alguna razón específica que llevó a la creación de una solución de TyE?	• Adopción de buenas prácticas. • Diferenciación con respecto a la competencia. • Enfoque proactivo para mejorar la transparencia y la rendición de cuentas. • Anticipación a las necesidades y expectativas de los y las clientes.
Propósito: ¿qué se pretende conseguir?	• Aumento de la concienciación de la persona usuaria sobre su interacción con un sistema de IA. • Facilitar la comprensión de los componentes y la gobernanza del sistema de IA. • Permitir la retroalimentación de consumidores, personas usuarias, socios o reguladores. • Involucrar a las personas usuaeias en la mejora del sistema o modelo de IA. • Ofrecer recomendaciones o predicciones del sistema de IA a las personas o entidades afectadas • Influir en el comportamiento futuro de las personas afectadas por las decisiones, recomendaciones o predicciones del sistema de IA, • Establecer responsabilidades y rendición de cuentas para un funcionamiento más transparente y explicable del sistema o modelo de IA y su gobernanza.

Continúa en la siguiente página...

Elementos	Posibles opciones[30]
Contenidos: ¿qué información y hasta qué punto debe compartirse?	• Justificación (información que describe cómo se tomó la decisión, recomendación o selección del sistema de IA). • Responsabilidad (información sobre quién participa en el desarrollo, gestión y aplicación del sistema de IA). • Seguridad y rendimiento (información sobre la precisión, viabilidad, seguridad, robustez de las decisiones y comportamiento), • Datos y modelos (información sobre la capacitación de bases de datos, modelos algorítmicos utilizados, etc.) • Equidad (información que garantice que el sistema de IA no está sesgado injustamente). • Impacto (información sobre los impactos del uso del sistema de IA en las decisiones de las personas).

- ***Fase 3. Aplicación de documentos personalizados***

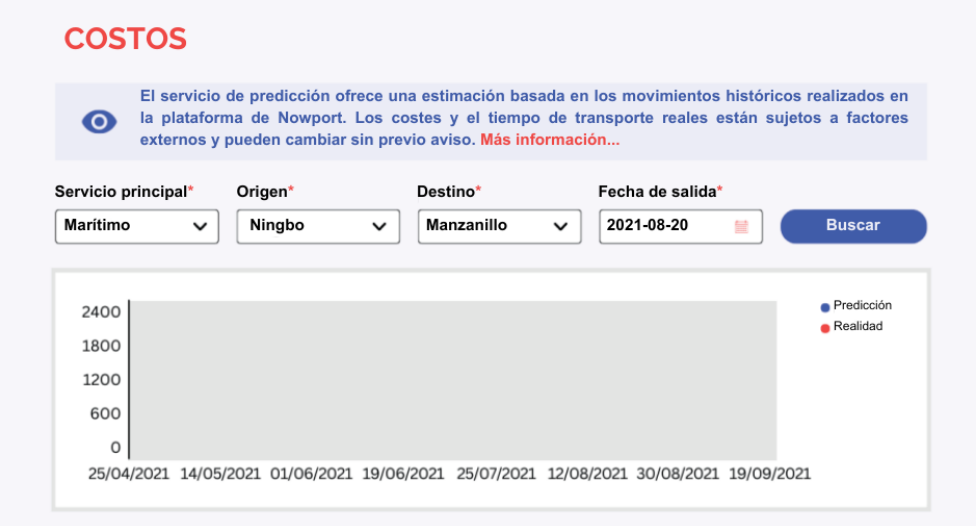
  Se elaboraron planes personalizados para cada empresa con actividades basadas en la información del marco normativo y el manual. Estos planes diferían en función de si las empresas desarrollaron su propio sistema A/TDA o contaban con un proveedor para hacerlo. Las empresas con desarrollos propios exploraron la calidad de su sistema y realizaron análisis de sesgos. A su vez, las empresas con sistemas de A/TDA de terceros exploraron las opciones y, en la medida de lo posible, los mismos temas que sus colegas, junto con su proveedor de sistemas de A/TDA.

- ***Fase 4. Creación de una solución de TyE***

  Con los conocimientos adquiridos durante el programa, en esta última etapa las empresas diseñaron y presentaron su propia solución de TyE a un grupo de personas de entre los socios para recibir retroalimentación y reiterar su solución. Los resultados pueden verse en el recuadro 2.

Recuadro 2. Soluciones de TyE desarrolladas por las empresas

La mayoría de las soluciones se centraban en explicar las razones para utilizar sistemas de A/TDA, así como el proceso de selección y preparación de datos. Para lograr esta explicación, la mayoría recurrió al uso de un lenguaje conciso con palabras sencillas que permita a las personas usuarias finales o sujetos comprender el proceso de toma de decisiones del sistema de una forma fácil e intuitiva. Lo consiguieron evitando términos técnicos y utilizando un lenguaje claro y conciso que pueda ser fácilmente comprendido por un público no experto.



Réplica de la plataforma de Nowport, la cual incluye su propuesta de solución de transparencia y explicabilidad (replicada por C Minds)

Las representaciones visuales son muy útiles para explicar los sistemas de IA, ya que proporcionan una forma sencilla e intuitiva de representar información, modelos y relaciones complejas. Dentro de estas soluciones, la implementación de material visual como representaciones gráficas, imágenes y vídeos explicativos ayudaron a que la descripción del sistema de A/TDA fuera más claro para la persona usuaria final o sujeto. Aunque la mayoría de ellas se centraron en el texto, algunas empresas implantaron y manifestaron su interés por el material visual.

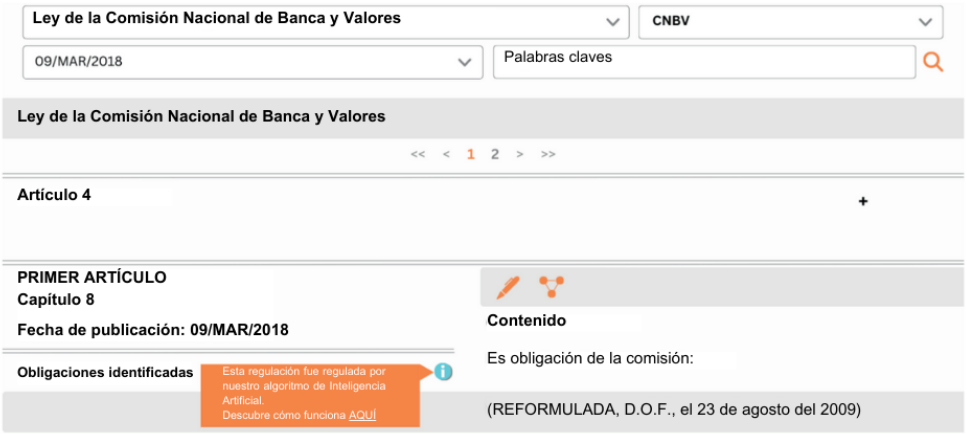
Continúa en la siguiente página...

Este fue el caso de Rhisco, que decidió crear una solución visual para sus usuarios y el público en general y así explicar su modelo de forma más comprensible. Esta solución constaba de dos niveles:

El primero consistía en un vídeo, así como una serie de imágenes y gifs, que explicaban el funcionamiento general del sistema A/TDA en términos generales.

Ejemplo del material visual creado por Rhisco para explicar el funcionamiento de su sistema A/TDA.

Este material visual aparecía automáticamente al iniciar sesión como persona usuaria de la plataforma. Además integraban una notificación para informar a la persona usuaria cuando el sistema A/TDA le hacía una recomendación.



El segundo nivel de la solución permitía a las personas usuarias solicitar más información sobre la justificación de una recomendación concreta realizada por el sistema.

La mayoría de las empresas participantes incluyeron su mensaje de explicabilidad como parte de la experiencia del usuario con el producto o servicio. Esta experiencia de usuario incluía la activación de notificaciones, mensajes y vídeos al acceder a la aplicación o página. Asimismo, la mayoría de las empresas optaron por crear mensajes que la persona usuaria o sujeto pudiera visualizar siempre que lo necesitara a través de un botón permanente que mostrara la información.

Criterios de evaluación

Dado que el objetivo del programa era probar la eficacia del prototipo de política pública, se evaluaron los siguientes criterios:

- **Claridad del marco normativo/propuesta de prototipo**

  Se refiere a la claridad de la propuesta y sus requisitos para los destinatarios y a la comprensión de los requisitos para su cumplimiento.

- **Eficacia de la propuesta de marco regulador**

  Dado que el propósito de las políticas públicas es atender problemas públicos específicos de manera competente, este criterio se refiere a la medida en que la propuesta contribuye a abordar el reto público en cuestión, en este caso, la falta de transparencia y explicabilidad de los sistemas de A/TDA.

- **Viabilidad del prototipo de política pública**

  Se refiere a la posibilidad, en términos de recursos técnicos, económicos, humanos y de tiempo, de implementar el marco normativo con el fin de evitar el surgimiento de **barreras innecesarias.**

Limitaciones del ejercicio

Si bien este ejercicio ofrece una serie de aprendizajes que sirven para el fortalecimiento de ejercicios similares y la creación de política pública en torno al tema de TyE de sistemas de A/TDA, hay que tener en cuenta algunas limitaciones, las cuales se enumeran a continuación:

- **Representatividad limitada en el tamaño de las empresas**

  Dado que la mayoría de las participantes en el programa eran pequeñas y medianas empresas (PYME)[32] o *startups*, las conclusiones y recomendaciones pueden no ser aplicables a todo tipo de empresas. Los socios recomiendan realizar ejercicios similares con empresas grandes y multinacionales. No obstante, al enfocarse en empresas nacientes y de madurez inicial, las recomendaciones que se desprenden del prototipo tienen el mérito de ajustarse a empresas con menos recursos técnicos, económicos y humanos (siendo MIPyMES) en México, así fomentando la competitividad en la industria. Según el Instituto Nacional de Estadística y Geografía (INEGI), el 99.8% de los negocios en México son MiPyMEs[33].

- **Representatividad limitada de las empresas con sistemas de terceros**

  Inicialmente eran 4 las empresas que utilizaban modelos de terceros, pero una de ellas abandonó ejercicio, quedando solo 3. Dentro de las razones se encuentran que el marco tenía más sentido para las empresas que habían desarrollado sus propios sistemas de A/TDA, puesto que ya habían participado en un ejercicio similar y dado que ya habían trabajado todo lo posible en TyE.

Evaluación del prototipo de política pública

Al hacer la evaluación del prototipo de política pública en TyE y sistemas A/TDA se puede concluir de forma general que fue un éxito. El objetivo de evaluar la adecuación del marco normativo y el manual técnico se ha cumplido, lo que provee recomendaciones claras de política pública a las instituciones reguladoras en México. La evaluación consideró tres criterios: claridad del prototipo de política, su eficacia y la viabilidad. A continuación se analiza detalladamente cada criterio.

Claridad del marco normativo

Las empresas que participaron en la evaluación consideraron que la descripción y el objetivo del prototipo normativo eran suficientemente claros. Esto se reflejó en su cumplimiento y comprensión de las actividades destinadas a mejorar la TyE de sus sistemas. Dentro del entendimiento del prototipo, las empresas identificaron como principal beneficio el trazar un piso parejo en cuanto a léxico y terminología entre participantes y el equipo de Open Loop México para que se pudiera hablar el mismo idioma.

Cuando se les preguntó cómo el programa les había brindado herramientas necesarias para traducir los artículos del marco normativo en acciones concretas, todas las empresas manifestaron tener una idea clara para aplicar las ideas y los artículos mencionados en el documento. Por ejemplo, helKi mencionó que el documento era útil para aclarar las implicaciones de la aplicación de las recomendaciones y servía de referencia para la acción. Sin embargo, algunas empresas sugirieron que los documentos podían ser más accesibles, señalando la falta de claridad en conceptos debido a la falta de detalle en las definiciones. En este sentido, las empresas recomendaron ilustrar la aplicación de cada principio con escenarios hipotéticos.

Eficacia de las políticas públicas

Con base en los resultados y las opiniones de las empresas participantes, se puede concluir que, si bien existe espacio para mejorar, el prototipo de política pública fue eficaz. Las empresas que completaron el programa cumplieron con éxito el propósito del prototipo que era crear y reforzar sus mecanismos de TyE, dando como resultado la implementación de soluciones adaptadas a sus contextos específicos.

Cuando se les pidió a las empresas que calificaran su cumplimiento de las expectativas del programa en una escala de 1 a 10, las empresas dieron un promedio de calificación de 9, lo que indica un alto nivel de satisfacción. También describieron su experiencia en el programa como un proceso retador pero útil para crear conciencia sobre la necesidad de desarrollar herramientas responsables de IA. Lo consideraron un proceso gratificante que trazó el camino para seguir participando en iniciativas similares.

Las empresas participantes destacaron contribuciones significativas del programa. Entre ellas, la reducción de los sesgos; una mejor comprensión de la implementación de TyE y una explicación mejorada del funcionamiento del sistema IA para aumentar la confianza de la persona usuaria. También se consideraron resultados valiosos la profundización en el análisis de sesgos; el refuerzo de la comunicación con las y los clientes y, la mejora de prácticas de TyE mediante pruebas. El programa facilitó la concientización sobre los sesgos y la transparencia, la comprensión de los peligros del uso de datos sensibles y la exploración de las prácticas de terceros en las plataformas de TyE. Además adquirieron conocimientos básicos sobre el desarrollo de soluciones TyE, conciencia sobre las limitaciones éticas y la comprensión de las mejores prácticas para mitigar los problemas de aplicación. Las empresas también informaron sobre la aplicación de una metodología ética, el acceso a conocimientos especializados sobre sistemas de A/TDA responsables, la adopción de un enfoque más transparente y la aplicación de una política más transparente en materia de datos personales.

Todas las empresas expresaron su compromiso de implantar prácticas de TyE en todos sus productos y servicios como un proceso continuo. Algunas empresas ya habían establecido mecanismos permanentes, como Nowports, los cuales contaban con procesos de revisión y comunicación continua para abordar los sesgos internos y externos.

Viabilidad

La aplicación del marco normativo y el manual requiere diversos recursos, entre ellos financieros humanos, técnicos y de tiempo, los cuales son cruciales para la evaluación de la viabilidad. Durante la aplicación del prototipo, las empresas tuvieron en cuenta los costos, principalmente en términos de inversión de tiempo y capital humano.

En cuanto a las limitaciones, las empresas expresaron que el tiempo invertido durante el programa no era excesivamente pesado. No obstante, debido a la limitada capacidad de mano de obra, les resultaba difícil priorizar las actividades del programa sobre otras responsabilidades. En términos de dificultad técnica, la puntuación promedio otorgada por los participantes fue de 5,6 sobre 10, siendo 10 considerado "extremadamente difícil". Sin embargo, esta clasificación varió dependiendo de la complejidad y sensibilidad del modelo utilizado por las empresas. Asimismo, algunas empresas mencionaron que, para gente sin conocimiento técnico, esta dificultad podría aumentar sin llegar a ser imposible.

Las empresas participantes también se enfrentaron a varios retos durante la implantación del prototipo. Entre ellos, la falta de tiempo debido a otras prioridades, la escasez de recursos humanos, dificultades en la asignación de recursos, la gestión del tiempo entre equipos, la resistencia de algunas áreas de las empresas, limitaciones de recursos y comprensión del sistema A/TDA, y la falta de conocimientos técnicos y documentación. Al afrontar estos retos e implementar las recomendaciones derivadas de la evaluación, estas contribuirán a la mejora continua del marco normativo, fomentando una cultura de sistemas de A/TDA responsables y garantizando el uso ético y transparente de las tecnologías de IA.

A partir de las experiencias de las empresas participantes, si bien el prototipo de política pública pareció ser eficaz, en general, se identificaron áreas de oportunidad para seguir fortaleciendo el marco normativo aún más. Para más información detallada, véase la sección 5.

5

Modificaciones específicas a la propuesta de marco normativo y manual

Con base en los resultados y los aprendizajes las empresas en la implementación del prototipo, se sugieren los siguientes ajustes para fortalecer aún más los documentos propuestos:

Marco normativo

Propuesta de modificación de textos a partir de las observaciones (ediciones y adiciones en negro)	**Observaciones del equipo de implementación**
Ajuste de la definición de transparencia: La transparencia puede definirse como la práctica de revelar cómo y por qué un sistema tomó una decisión concreta, cuando la supervisión de operación es posible. Las personas afectadas deberían comprender cómo se desarrolla, capacita, opera y despliega un sistema A/TDA.	**Aunque el marco tiene una sección dedicada a las pruebas enfocadas en TyE durante la aplicación del prototipo, una definición concisa de transparencia, separada del concepto de explicabilidad, sería muy útil para diferenciar los principios.**
Ajuste de la definición de explicabilidad: La explicabilidad de un sistema de A/TDA se refiere a que las personas impactadas por los resultados de un sistema entiendan, en términos simples, por qué se ha logrado ese resultado. Para ello, es necesario saber qué atributos o variables influyeron en la decisión final.	**Como en el caso anterior, también recomendamos añadir otra definición de explicabilidad para reforzar la posibilidad de distinguir entre TyE.**
Para muchas aplicaciones de aprendizaje automatizado, el modelo es tan complejo que no puede interpretarse, por lo tanto, **"localmente interpretable"** se refiere a una explicación de cómo se ha llegado a una conclusión específica (en contraposición a cómo el modelo toma las decisiones en general).	**Dado que empresas confundieron los distintos conceptos presentados con respecto a la interpretabilidad local y su definición, la información se reformuló para diferenciar mejor estos conceptos.**

Manual

Propuesta de modificación de textos a partir de las observaciones (ediciones y adiciones en negro)	Observaciones del equipo de implementación
La interpretabilidad **implica que una persona puede entender el proceso de toma de decisiones, especialmente cómo el sistema A/TDA ha llegado a una conclusión. Es la capacidad de determinar la causa y efecto del modelo.**	**Durante la implementación del prototipo las empresas compartieron que la definición de interpretabilidad era algo ambigua. Por ello, para evitar confusiones, la explicación se reformuló (véase en azul).**
La explicabilidad es la forma en que la mecánica interna de un sistema de decisión automatizado puede explicarse en términos humanos. La diferencia con la interpretabilidad es sutil. **Mientras que la interpretabilidad proporciona una amplia comprensión del funcionamiento de un sistema, la explicabilidad ofrece una comprensión de todos los atributos y variables que influyen en la toma de decisiones.**	**La confusión entre los diferentes conceptos requiere más y mejores definiciones que ayuden su diferenciación.**
Interpretabilidad global (o los modelos globalmente interpretables) • **Este nivel de interpretabilidad hace referencia a que un ser humano es capaz de comprender el modelo completo.** Para las decisiones que requieren plena responsabilidad y justificación, generalmente son preferibles los modelos que son globalmente interpretables. **Interpretabilidad local** • **La interpretabilidad local se refiere a un resultado específico de todo el modelo.** La principal crítica y desafío de los modelos de caja negra es que son difíciles –si no imposibles– de comprender para los humanos.	**Dado que se observó confusión por parte de las empresas sobre cómo identificar que sus modelos eran globalmente o localmente interpretables, se añadió una frase para recalcar esta distinción.**

Paso 1: Determina el riesgo del proceso de toma de decisiones.

El uso de sistemas de A/TDA conlleva un impacto que, dependiendo de la aplicación y alcance, puede afectar directamente a la vida de las personas y el medio ambiente. Los sistemas de A/TDA tienen usos que podrían categorizarse como de bajo impacto y de alto impacto; la principal diferencia se encuentra en cómo podría afectar la vida de las personas y sus derechos. Se considera que son de alto impacto cuando pueden dañar la salud o la seguridad de una persona o infringir en los derechos fundamentales garantizados.

Por ello, las organizaciones usuarias deben prever y analizar si existen posibles impactos negativos importantes en los derechos y libertades de las personas.

Después de realizar esta evaluación de riesgos y con base en los resultados de su evaluación de riesgos, determina cuáles riesgos plantea el proceso de toma de decisiones para los valores individuales y / o colectivos.

La importancia de llevar a cabo evaluaciones de riesgos (para mitigar los posibles impactos negativos) se añadió al paso 1.

Por último, aunque el prototipo de política pública se centró en el TyE de los sistemas de A/TDA, sería pertinente considerar la integración de información relativa a otros principios éticos de la IA, por ejemplo, como la rendición de cuentas y los procesos de objeción.

6

Recomendaciones para la formulación de políticas enfocadas en la transparencia y explicabilidad

A partir de los resultados obtenidos en la aplicación del prototipo de política pública y de las observaciones compartidas sobre el marco normativo, se ofrecen una serie de recomendaciones complementarias a los documentos originales (véanse anexos A y B). Estas recomendaciones están dirigidas a las instituciones que deseen desarrollar políticas relacionadas con la TyE de los sistemas de A/TDA.

1. **Promover proactivamente la Inteligencia Artificial como prioridad nacional, centrándose en la puesta en práctica de principios confiables de IA.**
2. **Desempeñar un papel proactivo en la gobernanza de la IA en México.**
3. **Desarrollar capacidades para una IA confiable en organismos gubernamentales no técnicos y, en particular, en transparencia y explicabilidad (TyE).**
4. **Aumentar la capacidad técnica de la IA confiable en México.**
5. **Invertir en la investigación y el desarrollo de la IA confiable.**
6. **Fortalecer la capacidad de desarrollo y adopción responsable de IA en la fuerza laboral mexicana.**
7. **Ampliar la conciencia cívica sobre la IA en México.**

1 Promover activamente la Inteligencia Artificial como una prioridad nacional, con un enfoque en la operacionalización de los principios confiables de la IA.

- Las y los formuladores de políticas públicas podrían utilizar los recursos, las mejores prácticas internacionales y herramientas para crear una estrategia nacional de IA[34].
- Esta estrategia podría esbozar los objetivos políticos de la IA alineados a los Principios de la IA de la OCDE y la UNESCO, así como con las políticas que podrían ser necesarias para alcanzar dichos objetivos. La estrategia también podría incluir medidas específicas para promover la transparencia y la explicabilidad de los sistemas de IA[35].
- Este ejercicio debería ser un esfuerzo multilateral liderado por organismos gubernamentales nacionales. Asimismo, en la creación de la estrategia también podrían participar el sector privado, la academia y la sociedad civil a través de ejercicios innovadores.

2 **Desempeñar un papel proactivo en la gobernanza de la IA en México**

- Las y los formuladores de políticas públicas en México podrían desempeñar un papel proactivo en la gobernanza del desarrollo y el uso de la IA en el país: i) organizando y promoviendo ejercicios gubernamentales experimentales para identificar y abordar las oportunidades y los retos de la IA, como prototipos de políticas públicas y sandboxes regulatorios (antes de que se establezcan políticas/regulaciones), así como hackatones y concursos para comprender mejor las oportunidades y los retos en este ámbito; ii) desarrollar un marco normativo claro y conciso para la IA, basado en las necesidades locales y buenas prácticas internacionales y; iii) promover la colaboración intersectorial para garantizar que el marco normativo de IA sea comprensivo y refleje los puntos de vista de todas las partes interesadas.

3 **Desarrollar capacidades para tener una IA confiable, especialmente en materia de transparencia y explicabilidad (TyE) en organismos gubernamentales no técnicos.**

- Organizar e implementar sesiones y talleres de fortalecimiento de capacidades sobre las oportunidades y riesgos de la IA, con enfoque en la TyE. Las y los formuladores de políticas públicas podrían colaborar con organizaciones de la sociedad civil y la academia para organizar y llevar a cabo estas sesiones y talleres, así como impartir cursos masivos en línea (MOOCs, por sus siglas en inglés) para nivelar los conocimientos de los y las formuladores de políticas públicas, así como las y los funcionarios públicos, sobre los riesgos y oportunidades de la IA, especialmente en relación con la TyE. El fortalecimiento de sus capacidades y conocimientos les permitiría participar mejor en las conversaciones sobre el tema.

- Crear espacios regulares de diálogo con funcionarios y funcionarias gubernamentales, desarrolladores de IA y otras partes interesadas para debatir sobre temas relacionados con la TyE. Por ejemplo, las y los formuladores de políticas públicas podrían formar un grupo de trabajo integrado por funcionarios y funcionarias públicos, desarrolladores de IA y otras partes interesadas para debatir cuestiones relacionadas con TyE. Estas conversaciones podrían ayudar a crear consenso sobre las mejores prácticas de diseño, desarrollo, despliegue y uso de la IA, así como para identificar las áreas en las que se necesita más orientación.

4 Aumentar la capacidad técnica de la IA confiable en México.

- Los y las formuladores de políticas públicas podrían considerar el desarrollo de un conjunto de normas/protocolos técnicos para sistemas de IA en consulta con desarrolladores de IA, empresas y otras partes interesadas en el ecosistema mexicano de IA, basándose en buenas prácticas internacionales para garantizar que incluyan prácticas "*human-in-the-loop*" (es cuando los sistemas de IA y aprendizaje automático se construyen con interferencia humana en diferentes etapas del ciclo) cuando sea relevante y estén alineados con un enfoque de la IA centrado en el ser humano.

- Explorar el desarrollo de un marco normativo de gestión de riesgos, basado en el contexto mexicano, el cual sea altamente coherente e interoperable con las mejores prácticas internacionales y los esfuerzos de estandarización[36] del diseño, desarrollo y despliegue de sistemas de IA confiables, para reducir el potencial de impactos negativos inesperados. Estos son especialmente relevantes si las empresas optan por cumplir con los principios de TyE. Esto podría ser liderado por las instituciones reguladoras en colaboración con el ecosistema mexicano de IA.

- Además de crear recursos locales, también se debe de considerar la posibilidad de reunir los recursos internacionales existentes de países, empresas y organizaciones multilaterales en una página web gubernamental que se actualice periódicamente.

5 Invertir en la investigación y el desarrollo de la IA confiable.

- Las y los formuladores de políticas públicas podrían establecer incentivos financieros y no financieros para promover proyectos de investigación sobre TyE de la IA a través de organismos gubernamentales, universidades públicas y privadas, en colaboración con la industria y la sociedad civil para garantizar un enfoque práctico. También podrían considerarse las colaboraciones transfronterizas. En particular estos actores podrían invertir en investigación sobre:

    - i) técnicas para hacer que los sistemas de IA sean más transparentes y explicables, lo que podría incluir la investigación sobre métodos para visualizar el proceso de toma de decisiones de los sistemas de IA, así como la investigación sobre métodos para explicar los fundamentos de las decisiones de la IA; e

    - ii) investigación y herramientas para identificar y mitigar sesgos en los sistemas de IA, así como marcos normativos generales de gestión de riesgos.

- Además de fomentar y financiar la investigación, el gobierno, la academia y demás partes interesadas en la IA podrían crear espacios para compartir las principales enseñanzas, recomendaciones y herramientas resultantes de las actividades de investigación.

6 Fortalecer la capacidad de desarrollo y adopción de la IA responsable en la mano de obra mexicana.

- Promover la inclusión de cursos y módulos sobre consideraciones éticas en el desarrollo y adopción de sistemas de IA en carreras técnicas vinculadas a la ciencia de datos, informática e inteligencia artificial, entre otras. Esto podría aplicarse en espacios de educación formal como universidades y otras instituciones de aprendizaje o cursos, incluidos los de aprendizaje permanente.

- Las carreras enfocadas en ciencias sociales y humanidades, en espacios educativos formales e informales, también podrían ofrecer cursos y módulos de introducción a los sistemas de IA para crear una fuerza laboral más diversa que pueda centrarse en la IA responsable desde diferentes perspectivas. Esto podría ser promovido por Organismos de Certificación, así como las y los formuladores de políticas públicas a través de organismos como el Instituto Nacional de Transparencia, Acceso a la Información y Protección de Datos Personales (INAI), en colaboración con la industria, la sociedad civil y la academia, creando programas de capacitación sobre la importancia y cómo construir sistemas de IA transparentes y explicables, especialmente para desarrolladores.

7 Incrementar la conciencia cívica sobre la IA en México.

- Las y los formuladores de políticas públicas podrían lanzar una campaña de concientización pública sobre los riesgos y oportunidades de los sistemas de IA, resaltando la importancia de la TyE en los servicios y productos de IA. Esta campaña podría ayudar a impulsar las prácticas de TyE, como una ventaja competitiva para las empresas y como una solicitud de las personas consumidoras a sus proveedores de productos y servicios.

- La y los formuladores de políticas públicas y los organismos locales de juventud y educación, podrían seguir promoviendo programas de alfabetización digital en escuelas y universidades como cursos de aprendizaje permanente enfocados en la IA (una vez que hayan comprendido los fundamentos digitales) en colaboración con la sociedad civil y la academia.

- Apoyar el desarrollo y despliegue de recursos de alfabetización digital y de IA en español, y trabajar junto con el gobierno y actores alrededor de la IA para promover la alfabetización de la AI.

7

Conclusión

Los sistemas de A/TDA se están incorporando cada vez más en diferentes sectores en México, pues se están volviendo cruciales para el crecimiento económico y el desarrollo. Esta adopción masiva está creando nuevas oportunidades que deben ser aprovechadas pero, al mismo tiempo, plantea nuevos retos. En este sentido, es clave seguir impulsando la innovación, pero esto debe hacerse responsablemente para no vulnerar los derechos y libertades de las personas.

Este prototipo de política pública sobre la TyE de los sistemas de A/TDA proporciona una visión valiosa sobre la relevancia e importancia de incorporar estos principios en el desarrollo y uso de estos sistemas. Como se observó durante la aplicación del prototipo de política pública, por un lado, la transparencia y la explicabilidad permiten a las empresas comprender mejor el funcionamiento interno de sus sistemas de A/TDA, por ejemplo, incluyendo los riesgos que pueden plantear debido a sesgos no tratados. Por otra parte, la transparencia y explicabilidad de los sistemas de IA/TDA pueden generar más confianza en las personas usuarias final y sujetos, ya que comprenden mejor el funcionamiento de cómo se obtienen las soluciones; además se obtiene la sensación de que el proveedor es responsable de la solución. Los autores creen que este programa. Los autores consideran que este programa es un buen ejemplo de la aplicación de los principios de T&E

El programa también demostró la eficacia de los prototipos de políticas públicas para promover el valor y la importancia de un tema en particular, en este caso la TyE, y para generar nuevas perspectivas que contribuyan a la conversación global sobre la gobernanza de los sistemas de A/TDA en América Latina y el Caribe. Los prototipos de políticas públicas son mecanismos valiosos para fomentar el aprendizaje dinámico y la colaboración entre las partes interesadas para establecer un diálogo abierto e informado en la región sobre un tema tan relevante y complejo como es el caso de la TyE. Los autores y las organizaciones que participan en el programa esperan que contribuya a la creación de marcos normativos de la IA prácticos, inclusivos, operativos, adaptados y adaptables a distintos contextos, y que sitúen el beneficio social como foco de su impacto, todo en colaboración con partes interesadas clave.

Finalmente, los autores y las organizaciones involucradas esperan que las lecciones aprendidas inciten al desarrollo de ejercicios similares en América Latina y el Caribe.

Bibliografía

Ad Hoc Expert Group (AHEG) for the Preparation of a Draft text of a Recommendation the Ethics of Artificial Intelligence. (2020). Outcome document: First draft of the recommendation on the ethics of artificial intelligence. UNESCO. Obtenido de: https://unesdoc.unesco.org/ark:/48223/pf0000373434

Brown, T., & Katz, B. (2011). Change by design. Journal of Product Innovation Management, 28(3), 381-383.

Buchanan, C. (2018). Prototipo de política. Gobierno del Reino Unido. Obtenido de: https://openpolicy.blog.gov.uk/2018/11/27/prototype-for-policy/

Fjeld, J., Achten, N., Hilligoss, H., Nagy, A., & Srikumar, M. (2020). Principled artificial intelligence: Mapping consensus in ethical and rights-based approaches to principles for AI. Berkman Klein Center Research Publication No. 2020-1. Obtenido de: http://dx.doi.org/10.2139/ssrn.3518482

Hamon, R., Junklewitz, H., & Sanchez Martin, J. I. (2020). Robustness and explainability of artificial intelligence. Obtenido de: https://publications.jrc.ec.europa.eu/repository/handle/JRC119336

Hébert, M. (2019). A pilot is not a prototype: How to test policy ideas before scaling. Apolitical. (2019). Obtenido de: https://apolitical.co/solution-articles/en/a-pilot-is-not-a-prototype-how-to-test-policy-ideas-before-scaling

Hernández, M. (2019). Estrategia Nacional de Inteligencia Artificial va por sentido ético y responsable [National Artificial Intelligence Strategy based on ethics and responsibility]. Forbes México. Obtenido de: https://www.forbes.com.mx/estrategia-nacional-de-inteligencia-artificial-va-por-sentido-etico-y-responsable/

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

Institute of Electrical and Electronic Engineers. (2022). The IEEE Global Initiative on Ethics of Autonomous and Intelligent Systems. Obtenido de: https://standards.ieee.org/industry-connections/ec/autonomous-systems/

Instituto Nacional de Estadística y Geografía (INEGI). (2019). Micro, pequeña, mediana y gran empresa: Estratificación de los establecimientos. [Micro, small and medium-size and large companies: Stratification of institutions]. Economic Census.

Kilpatrick, D. (2000). Definitions of public policy and the law. National Violence Against Women Prevention Research Center, Medical University of South Carolina. Obtenido de: https://main-web-v.musc.edu/vawprevention/policy/definition.shtml

Kontschieder, V. (2018). Prototype in Policy: What For?! (2018). Obtenido de: https://conferences.law.stanford.edu/prototype-for-policy/2018/10/22/prototype-in-policy-what-for/

Organization for Economic Cooperation and Development (OECD). (s.f.). Transparency and explainability (Principio 1.3). AI Policy Observatory. Obtenido de: https://oecd.ai/dashboards/ai-principles/P7

Organization for Economic Cooperation and Development (OECD). (2019). Recommendation of the Council on Artificial Intelligence. OECD Legal Instruments. Obtenido de: https://legalinstruments.oecd.org/en/instruments/OECD-LEGAL-0449

Reyes, E. (2020). Las empresas mexicanas no saben qué hacer con la Inteligencia Artificial [Mexican companies do not know what to do with Artificial Intelligence]. Expansión. Obtenido de: https://expansion.mx/tecnologia/2020/07/30/las-empresas-mexicanas-no-saben-que-hacer-con-la-inteligencia-artificial

Senate of the United States. (2022). Algorithmic Accountability Act. Obtenido de: https://www.wyden.senate.gov/imo/media/doc/Algorithmic%20Accountability%20Act%20of%202022%20Bill%20Text.pdf

United Nations Educational, Scientific and Cultural Organization (UNESCO). (2022). Recomendación sobre la ética de la inteligencia artificial [Recommendation on ethics of artificial intelligence]. Obtenido de: https://unesdoc.unesco.org/ark:/48223/pf0000381137_spa

White House Office of Science and Technology Policy. (junio 24 2022). Making automated systems work for the American People. The White House. Obtenido de: https://www.whitehouse.gov/ostp/ai-bill-of-rights/

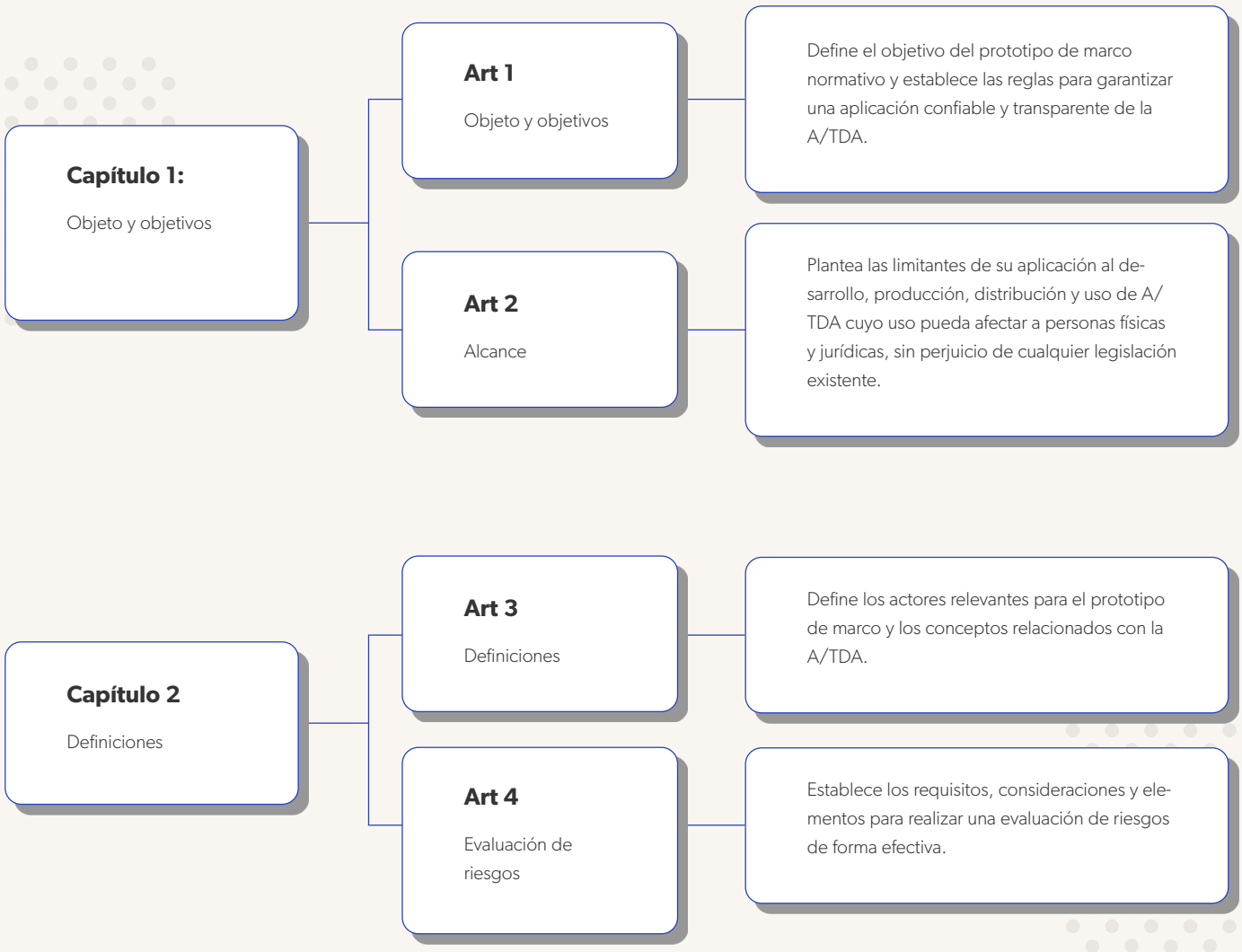
World Economic Forum. (2018). Agile governance – Reimagining policy-making in the Fourth Industrial Revolution. Obtenido de: https://www3.weforum.org/docs/WEF_Agile_Governance_Reimagining_Policy-making_4IR_report.pdf

Annexos

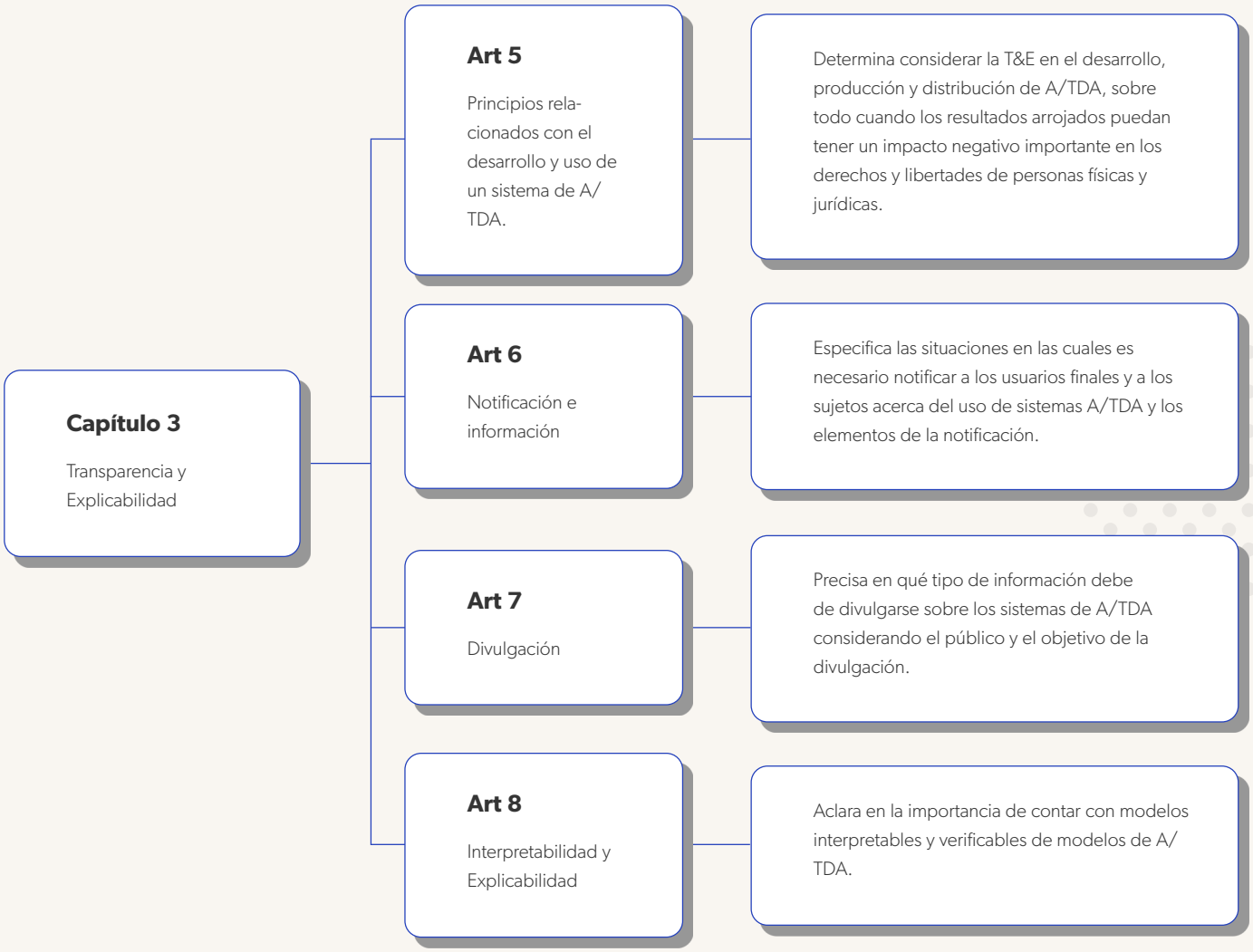
Anexo A. Propuesta de marco normativo para la transparencia y explicabilidad de los sistemas de A/TDM.

A continuación se presenta un resumen de la versión del marco normativo probada por las empresas, comenzando con una visión general del contenido antes de presentar el documento.

Contenido de la propuesta de marco normativo



Continúa en la siguiente página…



La propuesta completa se encuentra a continuación.

A medida que avanza el debate sobre el tema, los autores recomiendan actualizar el marco propuesto para utilizarlo en otro contexto.

NOTA: Este documento se utilizará únicamente para completar el ***Programa de Prototipos de Políticas Públicas sobre transparencia y explicabilidad*** *de la toma de decisión automatizada* ***de Open Loop*** *. El único propósito de este documento es obtener comentarios sobre su contenido y formato por parte de las empresas participantes del programa. Es un documento ficticio desprovisto de toda normatividad vinculante o legal. El Prototipo de Política de ninguna manera reemplaza las leyes y regulaciones existentes que pudieran aplicarse.*

El punto de partida para el Prototipo de Políticas Públicas sobre Transparencia y Explicabilidad de la Toma de Decisión Automatizada (en adelante, "Prototipo de Política" o "Prototipo") es que resulta neutral desde el punto de vista tecnológico (de ahí que en el texto se utilice la toma de decisión automatizada en lugar de IA / AA) y está basado en principios. De esta manera, el Prototipo de Política no se aplica a una tecnología, sector o contexto específico.

Este Prototipo de Política hace una distinción entre la aplicación de la toma de decisión automatizada (TDA o ADM por sus siglas en inglés) que tiene un impacto negativo importante en los dere-

chos y libertades de las personas físicas y jurídicas, y aquella que no. Para la primera categoría, se deben cumplir requisitos adicionales. Además, el Prototipo distingue entre sistemas que incluyen a un ser humano en el circuito (human-in-the-loop) y sistemas que no lo hacen (decisiones totalmente automatizadas). Para esta última categoría se prohíbe su uso en la mayoría de los casos que pudieran suponer un impacto negativo importante para el sujeto (aunque la probabilidad del mismo sea baja). Finalmente, una distinción relevante es la que existe entre modelos interpretables y modelos de caja negra. Cuando un ser humano no está involucrado, y la TDA se usa para propósitos que pueden afectar significativamente los derechos y libertades de las personas físicas, es necesario utilizar un modelo interpretable.

Preámbulos y consideraciones

Objeto y alcance

1 Los acelerados avances tecnológicos, especialmente en lo que respecta a la inteligencia artificial y la TDA, han generado nuevos desafíos para la protección de las personas físicas y jurídicas y los valores y principios asociados a ella. Los sistemas de TDA toman decisiones cada vez más significativas con respecto a las personas físicas y jurídicas, cuando antes estas decisiones eran tomadas por humanos. Algunas de estas decisiones son tomadas por sistemas opacos, inexplicables e incomprensibles para los sujetos afectados. Estas características de los sistemas de TDA pueden conducir a la falta de confianza en las decisiones tomadas por dichos sistemas.

2 Una TDA confiable requiere un marco regulatorio claro, fuerte y coherente, respaldado por una supervisión sólida y eficiente.

3 Este Prototipo establece un deber general de cuidado a todos los actores involucrados en el desarrollo, producción y distribución y uso de sistemas de TDAs para asegurar la transparencia y explicabilidad de la TDA.

4 Cuando la TDA se utiliza para fines que pueden afectar significativamente los derechos y libertades de las personas físicas, el sistema de TDA debe garantizar las decisiones legales, éticas, justas y confiables. Cuando un resultado es defendiblemente injusto o no ético, es necesario proporcionar un razonamiento para la toma de decisión (por ejemplo, medidas de equidad del modelo, métricas de equidad utilizadas); los usuarios deben tener derecho a impugnar. Dada la naturaleza específica de la A/TDA, este Prototipo de Política ayuda a garantizar un nivel coherente de protección contra decisiones dañinas, no éticas, erróneas o ilegales tomadas por estos sistemas. Este Prototipo de Política se entiende sin perjuicio de la legislación vigente sobre decisiones tomadas sobre personas físicas o jurídicas.

Definiciones

5 Este Prototipo de Política se refiere a varios actores en el campo de los sistemas de TDA, en particular desarrolladores, usuarios directos, usuarios finales y sujetos.

6 El desarrollador es la persona física o jurídica que desarrolló el sistema de TDA. Este actor puede sólo proveer el algoritmo de aprendizaje, pero es más probable que este sea la persona u organización responsable de seleccionar los datos (de entrenamiento) y los algoritmos de aprendizaje relevantes, así como la posterior creación y / o entrenamiento del modelo. Cuando el desarrollador y el usuario desarrollen conjuntamente un sistema de TDA, deberán de ponerse de acuerdo respecto a la obligación y la responsabilidad.

7 El usuario directo es la persona física o jurídica que implementa un sistema de TDA para lograr un objetivo particular. Generalmente, se tratará de una organización, como por ejemplo, las autoridades fiscales que detectan fraudes, las plataformas de redes sociales que brindan recomendaciones personalizadas automatizadas o los bancos que evalúan la solvencia de un cliente. El sistema de TDA implementado puede ser un sistema independiente o una parte integral de la entrega de un producto o servicio.

8 El usuario final es la persona física o jurídica que está destinada a utilizar el sistema de TDA, a diferencia de los actores involucrados en el desarrollo o la determinación de su uso. El usuario final sería el actor informado por la decisión del sistema de TDA y / o que tome una decisión basada en el resultado de la decisión automatizada. Por ejemplo, un médico que recibe asesoría sobre un tratamiento por parte de un sistema de TDA. El usuario final puede ser un empleado del usuario o independiente de éste, que utiliza el sistema de TDA como un producto o servicio del usuario.

9 El sujeto es una persona física o jurídica sometida a un sistema de TDA. La decisión que toma este sistema afecta de forma directa o indirectamente al sujeto. Por ejemplo, en el caso del médico (ver 8), sería el paciente el sujeto de la decisión.

10 Si bien cada actor tiene un rol diferenciado, en la práctica estos roles pueden coincidir. Por ejemplo, con los carros autónomos, el conductor podría considerarse tanto el usuario, como el usuario final y el sujeto. Al mismo tiempo, el fabricante de automóviles también puede considerarse el usuario. La designación del rol debe hacerse caso por caso, determinando qué rol está asociado con qué actor y cuáles son las responsabilidades asociadas con éste..

11 Los sistemas de TDA suelen ser demasiado complejos para comprenderlos en su totalidad. Este Prototipo de Política distingue entre transparencia (en el sentido de divulgación e información) y la explicabilidad de los modelos de TDA y las decisiones automatizadas. La explicabilidad de un modelo de toma de decisiones y de sus decisiones depende de la interpretabilidad del modelo. Los modelos simples de toma de decisiones (por ejemplo, regresión lineal simple, árboles de decisión) se consideran "globalmente interpretables": pueden entenderse completamente. Para muchas aplicaciones de aprendizaje automático, el modelo es tan complejo que no se puede interpretar. Cuando se puede interpretar una decisión individual de tal modelo, la decisión se considera "localmente interpretable".

12 El posible impacto negativo importante de una decisión automatizada en los derechos y libertades de las personas físicas y jurídicas se debe juzgar en función del contexto, la naturaleza, el propósito y el alcance de la solicitud. Al evaluar el impacto en las personas físicas y jurídicas, se tendrá en cuenta específicamente el hecho, si aplica, de que no se puede interpretar el modelo y no se pueden explicar las decisiones individuales.

13 Un impacto negativo importante en los derechos y libertades de las personas físicas y jurídicas pueden incluir la pérdida de vidas o lesiones, daños económicos o materiales, daños a la reputación profesional que interfieren con el sustento de una persona e interferencia con derechos fundamentales como el derecho a la igualdad de trato, el derecho a la privacidad, el derecho a la protección de datos personales y el derecho a la libertad de expresión. En el contexto de la TDA, se debe prestar especial atención a los daños económicos, psicológicos y sociales que pueden derivarse del uso de éstos. Deben considerarse como importantes tanto los derechos y libertades individuales como los colectivos. Respecto a esto último, cabe señalar que el impacto negativo en los derechos y libertades a nivel individual podría, por repetición en una colectividad, ser considerado como un impacto negativo importante.

14 El impacto de un sistema de TDA depende del contexto en el que se utiliza. Para cada contexto, quienes implementan sistemas automatizados de toma de decisiones deben evaluar qué daños individuales y colectivos son relevantes para su consideración a nivel individual y / o colectivo. Estos incluyen, entre otros, efectos (legales) que conducen a una pérdida de oportunidades económicas, por ejemplo, aunque no necesariamente, discriminación de precios, discriminación laboral o prácticas comerciales desleales; efectos que conducen a daños psicológicos como la autocensura, la pérdida de la autoestima y la pérdida de la autonomía personal; y daños colectivos como la pérdida de libertad y la inestabilidad económica o política.

Transparencia y explicabilidad

1 Los seres humanos deben ser conscientes de que están interactuando con una máquina, en particular cuando esto no es evidente en la interacción, por ejemplo, a causa de la inteligencia del sistema. Además, los sujetos deben ser conscientes de que la decisión sobre ellos es tomada por una máquina y no un humano, o una combinación de ambos. Por ende, tanto los usuarios finales como los sujetos deben ser informados de la existencia de un sistema de TDA cuando interactúan con dicho sistema o cuando se toman decisiones a través de este. Esta divulgación debe informarles sobre la existencia del sistema de TDA, su propósito, sus capacidades y limitaciones conocidas y el usuario que implementa el sistema. Además, los usuarios y sujetos deben poder hacer valer sus derechos de privacidad y protección de datos cuando sus datos personales se procesan como parte de un procedimiento de toma de decisiones.

2 Para permitir la comprensión del proceso de TDA y la rendición de cuentas, los aspectos relevantes del proceso deben documentarse internamente y posiblemente sea necesario que su importancia relativa deba informarse. Esto incluye, entre otros, los datos de entrenamiento, los algoritmos de aprendizaje y el modelo en sí. Pero también procesos como la selección de datos, la creación de modelos, la validación de modelos, las pruebas y la gobernanza del modelo (por ejemplo, creador, fecha de creación, última capacitación, etc.). El alcance y la profundidad de la divulgación, así como el público relevante para la divulgación, dependen del objetivo de la misma. Los objetivos pueden incluir explicabilidad, responsabilidad y supervisión interna y externa. Para proteger los derechos de propiedad intelectual y otros intereses legítimos del desarrollador y / o usuario, el público al cual se divulga la información puede estar obligado a la confidencialidad.

3 Las decisiones tomadas por sistemas de TDA deben ser, generalmente, comprensibles y verificables. Estos requisitos son obligatorios para los sistemas de TDA que pueden tener un impacto negativo importante en los derechos y libertades de las personas físicas y jurídicas. En estos casos, la regla general es usar solo modelos que puedan ser interpretados completamente por humanos (modelos globales o inherentemente interpretables). Sin embargo, si el usuario puede demostrar que modelos más complejos (por ejemplo, redes neuronales no intrínsecamente profundas) proporcionan una mejor precisión y rendimiento, entonces se puede utilizar un modelo que no sea interpretable de forma global o intrínseca. Esto solo está permitido si las decisiones individuales pueden explicarse con un grado considerable de precisión y si se han tomado otras medidas para reducir el riesgo.

4 Los estándares existentes para la explicación y / o justificación de decisiones deben aplicarse a la TDA. Por ejemplo, si en un veredicto legal requiere que el juez explique cómo se evaluaron las diferentes pruebas y cómo contribuyeron a la decisión final, un sistema de TDA usado en el tribunal debe cumplir el mismo estándar. En otras palabras, si la decisión automatizada está reemplazando una decisión humana, y la decisión humana solía requerir un nivel específico de información, la decisión automatizada deberá requerir el mismo nivel de explicación, si no mayor.

Capítulo 1: objeto y objetivos

Artículo 1: Objeto y objetivos

1.1 Este Prototipo de Política establece reglas para garantizar una aplicación confiable y transparente de la TDA.

Artículo 2: Alcance

2.1 Este Prototipo de Política se aplicará al desarrollo, producción, distribución y uso de sistemas de TDA cuyo uso pueda afectar a personas físicas y jurídicas.

2.2 Este Prototipo de Política se entiende sin perjuicio de cualquier legislación existente, en particular en el área de derechos fundamentales, protección de datos y prácticas comerciales desleales.

Definiciones

Artículo 3: Definiciones

(a) "Actores" significa los desarrolladores, usuarios, usuarios finales, sujetos y cualquier otra parte que contribuya al diseño, desarrollo, producción, distribución, capacitación y / o despliegue de sistemas automatizados de toma de decisiones y / o se vea afectado por dicho sistema o sus decisiones.

(b) "Algoritmo" significa una secuencia finita de instrucciones o un conjunto de reglas diseñadas para completar una tarea o resolver un problema.

(c) "Modelo" significa el resultado de utilizar un algoritmo de aprendizaje automático con datos de entrenamiento específicos. Este modelo es una representación matemática del dominio aprendido y se utiliza para mapear el proceso desde los insumos (*inputs*) hasta los resultados (*outputs*). El modelo es el componente principal de un sistema de TDA y es utilizado por un algoritmo para generar la decisión.

(d) "Decisión totalmente automatizada" significa una decisión tomada por un sistema de TDA que actúa sin ninguna intervención humana significativa.

(e) "Sistema de TDA" significa un proceso computacional que incluye uno derivado del aprendizaje automático, de estadísticas o de otra técnica de procesamiento de datos, que toma una decisión o facilita la toma de decisiones humana.

(f) "Usuario" significa la persona física o jurídica que implementa un sistema de TDA para lograr un objetivo particular.

(g) "Usuario final" significa la persona física o jurídica que utiliza el sistema de TDA para los fines previstos por el usuario.

(h) "Sujeto" significa la persona física o jurídica sometida directa o indirectamente a una decisión de un sistema de TDA.

(i) "Interpretabilidad" significa el nivel de comprensión del proceso de toma de decisiones, en particular la comprensión del modelo utilizado en la TDA.

(j) "Globalmente interpretable" significa que un individuo puede comprender todo el modelo de una vez, entendiendo todas las diferentes decisiones automatizadas que puede tomar el modelo. En este nivel de interpretabilidad se trata de comprender cómo el modelo toma decisiones, basado en una visión holística de sus características y de cada uno de los componentes aprendidos, tales como pesos, otros parámetros y estructuras.

(k) "Localmente interpretable" significa que una decisión automatizada individual puede entenderse o explicarse, precisando cómo un insumo específico llevó a un resultado específico.

(a) El 'impacto negativo importante en los derechos y libertades de las personas físicas y jurídicas' puede incluir daños, pérdida de la vida o lesiones, daños económicos, patrimoniales o psicológicos, e interferir con los derechos fundamentales como el derecho a la igualdad de trato, el derecho a la privacidad, el derecho a la protección de datos personales y el derecho a la libertad de expresión. Deben considerarse como importantes tanto los derechos y libertades individuales como los colectivos. El impacto negativo en los derechos y libertades de las personas, por repetición en una colectividad, podría considerarse un impacto negativo importante.

Artículo 4: Evaluación de riesgos

4.1 Previo al despliegue de un sistema de TDA, el usuario evaluará los riesgos del sistema de TDA previsto y su aplicación sobre los derechos y libertades de las personas físicas y jurídicas.

4.2 Si un sistema de TDA puede tener un impacto negativo importante en los derechos y libertades de las personas físicas o jurídicas, el usuario llevará a cabo una evaluación de impacto del sistema de TDA antes de su despliegue.

4.3 Se exigirá una evaluación de impacto del sistema de TDA, referido en el apartado 4.2, en caso de:

(a) posible sesgo injusto o discriminador hacia los sujetos, incluyendo, entre otros, discriminación de precios, discriminación laboral o acceso diferencial injusto a los servicios;

(b) una posible pérdida de control o agencia del sujeto, incluida la manipulación económica o psicológica;

(c) la aplicación a gran escala de la TDA que pueda afectar a las comunidades o la sociedad en su conjunto;

(d) el procesamiento de datos sistemático y extenso o a gran escala que presente un alto riesgo para los derechos de protección de datos del sujeto, incluida la elaboración de perfiles y el seguimiento sistemático.

4.4 Una evaluación de impacto del sistema de TDA deberá contener al menos:

(a) una descripción detallada del sistema de TDA, su diseño, su capacitación, datos y su propósito;

(b) una evaluación de la calidad, integridad y representatividad de los datos utilizados para entrenar el modelo subyacente;

(c) una evaluación de los riesgos para las personas físicas y jurídicas, con especial atención a los sujetos y usuarios finales; y,

(d) las medidas previstas para abordar los riesgos, incluidas las salvaguardias, las medidas de seguridad, la periodicidad de las revisiones y los mecanismos de protección de los derechos y libertades de los usuarios finales y sujetos y para demostrar el cumplimiento de este Prototipo, teniendo en cuenta los derechos e intereses legítimos de los interesados.

4.5 En aquellos casos en los que la evaluación de impacto de la TDA indique que la aplicación puede generar un alto riesgo para los derechos naturales y las libertades de las personas físicas y jurídicas y que estos riesgos no pueden ser mitigados, el usuario deberá buscar la aprobación de la autoridad supervisora antes de la implementación.

Transparencia y explicabilidad

Artículo 5: Principios relacionados con el desarrollo y uso de un sistema de TDA.

5.1 En el desarrollo, producción, distribución y uso de sistemas de TDA, los actores deberán considerar la transparencia y explicabilidad de sus sistemas de TDA teniendo en cuenta el contexto, alcance, propósito y naturaleza de la aplicación.

5.2 Cuando un sistema de TDA tome decisiones que puedan tener un impacto negativo importante en los derechos y libertades de las personas físicas o jurídicas, el usuario tomará las medidas técnicas y organizativas necesarias para garantizar que el uso del sistema sea transparente y los resultados explicables.

Artículo 6: Notificación e información

6.1 Los usuarios y desarrolladores, cuando corresponda, deberán notificar a los usuarios finales y a los sujetos acerca del uso de sistemas automatizados de toma de decisiones en aquellas instancias en las que:

(a) su uso pueda tener un impacto negativo importante en sus derechos y libertades, o

(b) cuando el sistema de TDA interactúe con el usuario final o sujeto tal como lo haría un ser humano.

6.2 Los usuarios deben proporcionar información significativa a los usuarios finales y a los sujetos sobre:

(a) La finalidad del sistema de TDA y, en su caso, la justificación de su uso en lugar de la toma de decisiones humana,

(b) el posible impacto del sistema de toma de decisiones sobre los derechos y libertades de los usuarios finales y sujetos,

(c) la lógica del proceso de toma de decisiones de acuerdo con los requisitos establecidos en el artículo 7,

(d) su derecho a impugnar decisiones automatizadas y la forma en que estos derechos pueden ejercerse.

6.3 La información descrita en el punto 6.2 será clara, concisa, accesible y de fácil lectura. La información para los usuarios finales puede ser más detallada y adaptada según sea necesario para que facilite significativamente su comprensión al grupo al que se dirige.

Artículo 7: Divulgación

7.1 Teniendo en cuenta el contexto, la naturaleza y el alcance de la aplicación y los riesgos que puede plantear el sistema de toma de decisiones automatizado, los desarrolladores y usuarios de sistemas de toma de decisiones automatizados deberán divulgar los elementos relevantes al desarrollo, operación y uso de dichos sistemas.

7.2 Los elementos relevantes mencionados en el punto 7.1 incluyen, pero no se limitan a:

(a) el fundamento para la automatización de la toma de decisión,

(b) los datos de entrenamiento, prueba y validación,

(c) los algoritmos utilizados,

(d) el modelo de toma de decisiones,

(e) el proceso de selección y preparación de datos,

(f) el proceso de formación, selección, validación y prueba del modelo,

(g) el proceso de gestión y mantenimiento del modelo en funcionamiento.

7.3 La divulgación de los elementos relevantes de un sistema de TDA debe tener en cuenta al público, los medios de comunicación y los objetivos de la divulgación. Esto podría ser, entre otras cosas:

(a) el tema, para dar una idea de la lógica del proceso de toma de decisiones;

(b) las unidades internas de la organización responsables de la gestión de riesgos y cumplimiento y auditoría interna para el ejercicio de sus funciones;

(c) los auditores externos para fines de auditoría o verificación de terceros;

(d) las Autoridades Supervisoras con fines de cumplimiento, investigaciones y fiscalización general.

Artículo 8: Interpretabilidad y explicabilidad

8.1 Las decisiones de un sistema de TDA deberán ser interpretables y verificables, teniendo en cuenta la naturaleza, el alcance, el contexto y el propósito del sistema de TDA.

8.2 Cuando las decisiones totalmente automatizadas puedan causar un impacto negativo importante en los derechos y libertades de las personas físicas y jurídicas sometidas al proceso de toma de decisiones o se vean afectadas por él, se debe utilizar un modelo globalmente interpretable.

8.3 El punto 8.2 no se aplicará cuando:

(a) El usuario puede argumentar que un modelo no globalmente interpretable es estrictamente necesario para el propósito de la TDA; sólo en aquellos casos en los que el usuario pueda demostrar que modelos más complejos (por ejemplo, redes neuronales profundas) proporcionarán una mejor precisión y rendimiento, se podrá utilizar un modelo no globalmente interpretable Sin embargo, esto únicamente está permitido si las decisiones individuales pueden explicarse con un grado significativo de precisión, y si se han implementado otras medidas para reducir el riesgo.

(b) El usuario puede proporcionar explicaciones para decisiones individuales con un grado suficiente de precisión, por ejemplo, utilizando métodos de interpretación locales;

(c) El usuario ha tomado las medidas técnicas y organizativas necesarias para proteger los derechos e intereses del sujeto.

8.4 Los usuarios finales y los sujetos recibirán explicaciones de las decisiones individuales en un formato claro, conciso, accesible y comprensible.

Anexo B. Propuesta de manual práctico para la adopción de los principios de transparencia y explicabilidad de los sistemas de IA/ADM

El manual podría complementar cualquier legislación futura y ser la base de instrumentos de corregulación y derecho indicativo: códigos de conducta, códigos de práctica, normas, certificaciones, pautas industriales, etc.

En esta sección establecemos formas de cumplir con el prototipo de política pública propuesta. Implementando los elementos del manual, una organización está en una buena posición para cumplir con los requisitos del prototipo.

Transparencia y explicabilidad de la AI : ¿por qué hacerlo?

Con base en la evaluación de la literatura especializada en IA, podemos concluir que no existen definiciones universalmente aceptadas sobre los requisitos como transparencia, interpretabilidad, auditabilidad, explicabilidad, comprensibilidad y trazabilidad, y muchas veces son utilizadas de diferentes maneras. Además, en la mayoría de los documentos analizados, no está claramente definido cuál es el propósito específico de un requisito en un contexto determinado.

Vale la pena establecer (bajo un contexto determinado) cómo deben interpretarse los requisitos de transparencia y explicabilidad. Las siguientes preguntas deben considerarse a la hora de decidir cómo se deben cumplir los requisitos legales y éticos de transparencia, interpretabilidad, etc:

- ¿Deberían restringirse ciertos usos previstos para el uso de la IA? ¿Por qué?
- ¿Cuál es el potencial problema causado por la TDA opaca / inescrutable?
- En un contexto dado, ¿cuál es el impacto de este potencial problema?
- ¿Este impacto se siente a nivel individual, colectivo o ambos?
- ¿Cómo reducirá el impacto la comprensión del proceso automatizado de toma de decisiones a través de la transparencia / interpretabilidad, etc.? Es decir, ¿qué objetivos estamos tratando de lograr aumentando la comprensión del proceso de toma de decisiones?
- ¿Cuál es el requisito para lograr nuestro objetivo subyacente?
- ¿Existen opciones alternativas que sean más efectivas para lograr estos objetivos?
- ¿Cuál es la mejor forma de darse cuenta / comprender / considerar al público relevante?
- ¿Cómo explicarle al usuario final o al sujeto la manera en que funciona la decisión automatizada en un contexto dado?

Capturamos estas preguntas y sus posibles respuestas en la Tabla 1.

¿Qué problema potencial estás intentando resolver con tu explicación?

Potencial problema	Objetivo	Requerimiento	Soluciones posibles	Foco en transparencia/explicabilidad	Público objetivo principal
Se desconoce si se está utilizando un sistema automatizado de toma de decisiones o se desconoce qué hace.	Notificar e informar a los usuarios, usuarios finales y sujetos del uso de un sistema de TDA (TDA)	Notificar a los usuarios, usuarios finales y sujetos que están utilizando una TDA o que están sujetos a decisiones tomadas por un TDA y proporcionarles información (de alto nivel) sobre su funcionamiento.	Declaraciones de información, notificaciones y advertencias	Sistema TDA en su totalidad	Usuarios, usuarios finales, sujetos, autoridad supervisora, sociedad en general
No está claro si un modelo está funcionando correctamente (en la medida en que es posible saberlo) y si toma decisiones precisas / justas / oportunas / proporcionadas en un entorno real.	Tener en cuenta el funcionamiento correcto del modelo (es decir, en términos de equidad de seguridad, etc.)	Proporcionar información sobre el funcionamiento de un modelo (transparencia, interpretabilidad, auditabilidad, trazabilidad)	Evaluación de modelos, interpretabilidad global, interpretabilidad local.	Datos de entrenamiento, algoritmos de aprendizaje, modelo entrenado, resultados del modelo, datos de entrada	Usuario, usuario final, sujeto, autoridad de control, auditores, sociedad en general
No está claro cómo se tomó una decisión en particular.	Mejorar la comprensión del alcance de la TDA y las razones detrás de una decisión en particular.	Explicar por qué se tomó una decisión en particular en un caso particular (explicabilidad)	Interpretabilidad global, interpretabilidad local	Datos de entrada, resultados del modelo	Usuario, usuario final, sujeto, autoridad supervisora, auditor, sociedad en general
No está claro si la decisión que se tomó cumple con los requisitos (legales o de otro tipo) para una decisión justificada.	Proporcionar una justificación para una decisión en particular.	Proporcionar una explicación de la decisión y prueba de que la decisión se tomó siguiendo el procedimiento / estándar adecuado (legal o de otro tipo) para este proceso de toma de decisiones (explicabilidad, trazabilidad)	Evaluación de modelos, interpretabilidad global, interpretabilidad local.	Datos de entrenamiento, algoritmos de aprendizaje, modelo entrenado, resultados del modelo, datos de entrada	Usuario, usuario final, sujeto, autoridad de control, sociedad en general

Continúa en la siguiente página...

Potencial problema	Objetivo	Requerimiento	Soluciones posibles	Foco en transparencia/explicabilidad	Público objetivo principal
Se desconocen las razones de una toma de decisión en particular, lo que dificulta la impugnación o el cuestionamiento de la decisión.	Ayudar a impugnar una decisión	Comprender por qué se tomó una decisión en particular en un caso particular (explicabilidad)	Interpretabilidad local	Datos de entrada, resultados del modelo	Sujeto
Se desconocen las razones de una toma de decisión en particular, lo que dificulta determinar cómo cambiar el resultado de la decisión en el futuro.	Alterar el comportamiento futuro para recibir potencialmente un resultado preferible o deseado	Comprender por qué se tomó una decisión en particular (explicabilidad)	Interpretabilidad local, explicaciones contrafácticas	Datos de entrada, resultados del modelo	Sujeto
No está claro cómo funciona todo el sistema de inteligencia artificial y por qué los usuarios deberían de confiar en él.	Que la gente comprenda que un producto o experiencia en particular está impulsado por IA y cómo funciona esa IA	Confianza / aceptación y dependencia de la sociedad: este tipo de explicación está diseñada para generar confianza y aceptación por parte de la sociedad. Por ejemplo, si el sistema proporciona un resultado inesperado , la explicación puede ayudar a los usuarios a comprender por qué se generó este resultado . También puede proporcionar una mayor sensación de confianza en el sistema si se puede proporcionar la justificación.	Explicabilidad a nivel de modelo y sistema, así como explicabilidad a nivel de proceso para educar acerca de cómo se realizan los análisis internos de costo-beneficio.	Un sistema TDA que incluye entradas del modelo, resultados finales, impactos potenciales	Usuarios finales, sujetos, autoridad de control, sociedad en general

Tabla 1 Comprender la TDA: una guía para la explicabilidad basada en el problema que estás resolviendo

Otra forma de abordar estas preguntas y llegar a las respuestas correspondientes es contextualizar a la transparencia y explicabilidad a sus componentes fundamentales: público, contexto, propósito y contenido.

Elementos de la transparencia y explicabilidad

Público Objetivo	Contexto	Propósito	Contenido
Regulador	Investigación	Concientizar	Justificación
Socio comercial	Relación entre empresas	Comprender las características del producto o servicio	Responsabilidad
Consumidor	Reclamación / queja	Permitir la retroalimentación	Seguridad y rendimiento
Sociedad	USO general	Involucrar a los usuarios	Datos y modelo
Propia empresa		Habilitar recurso/ renuncia	Equidad (*fairness*)
		Modificar el comportamiento futuro	Impacto
		Rendir cuentas sobre su operación correcta	

Tabla 2. Elementos de la transparencia y explicabilidad

Publico objetivo: los destinatarios / a quién va dirigida la explicación.

Contexto: las razones por las que una explicación es (o no) importante y se está solicitando.

Propósito: qué está tratando de lograr la explicación; el objetivo y / o la motivación principal por la que se proporciona una explicación de la IA.

Contenido: la información proporcionada a los destinatarios de la explicación; en qué información en particular compartir con el público.

Requisitos del prototipo de política pública

La tabla 1 demuestra, a través de los potenciales problemas y objetivos descritos, diferentes requisitos y posibles soluciones. Con base en esto, se han descrito en el Prototipo de Política Pública tres obligaciones legales y éticas separadas pero relacionadas entre sí: **1)** notificación e información (artículo 6), **2)** interpretabilidad y explicabilidad (artículo 8) y **3)** divulgación (artículo 7).

Notificación e información (artículo 6)

El Prototipo de Política Política requiere que los usuarios notifiquen a los usuarios, usuarios finales y sujetos sobre el uso de un sistema de toma de decisión automatizada en dos situaciones específicas:

- Cuando el uso del sistema de TDA puede tener un impacto significativo en los derechos y libertades o percepciones (por ejemplo, efecto de cámara de eco, distorsiones de la realidad);

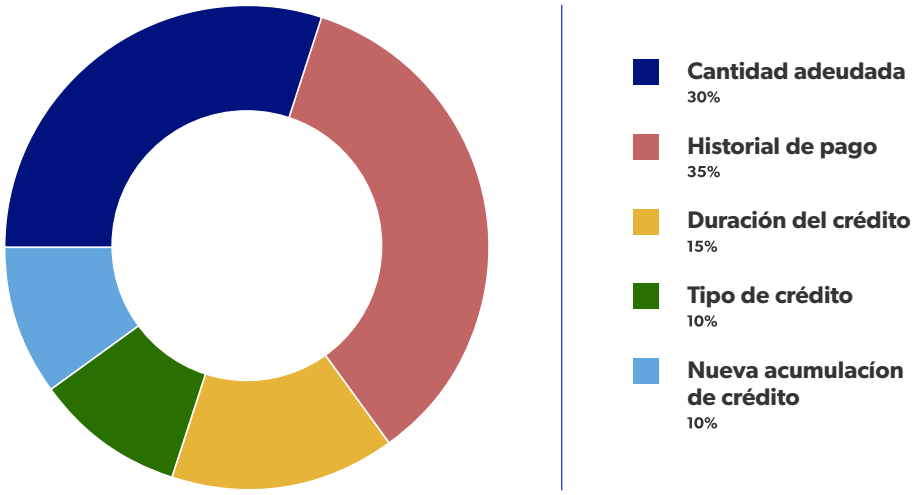
- Cuando el sistema de TDA interactúa con el usuario final o sujeto como lo haría un ser humano.

Los usuarios (y, dependiendo del caso, los desarrolladores) pueden cumplir con la obligación de notificación de diversas formas. Por ejemplo, proporcionando un texto pequeño o un logotipo (generalmente aceptado) para señalar la existencia de una TDA (TDA).

El contenido del requisito de información depende del contexto, y se relaciona con el requisito de divulgación del artículo 7 y el requisito de explicabilidad del artículo 8. El Prototipo de Política Pública establece cuatro requisitos. Para cumplir con el requisito de información no es necesaria en todos los casos la divulgación total de los algoritmos, etcétera, a los sujetos, ya que no les ayudará a comprender el proceso de la TDA. Más bien, el sujeto debe comprender por qué se están tomando decisiones automatizadas, cómo esto puede afectarlo, cuál es la lógica de la TDA, qué datos se utilizan y cómo se pueden impugnar, ser excluido, o estar sujeto a la aplicación de otros derechos. Para la lógica de la toma de decisiones, es importante que el sujeto pueda entender cómo se toma una decisión y cómo su situación particular es impactada por el algoritmo / modelo.

La explicación dada por *Dupaco Credit Union* sobre cómo se calcula un puntaje crediticio puede servir como un buen ejemplo de características que podrían informar una decisión automatizada.[1] *Dupaco* explica.

1. Cómo se construye una puntuación de la siguiente manera:



Cantidad adeudada
El porcentaje de límites de crédito disponibles

Historial de pago
Se le da más importancia a la forma en que se pagan las cuentas: ¿los pagos se realizan a tiempo o están atrasados?

Duración del crédito
La cantidad de tiempo que se ha establecido una línea de crédito.

Tipo de crédito
Manejar responsablemente una combinación de crédito. Los préstamos en cuotas aumentan las calificaciones, mientras que los préstamos renovables tienden a bajar las calificaciones

Nueva acumulación de crédito
Número de consultas de crédito y fechas de apertura de nuevas líneas en los últimos 12-18 meses

[1] https://www.dupaco.com/learn/master-your-money/credit-credit-scores/#1550677639908-d7dfc9db-5fae

Qué significa en la práctica tener un puntaje particular (impacto / consecuencias):

Puntaje	Unión de crédito
720 y mayor	Si obtienes 720 o más y tendrás buenas posibilidades de obtener préstamos con las mejores tasas de interés. Estos préstamos pueden requerir menos documentación y papeleo, y potencialmente menos (o incluso ningún) pago inicial o garantía.
680–720	Si calificas en este rango, por lo general podrás negociar buenos términos.
620–680	Si te encuentras en este rango te aplicarán las reglas "estándar" de la compañía, lo cual te dará menos flexibilidad para elegir mejores préstamos o servicios.
580–620	Serás evaluado con ojo crítico y necesitarás factores compensatorios para que la mayoría de las empresas aprueben préstamos o servicios.
Debajo de 500	No es una buena calificación. Por lo general, se te pedirá que proporciones un pago inicial / garantía sustancial y / o pagues una tasa de interés más alta.

Otra opción es utilizar "etiquetas de información nutricional". El proyecto *Data Nutrition* proporciona un prototipo en su sitio web: https://datanutrition.org/.

Interpretabilidad y explicabilidad (artículo 8)

El artículo 8 del Prototipo de Política Pública se enfoca en interpretabilidad y explicabilidad de la TDA.

Comencemos por definir cada uno de estos conceptos:

- La **interpretabilidad** es la medida en que uno puede predecir lo que va a suceder, dado un cambio en los parámetros de entrada o algorítmicos. O, para decirlo de otra manera, se trata de poder distinguir los cambios de una decisión cuando la entrada cambia sin saber necesariamente por qué.

- La **explicabilidad**, a su vez, es la medida en que la mecánica interna de un sistema de decisión automatizado se puede explicar en términos humanos. La diferencia con la interpretabilidad es muy sutil, pero considéralo así: la interpretabilidad se trata de ser capaz de discernir la mecánica sin saber necesariamente por qué. La explicabilidad es poder explicar literalmente lo que está sucediendo.[2]

Este artículo tiene como objetivo específico abordar el "problema de la caja negra" en la TDA. La mayor crítica con la TDA de alto impacto basada en IA / Aprendizaje Automático (AA) es que los

resultados suelen ser imposibles de explicar y verificar porque el modelo no se puede interpretar debido a su inherente complejidad. Por lo tanto, los usuarios deberían considerar primero, según el contexto y su aplicación, si se pueden justificar los modelos complejos de aprendizaje automático. Sabiendo que la forma más eficaz de evitar los retos inherentes a los modelos de caja negra es no utilizarlos, a continuación se profundiza en el tema de interpretabilidad.

Interpretabilidad global (o modelos globalmente interpretables)

Para las decisiones que requieren plena responsabilidad y justificación, generalmente son preferibles los modelos que son globalmente interpretables. El gran inconveniente es que el proceso de TDA para modelos globalmente interpretables se limita al uso de sistemas basados en reglas y modelos predictivos simples. Esto puede implicar, en ciertos casos, encontrar un balance entre la interpretabilidad, por un lado, y la precisión, eficacia y eficiencia, por el otro. Sin embargo, para ciertos tipos de decisiones (por ejemplo, para determinar si alguien es culpable de un delito o seleccionar automáticamente un tratamiento médico), la interpretabilidad global puede ser un requisito obligatorio.

En la literatura, estas "líneas rojas" aún no están claras a pesar de que los requisitos establecidos por diferentes cuerpos legisladores (por ejemplo, el Congreso de los Estados Unidos, la Comisión Europea y el Consejo de Europa) pueden implicar prohibiciones de este tipo en la práctica. Además, en el contexto del derecho administrativo y el derecho penal, se considera un principio general el hecho que las decisiones que no puedan explicarse y / o no vayan acompañadas de una justificación son inválidas.

Interpretabilidad local

La principal crítica y desafío de los modelos de caja negra es que son difíciles –si no imposibles– de comprender para los humanos. El campo naciente de la IA explicable (xIA) se ocupa de la explicabilidad de la TDA, más específicamente la de los modelos de caja negra. Este campo de investigación se está expandiendo rápidamente y es importante mantenerse al día con los diversos enfoques que se están desarrollando y presentando para que los modelos de caja negra sean cada vez más explicables.

Para obtener una descripción general extensa (aunque no exhaustiva) de estos enfoques, consulta el Anexo 1 "Transparencia y explicabilidad de la IA: orientación técnica". Según nuestra revisión de la literatura al respecto, existen diferentes métodos para lograr el objetivo de la explicabilidad, como por ejemplo la creación de un modelo proxy que se comporte de manera similar al modelo original, pero de una manera que sea más fácil de explicar, mediante la creación de un mapa de prominencia (*salience map*) para resaltar la parte del cálculo que sea más relevante, o mediante la extracción automática de reglas. Lo que estos métodos tienen en común es que proporcionan interpretabilidad local; es decir, no explican todo el modelo, sino cómo se llegó a una determinada conclusión. A continuación se presentan y describen varios modelos para lograr el objetivo de la explicabilidad:

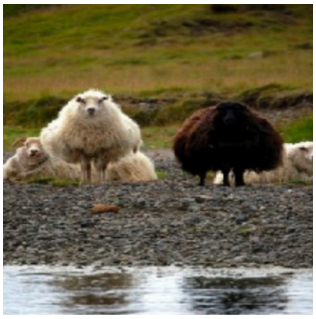
Modelos proxy

Los modelos proxy, también llamados modelos sustitutos locales, son modelos interpretables que se utilizan para explicar predicciones individuales de modelos de AA de caja negra (Molnar, 2019).

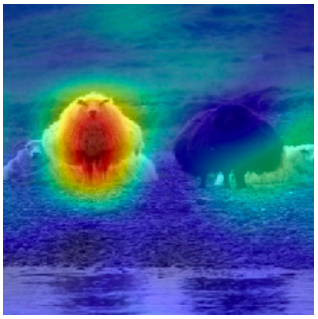
El ejemplo más conocido de este enfoque de modelo proxy es LIME (Explicación agnóstica del Modelo Interpretable Localmente) (Ribeiro, et al., Sin fecha). LIME genera un nuevo conjunto de datos que consta de muestras permutadas y las predicciones correspondientes del modelo de caja negra. En este nuevo conjunto de datos, LIME entrena un modelo interpretable, que se pondera por la proximidad entre las instancias de muestreo y la instancia de interés (Molnar, 2019). También se pueden utilizar los árboles de decisión para aproximar redes neuronales. Finalmente, SHAP (Explicaciones aditivas de Shapley) también proporciona una explicación local para el resultado de un modelo (Lundberg & Lee, 2017).

Mapeo de prominencia

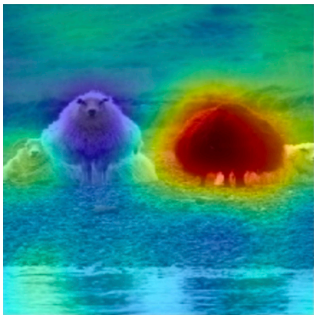
En el campo del reconocimiento de imágenes, el concepto de mapa de prominencia contribuye a una mejor comprensión. Para una clasificación dada de una imagen, se puede trazar un mapa de prominencia para mostrar qué partes distintivas de la imagen se usaron para llegar a una conclusión particular. Esto ayuda a los usuarios a comprender qué partes de la imagen fueron más relevantes para la clasificación. Un ejemplo de este enfoque es RISE (muestreo de entrada aleatorio para la explicación de modelos de caja negra) (Petsiuk, et al., 2018).



(a) Borrego 26%, Vaca 17%

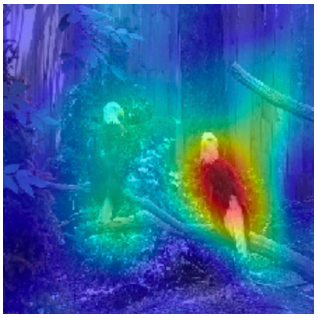


(b) Mapa de importancia de "borrego"

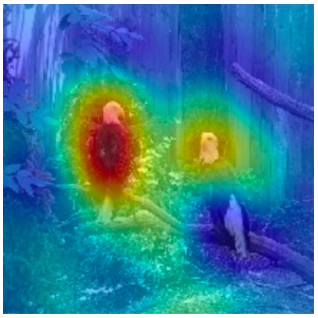


(c) Mapa de importancia de "vaca"

(d) Pájaro- 100%, Persona- 39%



(e) Mapa de importancia de "pájaro"



(f) Mapa de importancia de "persona"

Extracción automática de reglas

La extracción automática de reglas es un método que, usando una red neuronal entrenada y los datos sobre los que fue entrenada, produce una descripción de la hipótesis de la red. Ésta última es comprensible y se aproxima mucho al comportamiento predictivo de la red (Craven, 1996; Gilpin, et al., 2018). Mediante la extracción de reglas se construye un modelo más simple que imita el comportamiento de la red neuronal compleja y profunda (Jacobsson & Ziemke, 2005). Las técnicas de

extracción de reglas se pueden clasificar en **1)** descomposicionales **2)** pedagógicas y **3)** eclécticas. **1)** Las técnicas de descomposición implican analizar los pesos entre las unidades y la función de activación ("mira dentro" de la red) para extraer reglas. **2)** Las técnicas pedagógicas tratan a la red como una "caja negra" y extraen reglas examinando la relación entre las entradas y las salidas. **3)** Los enfoques eclécticos incorporan técnicas de descomposición y pedagógicas juntas (Biswas, et al., 2017). Un ejemplo de técnica de extracción automática de reglas es el algoritmo TREPAN, el cual deriva un árbol de decisiones de una red neuronal (Craven y Shavlik, 1996).

Explicaciones contrafácticas

Wachter y col. (2017) argumentan que una explicación para una decisión automatizada suele ser insuficiente, en particular para los sujetos de un proceso de toma de decisiones. Además, argumentan que la interpretabilidad local puede ser difícil de implementar en la práctica. Proponen una forma novedosa de aumentar la comprensión y la confianza en la TDA: explicaciones contrafácticas. Una explicación contrafactual le da al sujeto una visión de cómo su situación tendría que ser diferente para que ocurra un resultado deseable. Son posibles múltiples contrafactuales, ya que pueden existir múltiples resultados deseables (Wachter, et al., 2017). Una decisión contrafactual podría verse así:

> *"Te han negado un préstamo porque tus ingresos (30 mil euros) son demasiado bajos. Si tus ingresos hubieran sido de 50,000 euros, tu préstamo habría sido aprobado."*

Este enfoque proporciona a los interesados explicaciones valiosas para comprender una decisión determinada, motivos para impugnarla y consejos sobre cómo el interesado puede cambiar su comportamiento o situación para posiblemente recibir la decisión deseada (por ejemplo, aprobación de préstamo).

Divulgación (artículo 7)

Finalmente, se requiere la divulgación efectiva de los elementos de un sistema de TDA y los procesos y procedimientos para crear el sistema. Los elementos relevantes mencionados en el Prototipo de Política Pública incluyen, entre otros:

- el fundamento del uso de TDAs,
- los datos de entrenamiento, esto puede incluir la descripción, origen y consentimiento cuando corresponda, tipo (personal, sensible), nivel de actualización, mecanismos de seguridad para su protección y resguardo de dichos datos.
- los algoritmos de AA utilizados,
- el modelo de toma de decisiones,
- el proceso de selección y preparación de datos,
- el proceso de entrenamiento, selección, validación y prueba del modelo
- el proceso de gestión del modelo en funcionamiento.

Cuando se trata de divulgación, es de particular importancia el objetivo de la divulgación y el público al que se divulga el contenido. Por lo tanto, para cumplir con el requisito del artículo 8, un usuario debe establecer primero qué se puede esperar de él en términos de divulgación. La tabla 1 y la tabla 2 pueden usarse para evaluar cuál es el requisito de divulgación y cuál es el público relevante.

En términos generales, la divulgación sólo es relevante para los sistemas de TDA que tienen un impacto negativo importante en los derechos y libertades de las personas físicas y jurídicas. En estas áreas en particular, el usuario debe ser capaz de rendir cuentas de los resultados. Un tercero (por ejemplo, un auditor o una autoridad supervisora) puede desear evaluar el funcionamiento del TDA y los procesos para desarrollarlo, implementarlo y mantenerlo. En otras palabras, un operador del sistema debe considerar que un tercero evalúe la operación los casos que puedan tener un impacto negativo importante en los derechos y libertades de las personas físicas y jurídicas. Por lo tanto, el requisito de divulgación está estrechamente relacionado con los requisitos en el área de auditabilidad.

Proceso de transparencia y explicabilidad

Con base en nuestro análisis de la literatura sobre transparencia y explicabilidad, sugerimos el siguiente proceso para abordar los requisitos relacionados con mejorar la comprensión de la toma de decisiones.

Paso 1: Determina el riesgo del proceso de toma de decisiones.

Con base en los resultados de su evaluación de riesgos, determina cuáles riesgos plantea el proceso de toma de decisiones para los valores individuales y / o colectivos.

Paso 2: Determina si la interpretabilidad global es un requisito

Con base en esta evaluación, determina qué nivel de información, divulgación e interpretabilidad se requiere.

Si el proceso de toma de decisiones tiene un impacto negativo importante y requiere una justificación sólida (por ejemplo, porque afecta la posición jurídica del sujeto), considera si se requiere un modelo globalmente interpretable.

Paso 3: Determina los objetivos de comprensión y el público objetivo asociado

Determina cuál es el objetivo de la provisión y divulgación de información. En función de los diferentes objetivos, determina a qué público objetivo debe dirigirse (consulta la tabla 2 para conocer los tipos de público).

Paso 4: Determina la divulgación adecuada por público objetivo

Determina qué forma de divulgación o explicabilidad es necesaria y deseable para cada público objetivo.

Cuando se trata de interpretabilidad y explicabilidad, adapta la explicación y el mensaje al público objetivo y su capacidad para comprender y evaluar la información que les proporciona. Ten en cuenta factores como la experiencia y la limitación de tiempo. A continuación se comparten preguntas que te pueden ayudar en crear una buena explicación:

- **Identificar y estudiar al público**

  Es fundamental responder a la pregunta "¿A quién afecta directamente el producto o servicio basado en IA?" para que el material de comunicación y apoyo diseñado y construido sea legible, comprensible y personalizado para el público objetivo.

  De manera similar, se deben tener en cuenta las diferencias normativas / legales y del contexto humano de cada geografía, especialmente para un producto o servicio global, para evitar una comunicación genérica y vacía.

- **Definir el nivel de detalle para cada público**

  Cada grupo tiene un nivel de alfabetización técnica y digital diferente según su nivel educativo o el interés personal. Por lo tanto, es importante definir la cantidad de información y la complejidad que se presentará al usuario final. Es importante acordar la información que se va a comunicar (datos utilizados, información del modelo, nivel de participación humana, situaciones en las que se usa la IA, impacto de la decisión o predicción de la IA, entre otros) y cómo se comunicará esa información; que a su vez debe estar alineado con los propósitos de comunicación y el uso ético de los principios de la IA.

Paso 5: Implementa las medidas técnicas y organizativas para aumentar la comprensión

Finalmente, implementa las medidas técnicas y organizativas necesarias. Elige métodos de interpretabilidad y explicabilidad (ver Anexo 1) basados en el impacto de la toma de decisión, el tipo y nivel de transparencia / explicabilidad requerida, y el público relevante como se describe en este documento.

Consulta la tabla 1 y la tabla 2 para determinar cuáles son los problemas potenciales y el público en el que te debes enfocar, junto con los diferentes tipos de contexto, propósito y contenido subyacentes al requisito de transparencia y explicabilidad.

Sugerencias para complementar e ir más allá de la transparencia y la explicabilidad, hacia una IA ética

Permear la práctica de la ética en el uso de sistemas basados en IA

Implementa talleres, cursos y capacitaciones sobre el uso ético de la IA en las diferentes áreas de la empresa o institución involucradas en el diseño, desarrollo, implementación, mantenimiento, soporte, mejora o cualquier actividad relacionada con el producto o servicio que dependa del uso de la IA. Todo el equipo, de forma transversal, debe estar informado sobre el estado del arte del desarrollo y uso responsable de la IA, así como las prácticas recomendadas en materia de transparencia y explicabilidad. Además, es importante que te asegures que las herramientas cumplan con los estándares éticos, para que las mismas herramientas de desarrollo que utilizan los ingenieros contribuyan a cumplir las prácticas recomendadas.

Preservar una cultura basada en el uso ético de la IA

Este proceso no termina una vez que se completan los pasos descritos en este manual; es un proceso continuo que nos permite preservar una cultura que fomenta los principios éticos en el uso de la IA, que también responde a los cambios en la tecnología, las mejoras o cambios en el producto o servicio, los nuevos objetivos comerciales, el nuevo usuario y las expectativas del mercado, así como los cambios organizativos, las nuevas leyes y normativas y los cambios sociales y culturales.

Recursos adicionales

Sección 1

Guías existentes

- Explaining decisions made with AI - UK's Information Commissioner's Office, 2019
- Companion to the Model AI Governance Framework: Implementation and Self-Assessment Guide for Organizations – World Economic Forum, Info-communications Media Development Authority of Singapore(2020)

Sección 2

Cuando tus sistemas de IA son proporcionados por terceros

Cuando la organización utiliza tecnología de IA proporcionada por terceros, se recomienda obtener del proveedor un modelo de explicabilidad; las empresas que ofrecen plataformas de aprendizaje automático (como Facebook, Google, IBM, Amazon y Microsoft) están comenzando a incluir herramientas de explicabilidad.

Sección 3

Herramienta Recomendada

- "People-centric Approaches to Notice, Consent, and Disclosure" de TTC Labs y Singapore Infocomm Media Development Authority.

Herramientas útiles para la gobernanza

- "Explaining Decisions with AI. Part 3: What explaining AI means for your organisation" UK's Information Commissioner's Office (2019)

Cursos recomendados

- Big Data, Artificial Intelligence and Ethics, Universidad de California, Davis.
- Ethics and Big Data, The Linux Foundation.

Evaluación de impacto y riesgo

Esta debe incluir, entre otros:

- describir la naturaleza, alcance, contexto y propósitos del procesamiento:
- evaluar las medidas de necesidad, proporcionalidad y cumplimiento;
- identificar y evaluar el riesgo para las personas; e
- identificar cualquier medida adicional para mitigar esos riesgos.

Recomendación para cumplir con el principio de rendición de cuentas

- **Desarrollo de mecanismos y procesos de impugnación y revisión de decisiones.**

  Es esencial, especialmente en un sistema con un humano involucrado (*human-in-the-loop*), que existan procesos y herramientas para que los usuarios finales que se sientan insatisfechos con la decisión tomada por el sistema de IA puedan pasar por un proceso de impugnación o solicitar que una decisión se someta a revisión humana para auditar el proceso. En caso de ser un resultado con un impacto negativo importante en la vida de una persona, se recomienda que la decisión final no se automatice.

- **Comunicar los mecanismos de revisión humana y cuestionar las decisiones.**

  Los usuarios deben tener claro cómo pueden impugnar las decisiones tomadas por IA en caso de que éstas los pongan en desventaja y ellos crean que es un error. Estos mecanismos siempre deben ser fácilmente accesibles, visibles y presentes, para que el usuario pueda decidir entre cuestionar una decisión, solicitar una explicación o bien una revisión humana.

- **Abrir canales de retroalimentación**

  Por último, ningún ejercicio de diseño centrado en el usuario estaría completo sin un proceso de retroalimentación continuo. Los usuarios deben contar con mecanismos claros y siempre disponibles para emitir comentarios, sugerencias o quejas sobre el desempeño o servicio brindado por el sistema basado en IA.

Metodologías de análisis de impacto existentes

- Data Protection Impact Assessments (DPIAs), UK's Information Commissioner's Office
- Human rights impact assessments (HRIAs)
- Human Rights Due Diligence, United Nations
- "AI Impact Assessment: A Policy Prototyping Experiment" (2021), en https://openloop.org/wp-content/uploads/2021/01/ AI_Impact_Assessment_A_Policy_Prototyping_Experiment.pdf
- ICO: AI and Data Protection Risk Mitigation and Management Toolkit

Guías técnicas recomendadas (además del Anexo 1: "AI Transparency and Explainability: Technical Guidance")

"Individual Explanations in Machine Learning Models: A Survey" (Alfredo Carrillo, Luis F. Cantú y Alejandro Noriega, 2020).

En los últimos años, ha estado aumentando el uso de modelos estadísticos sofisticados que influyen en las decisiones en ámbitos de gran relevancia social. En las aplicaciones del mundo real, principalmente en dominios en los que las decisiones pueden tener un impacto social, es necesario interpretar estos modelos. Este estudio revisa los métodos más relevantes y novedosos para abordar los problemas de explicación de casos individuales en el AA. En particular, pretende proporcionar una guía para los científicos de datos sobre la investigación de métodos apropiados para resolver la necesidad de modelos de explicabilidad.

"Individual Explanations in Machine Learning Models: Case Study" with Prosperia.ai (Alfredo Carrillo, Luis F. Cantú y Alejandro Noriega, 2020).

El caso de estudio tiene dos objetivos principales. Primero, exponer los desafíos que enfrentan los equipos técnicos en el mundo real y cómo utilizan métodos de explicación relevantes y novedosos. En segundo lugar, presentar un conjunto de estrategias que mitiguen dichos desafíos.

Técnicas de explicabilidad:

Estrategias y herramientas tales como modelos sustitutos, diagramas de dependencia parcial, importancia / interacción de variables globales, análisis de sensibilidad, explicaciones contrafácticas o sistemas autoexplicativos y basados en la atención son ejemplos de técnicas de explicabilidad. Sumar documentación sobre cómo se construyó, entrenó y probó el sistema de IA ayudará a mejorar tu explicabilidad. El equipo técnico también podría considerar la construcción de modelos predictivos que imiten condiciones reales o entrenar versiones más simples del modelo (por ejemplo, regresiones lineales o árboles de decisión en lugar de una red neuronal) para simplificar la explicación. Además, algunas de estas opciones también podrían resultar útiles:

a. Usar visualizaciones para explicar predicciones o decisiones individuales.

b. Explicar las características o pesos de cada entrada.

c. Tener en cuenta las investigaciones y métodos desarrollados por terceros en campos académicos y científicos que permitan el análisis de modelos de IA desde múltiples perspectivas: por ejemplo por alcance, global o local; por generalidad, agnóstico del modelo o específico del modelo; o por nivel de interpretabilidad, post-hoc o ad-hoc (específicamente, métodos intrínsecamente interpretables).

Técnicas de diseño (interfaces de usuario y experiencias).

Es importante recordar que el usuario final siempre tendrá interacción con la IA a través de algún tipo de interfaz (ya sea gráfica, basada en audio o de cualquier otro tipo), por lo que es importante

crear una experiencia de usuario que no solo sea fácil de usar e intuitiva, pero que también comunique de manera adecuada y elegante toda la información, los mecanismos y los procesos que conlleva el uso de la IA. Por ello, se recomienda utilizar metodologías de co-diseño y diseño centrado en el usuario para maximizar la calidad del resultado final, así como seguir estos principios de diseño:

a. La transparencia no es lo opuesto a la simplicidad: es posible comunicarse con detalle siempre que se haga en el momento y lugar adecuados, es decir, de forma progresiva a lo largo del recorrido de uso del sistema de TDA.

b. Una organización debe conocer los diferentes niveles de conocimientos técnicos sobre la TDA y adaptarse en consecuencia. Cuando los usuarios entregan datos, deben percibirlos como un intercambio de valor para que se comprenda la relación entre la entrada de sus datos y los resultados de esa entrada de datos.

c. La confianza se construye siendo proactivo, contextual y transparente, lo que permite a las personas comprender en todo momento cómo se utilizan sus datos, cuáles son sus opciones de privacidad y cómo y dónde actualizarlos.

d. La transparencia por sí sola no es suficiente, debe ir acompañada de 'controles' (cuando corresponda) que permitan a los usuarios finales tomar decisiones sobre la TDA, el uso de sus datos y darles la opción de, según el nivel de impacto de la TDA en cuestión: impugnar, apelar u optar por no participar en la TDA.

Recomendaciones adicionales:

Comunicar y controlar el reentrenamiento de la IA.

Cuando el usuario final proporciona nuevos datos o realiza acciones que alimentan un proceso de entrenamiento constante para la IA, es fundamental comunicarse con los usuarios finales: por un lado, desde el punto de vista del uso de datos; por otro, desde el punto de vista de la seguridad de la IA, es decir, mantener el sistema de IA a salvo (en la medida de lo posible) de usuarios y contenidos malintencionados. Debido a esto, las organizaciones deben considerar cubrir estos temas en sus términos y condiciones de uso.

Una de las tareas clave de este proceso es convertir la complejidad técnica que implica comprender cómo funciona un sistema basado en IA, y lograr una comunicación clara y concisa. Se recomienda que los contenidos utilizados para esto incluyan los beneficios claros y específicos que obtiene el usuario final al usar un sistema compatible con IA, así como las posibles consecuencias negativas.

Posibles tensiones que pueden surgir por la implementación de este manual.

La implementación de prácticas de transparencia y explicabilidad (T&E) también implica un proceso continuo para equilibrar sus oportunidades y riesgos (también llamados tensiones) para la organización, por ejemplo:

1. T&E vs. permitir que las personas malintencionadas actúen de manera más efectiva, manipulando el sistema y manipulando algoritmos para sus propios fines.

2. T&E vs. efectividad / precisión: los métodos modernos de IA, especialmente el aprendizaje profundo no interpretable de manera inherente, pueden, en algunos casos, volverse más efectivos pero al mismo tiempo más difíciles de entender.

3. T&E vs. divulgación de posibles problemas de propiedad intelectual: la transparencia total de los algoritmos, es decir, la divulgación del código fuente, plantea importantes problemas legales desde la perspectiva de la propiedad intelectual y el secreto comercial, al igual que la divulgación de otros tipos de información patentada (por ejemplo, software, patentes).

4. T&E vs. significado / comprensión real: una explicación demasiado detallada puede ser incomprensible para los usuarios y no contribuya a su comprensión de la manera en que se manejan sus datos y / o cómo se toman las decisiones / recomendaciones / predicciones.

5. T&E vs. múltiples actores: las obligaciones de divulgación son muy diferentes para los desarrolladores que para los usuarios.

6. T&E vs protección de datos: la IA moderna puede actuar en contra de algunos de los principios de protección de datos (minimización, especificación del propósito, etc.).

Notas finales

1. El término toma de decisiones automatizada A/TDA es entendido y utilizado de distintas maneras por ingenieros, ingenieras y legisladores. Además, el término A/TDA se utiliza de distintas maneras según la legislación específica en la que se emplee. Mientras que las y los ingenieros utilizan el término en sentido amplio, describiéndolo como un sistema que aprovecha el aprendizaje automático para producir un resultado (decisión), las y los formuladores de políticas públicas y reguladores utilizan el término de una manera más granular, limitándolo a tipos específicos de resultados (decisiones) que observan criterios estrechamente definidos. El concepto jurídico restringido de A/TDA se encuentra sobre todo en la normativa sobre privacidad. La interpretación jurídica más amplia de la A/TDA se utiliza a veces en la nlegislación que en otros lugares se denominaría "legislación sobre IA". Nota: Para efectos de este ejercicio de creación de prototipos de políticas, y en lo que respecta a las pruebas reales del prototipo, el término A/TDA se utiliza en el sentido más amplio, desvinculado de cualquier criterio de calificación. El objetivo era la operacionalización de practicas concretas de los principios de transparencia y explicabilidad y no en el tipo específico, la relevancia o el impacto de las decisiones producidas o respaldadas por los sistemas de A/TDA.

2. Véanse, por ejemplo, las bases para una Estrategia Nacional de IA en México desarrollada por C Minds, Oxford Insights y la Embajada Británica (Inglés - Español); la Agenda Nacional Mexicana de IA desarrollada por IA2030Mx (Inglés - Español), entre otros recursos.

3. Por ejemplo, recomendar a las y los desarrolladores de IA que faciliten documentación sobre el funcionamiento de sus sistemas, entre otras prácticas de TyE.

4. Véase, por ejemplo, el Marco de Gestión de Riesgos (RMF por sus siglas en inglés) del NIST creado por el Centro de Recursos de Seguridad Informática el cual "proporciona un proceso completo, flexible, reproducible y cuantificable de 7 pasos que cualquier organización puede utilizar para gestionar la seguridad de la información y el riesgo para la privacidad de organizaciones y sistemas". Más información en: https://www.nist.gov/itl/ai-risk-management-framework

5. Reyes, E. (2020). Las empresas mexicanas no saben qué hacer con la Inteligencia Artificial. Expansión. Obtenido de https://expansion.mx/tecnologia/2020/07/30/las-empresas-mexicanas-no-saben-que-hacer-con-la-inteligencia-artificial

6. Grupo independiente de expertos de alto nivel sobre IA. (2019). Ethics guidelines for trustworthy AI. [Directrices éticas para una IA confiable. Comisión Europea]. Obtenido de https://digital-strategy.ec.europa.eu/en/library/ethics-guidelines-trustworthy-ai

7. Organización para la Cooperación y el Desarrollo Económicos (OCDE). (2019). Recommendation of the Council on Artificial Intelligence [Recomendación del Consejo sobre Inteligencia Artificial]. OCDE. Instrumentos Legales. Obtenido de https://legalinstruments.oecd.org/en/instruments/OECD-LEGAL-0449

8. Parlamento Europeo. (2023). PROPUESTA de enmiendas de transacción sobre el proyecto de la Propuesta de Reglamento del Parlamento Europeo y del Consejo sobre normas armonizadas relativas a la Inteligencia Artificial (Ley sobre Inteligencia Artificial) y por el que se modifican determinados actos legislativos de la Unión. Parlamento Europeo. Obtenido de https://www.europarl.europa.eu/resources/library/media/20230516RES90302/20230516RES90302.pdf

9. Instituto de Ingenieros Eléctricos y Electrónicos . (2022). La Iniciativa Global sobre Ética en Sistemas Autónomos e Inteligentes. Obtenido de https://standards.ieee.org/industry-connections/ec/autonomoussystems/

10. La Organización de las Naciones Unidas para la Educación, la Ciencia y la Cultura (UNESCO). (2022). Recommendation on ethics of artificial intelligence.[Recomendación sobre la ética de la inteligencia artificial]. Obtenido de https://unesdoc.unesco.org/ark:/48223/pf0000381137_spa

11. Supra nota 6

12. OCDE. Recommendation of the Council on Artificial Intelligence. (2019). https://legalinstruments.oecd.org/en/instruments/OECD-LEGAL-0449

13. Kilpatrick, D. (2000). Definitions of public policy and the law. [Definiciones de política pública y Derecho]. Centro Nacional de Investigación para la Prevención de la Violencia contra las Mujeres, Universidad de Medicina de Carolina del Sur. Obtenido de https://mainweb-v.musc.edu/vawprevention/policy/definition.shtm

14. Foro Económico Mundial (2018). Agile governance – Reimagining policy-making in the Fourth Industrial Revolution. Obtenido de https://www3.weforum.org/docs/WEF_Agile_Governance_Reimagining_Policy-making_4IR_report.pd

15. Design Thinking es un proceso utilizado para resolver problemas priorizando las necesidades de las personas consumidoras. Utiliza pruebas que determinan cómo se relacionan las personas consumidoras con los productos o servicios, así como un enfoque que se pueda reproducir y práctico para crear soluciones innovadoras. Para más información, visite: https://designthinking.ideo.com/

16. Brown, T., & Katz, B. (2011). Change by design. Journal of Product Innovation Management, 28(3), 381-383.; Villa Alvarez, D., Auricchio, V., & Mortati, M. (2020). Design prototype for policymaking. Obtenido de: https://dl.designresearchsociety.org/cgi/viewcontent.cgi?article=1168&context=drsconference-papers; Kontschieder, V. (2018). Prototype in policy: What for?! (2018). Obtenido de https://conferences.law.stanford.edu/prototype-for-policy/2018/10/22/prototype-in-policy-what-for/

17. Buchanan, C. (2018). Prototype for policy. UK Government. Obtenido de https://openpolicy.blog.gov.uk/2018/11/27/prototype-for-policy/

18. Hébert, M. (2019). A pilot is not a prototype: How to test policy ideas before scaling. Apolitical. (2019). Obtenido de https://apolitical.co/solution-articles/en/a-pilot-is-not-a-prototype-how-to-test-policyideas-before-scaling; también véase https://openloop.org/reports/2022/12/Experimental_governance_emerging_technologies_Chapter1.pdf

19. Villa Alvarez, D., Valentina A., y Marzia M. "Design prototyping for policymaking." (2020) https://dl.designresearchsociety.org/cgi/viewcontent.cgi?article=1168&context=drs-conference-papers

20. En este caso, medimos la claridad ¿hasta qué punto entienden las empresas participantes los requisitos establecidos en el prototipo?, la eficiencia (¿hasta qué punto contribuye el prototipo a alcanzar el objetivo político general?) y la viabilidad (¿hasta qué punto los beneficios compensan los costos de alcanzar los objetivos del prototipo de política pública?) Para más información sobre estas definiciones, visite: https://openloop.org/wp-content/uploads/2021/01/AI_Impact_Assessment_A_Policy_Prototype_Experiment.pdf

21. Una coalición multisectorial integrada por personas expertas, instituciones académicas, empresas, startups, organismos públicos y otros actores del ecosistema digital y de IA en México con el propósito de desarrollar un marco adecuado para promover recomendaciones y mejores prácticas de IA. VéaseHernández, M. (2019). Estrategia Nacional de Inteligencia Artificial va por sentido ético y responsable Forbes México. Obtenido de https://www.forbes.com.mx/estrategia-nacional-de-inteligencia-artificial-va-por-sentido-etico-y-responsable/

22. Previo a la construcción de una Agenda Integral e Incluyente de Inteligencia Artificial para México, la Coalición IA2030Mx lanzó una Encuesta Nacional de Inteligencia Artificial para conocer mejor los principales retos de la transformación digital en México y las diferentes propuestas a favor de la revolución de la IA en beneficio de todos y todas. Para más información, visite https://www.ia2030.mx/consulta

23. Supra nota 6

24. Los datos de entrenamiento son los que se utilizan para entrenar un modelo. En los sistemas de *Machine Learning*, los algoritmos aprenden a partir de los datos que se les suministran.

25. Fjeld, J., Achten, N., Hilligoss, H., Nagy, A., & Srikumar, M. (2020). Principled artificial intelligence: Mapping consensus in ethical and rights-based approaches to principles for AI. [Inteligencia artificial basada en principios: Consenso sobre los enfoques éticos y basados en los derechos de los principios de la IA]. Centro de Investigación Berkman Klein. Publicación No. 2020-1. Obtenido de http://dx.doi.org/10.2139/ssrn.3518482

26. Grupo Ad Hoc de Expertos (AHEG por sus siglas en inglés) para la preparación de una propuesta de texto de recomendación sobre la ética de la inteligencia artificial. (2020). Documento final: First draft of the recommendation on the ethics of artificial intelligence. UNESCO. Obtenido de https://unesdoc.unesco.org/ark:/48223/pf0000373434

27. Organización para la Cooperación y el Desarrollo Económicos (OCDE), (s.f.) Transparencia y explicabilidad (Principio 1.3). Observatorio de Políticas de IA. Obtenido de https://oecd.ai/dashboards/ai-principles/P7

28. Véase IA interpretable. (2022, septiembre). ¿Qué es la interpretabilidad?

29. Véase también TTC Labs. People-Centric Approaches to Algorithmic Explainability. Obtenido de https://www.ttclabs.net/report/people-centric-approaches-to-algorithmic-explainability

30. Dscout es una plataforma etnográfica móvil de investigación cualitativa que puede utilizarse para obtener información sobre las experiencias de las personas usuarias. Para más información, visite: https://help.dscout.com/hc/en-us/articles/360038171372-dscout-Overview-for-Researchers/.

31. Para la selección del público y el contexto, sólo se podía elegir una opción, pero para propósito y contenido se podía elegir más de una.

32. Entre las mayores empresas consultadas, las principales razones compartidas para no participar en el programa fueron: 1) la considerable inversión de recursos y tiempo que requeriría, y 2) la falta de una normativa vigente que obligara a reforzar el T&E de sus sistemas de A/TDA.

33. Instituto Nacional de Estadística y Geografía (INEGI). (2019). Micro, pequeña, mediana y gran empresa: Estratificación de los establecimientos. Censo Económico, p. 20.

34. Supra nota 2

35. Supra nota 3

36. Supra nota 4