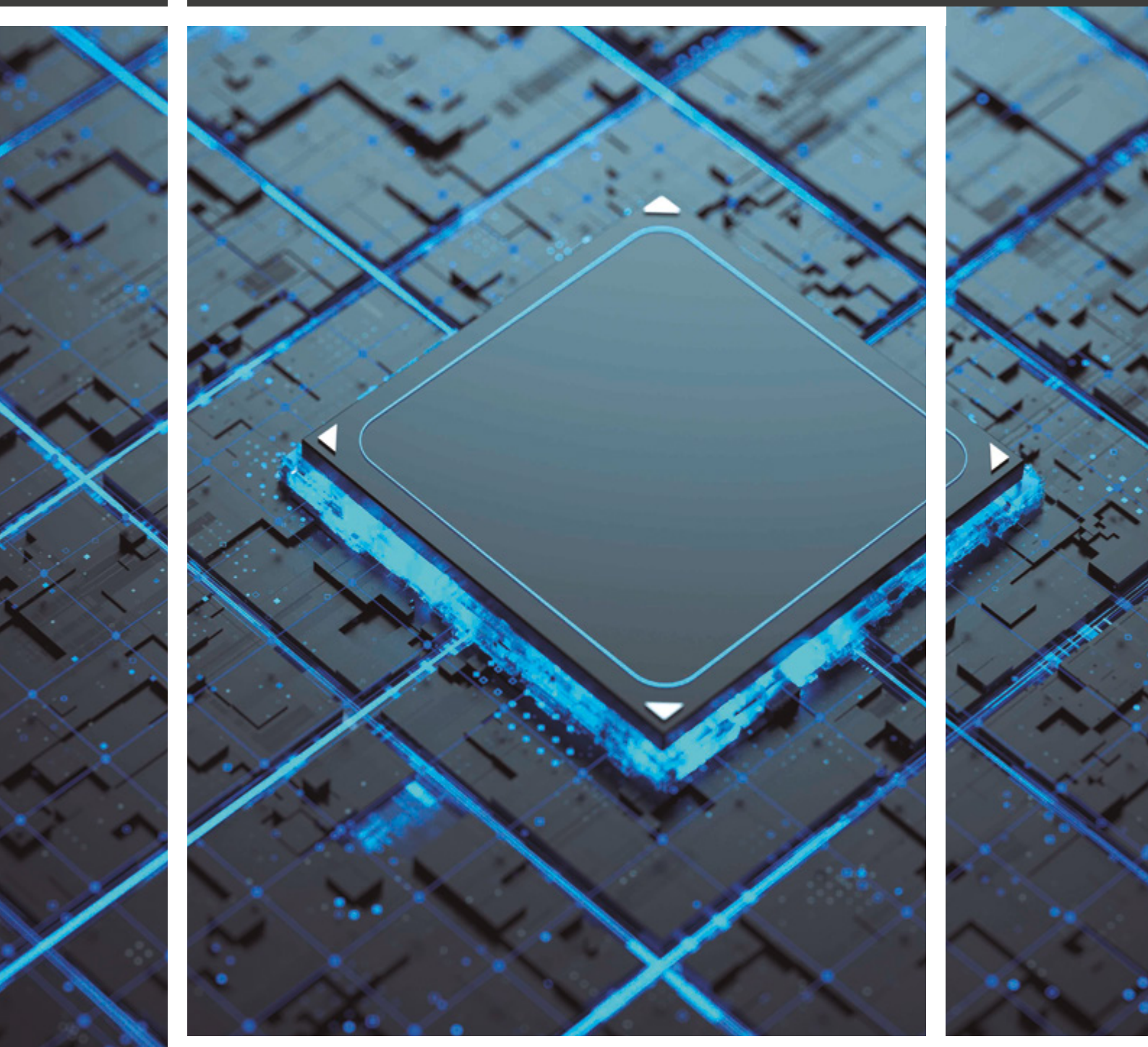
GuIA de buenas prácticas en el uso de la inteligencia artificial ética

OdiseIA. Observatorio del Impacto Social y Ético de la Inteligencia Artificial

pwc

GUIA
UNA INICIATIVA DE ODISEIA

ODISEIA



Contenido

0

Prólogo

Juan Manuel Belloto Gómez.
Director de OdiseIA responsable de GuIA

Contexto

Tras la eclosión de internet a finales de los 90, hemos asimilado tecnologías que han cambiado por completo nuestra forma de entender empresa y sociedad.

La inteligencia artificial (IA) es una tecnología con muchas más décadas de desarrollo conceptual (aproximadamente desde los años 60), que ya está preparada para entrar en juego gracias a la cantidad ingente de datos disponibles y a la capacidad para procesarlos.

La IA es una tecnología diferencial respecto de esas que hemos asimilado en las últimas décadas porque tiene la capacidad de aprender por si misma, hacer predicciones, puede tomar decisiones, y emular capacidades cognitivas del ser humano.

Por primera vez tenemos a nuestra disposición una tecnología que tiene similitudes a las capacidades del ser humano. Es por tanto natural pensar que para su correcta adopción sea necesario tener en cuenta los principios éticos (por las implicaciones de sus capacidades) como aspectos regulatorios (por los riesgos que puedan derivarse de su uso). De igual manera que dichos principios y normativas son necesarias para las personas en una sociedad de bienestar.

Cuando hablamos de ética en la IA no solo hablamos del enorme potencial que tiene su uso para hacer bien en la sociedad (*AI For Good*). Este potencial lo tiene cualquier tecnología o cualquier otra herramienta, producto o servicio disponible en la sociedad.

Cuando hablamos de IA ética también hablamos de competencias que normalmente se requieren a un ser humano y a las empresas en su desempeño diario profesional como, por ejemplo:

- Que no esté sujeta a estereotipos sociales, como la discriminación de minorías, razas o géneros.
- Que sea transparente, y que pueda explicar sus razonamientos.
- Que pueda asegurar la seguridad física de los trabajadores.
- Que sea segura y no sea engañada en su aprendizaje y de esta manera pierda fiabilidad.
- Que gestione de manera correcta la privacidad de los datos de sus clientes y empleados.

La ética por tanto es necesaria en la IA de igual manera que lo es en las personas. Las empresas, al igual que tienen sus políticas de comportamiento para sus empleados, tienen la responsabilidad de que sus soluciones inteligentes estén desarrolladas bajo principios éticos.

Pero, al igual que ocurre en la sociedad, la ética es condición necesaria pero no suficiente. Es necesario acompañarla de legislación. Y por primera vez desde la revolución digital en la que nos hallamos inmersos en este siglo, la legislación no se está dejando esperar. Un claro síntoma de su importancia en la IA.

Por ejemplo, en Diciembre de 2020, la Secretaría de Estado de Digitalización e Inteligencia Artificial (SEDIA) publicó la Estrategia Nacional de Inteligencia Artificial (ENIA). Dicha estrategia dedica por entero uno de sus seis ejes a la inteligencia artificial ética y normativa.

En Abril de 2021, la Comisión Europea presentó su propuesta de regulación sobre inteligencia artificial (AI Act), un proyecto de ley que pretende establecer una regulación sobre la IA, y a la que las empresas deberán comenzar a ajustarse en menos de dos años.

En resumen:

- Las empresas tienen la responsabilidad de que su inteligencia artificial esté desarrollada bajo principios éticos.
- Ya están en marcha marcos normativos que instarán a las empresas a cumplir pautas en este ámbito. Es fundamental que las empresas estén preparadas para cuando entren en vigor sus recomendaciones y obligaciones.
- Ética y normativa se retroalimentan.

El estado del arte de la IA ética

Existen muchas investigaciones alrededor de la IA ética. Nosotros hemos identificado más de 150 en todo el mundo y analizado con detalle 27 de ellas. En un ejercicio que nos ha permitido generar un marco ético-normativo integrado con otro tecnológico y que se describe en el presente documento.

La conclusión a día de hoy (Febrero de 2022), sigue siendo la misma que hizo la Universidad de Harvard en su informe *"Principled Artificial Intelligence: Mapping Consensus in Ethical and Rights-based Approaches to Principles for AI"*.

> Existe una amplia brecha y complicada articulación entre conceptos de alto nivel y su aplicación en el mundo real

Es por tanto necesario profundizar en los planteamientos que permitan a las empresas asimilar la inteligencia artificial ética y normativa de manera pragmática en su día a día. **Aterrizar** los conceptos éticos mediante recomendaciones, guidelines y modelos de gobierno que ayuden a las empresas a saber **cómo hacerlo**. Incluso **aterrizar** los conceptos éticos mediante tecnologías que permiten automatizar su gestión.

Por este motivo, en Noviembre de 2020 desde OdiseIA lanzamos la iniciativa GuIA

La historia de GuIA

GuIA es una iniciativa de OdiseIA que fue lanzada públicamente el 27 de Noviembre de 2020, en un evento *on-line* que contó con el apoyo y la visión estratégica de la SEDIA de la mano de su responsable, Carme Artigas. En el evento además participaron responsables de IA ética de Google, Microsoft, IBM y Telefónica.

El evento fue una declaración de intenciones y objetivos, a la que siguieron seis meses para definir la estrategia para conseguirlos, dotando además a la iniciativa de estructura conceptual y operativa.

Tuvimos la suerte de contar con el interés de varias compañías que quisieron acompañarnos en esta aventura, señal indicativa del interés y confianza en OdiseIA en general y por la iniciativa GuIA en particular. Y señal de lo más importante: la necesidad de su objetivo.

En Junio de 2021 comenzamos su *delivery* de la mano de PwC España, quienes unas semanas antes nos presentaron una excelente propuesta técnica, como el resto de las presentadas. Además, nos propusieron realizar el proyecto de manera 100% pro-bono, un claro ejemplo del compromiso de PwC por tener impacto positivo en empresa y sociedad.

Cuál es el objetivo de GuIA

El objetivo de GuIA es **generar un ecosistema** donde cualquier entidad se pueda integrar para compartir y conocer **las mejores prácticas en inteligencia artificial** atendiendo a principios éticos y preceptos normativos.

De manera **verticalizada** a los casos de uso de cada sector empresarial.

Aterrizando los principios éticos y su normativa asociada con las tecnologías y Modelos de Gobierno que permiten su gestión (Visión 360).

En todo el ciclo de vida de las soluciones Inteligentes (visión end-to-end).

Una iniciativa que pretende aportar su granito de arena a varias de las acciones estratégicas del eje 6 de la ENIA, dedicado por entero a la inteligencia artificial ética y normativa.

El *output* de GuIA no es solo un informe. Es una iniciativa que tiene tres pilares estratégicos para conseguir su objetivo:

- **Investigación**. Cuyo primer resultado es el presente informe.
- **Divulgación y formación**. En escuelas de negocio y Universidades, para que los profesionales de hoy y del futuro conozcan con detalle la importancia de una IA ética. Una forma más de aterrizar la IA ética, consiguiendo que los conceptos permeen.
- **Colaboración**. El objetivo final, el de generar comunidad alrededor de la IA ética para que las entidades que no hayan participado inicialmente en el proceso de Investigación puedan consultar las buenas prácticas definidas y aportar las suyas propias.

El plan de acción de GuIA

Hemos comenzado la iniciativa con empresas tecnológicas como Google, Microsoft e IBM porque ellas son el primer eslabón de la cadena, las creadoras de muchos de los productos y servicios inteligentes que después son utilizados en las empresas directamente o a través de soluciones creadas internamente y/o por terceros integradores.

Estas compañías llevan realizando desde hace varios años un notable esfuerzo por ofrecer tecnologías, herramientas y recomendaciones alrededor de la IA ética.

Además, hemos contado con la experiencia de Telefónica, una compañía pionera en la adopción de la inteligencia artificial en general e inteligencia artificial ética en particular.

Este documento recoge sus planteamientos de manera resumida, ordenada y homogeneizada, para facilitar al lector el entendimiento de los mismos. Es una GuIA válida para todos los sectores empresariales.

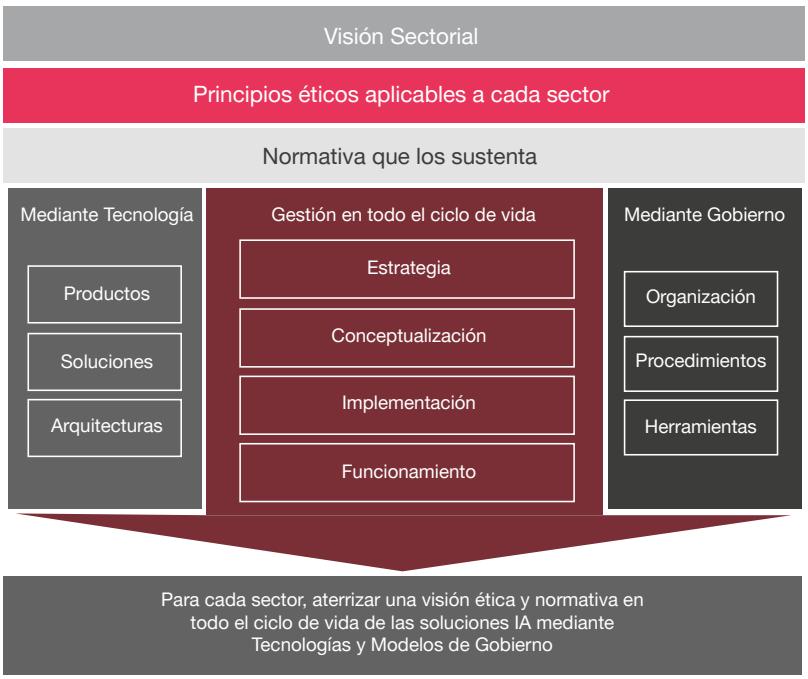
Un contenido lo suficientemente modular como para poder enlazarlo directamente con los casos de uso de una Fase 2 que ya hemos comenzado y que detallamos al final de este documento. Una fase en la que adaptaremos GuIA a 10 sectores empresariales de la mano de más de 50 empresas. Cerraremos así el círculo y aterrizaremos también los conceptos de IA ética desde un punto de vista de negocio.

Qué contiene este documento

El objetivo de este documento GuIA de buenas prácticas para una inteligencia artificial ética y normativa es mostrar a las empresas **de manera aterrizada a cómo adoptar** esta tecnología emergente atendiendo a los principios éticos y su normativa asociada.

No es un informe de análisis de estado con consideraciones globales. El análisis de estado ha sido una herramienta que hemos utilizado para destilar un conjunto de 8 principios éticos, los más utilizados de entre 27 iniciativas de todo el mundo (seleccionadas de entre 150). Un marco ético al que después hemos dotado de un marco jurídico, para finalmente aterrizarlos con las herramientas y tecnologías que permiten la gestión de los principios éticos y su normativa asociada.

Es nuestro marco ético-jurídico-tecnológico, nuestro ***FrameWork* GuIA**.



En definitiva, **un aterrizaje** desde el principio y su normativa asociada hasta, en muchos casos, el algoritmo necesario para implementarlo, quedándonos a las puertas del *coding*.

A quien va dirigido este documento

Hasta hace relativamente poco las áreas tecnológicas eran transcriptores de las necesidades de negocio, ya que sus necesidades se soportaban en tecnologías de gestión con muchos años de recorrido. Las áreas de negocio sabían lo que la tecnología podía ofrecer, y las áreas de IT no necesitaban explicar nuevas grandes capacidades de la misma.

La aparición de nuevas tecnologías en la última década ha provocado la necesidad de comunicación fluida y constante entre las áreas de negocio y las de tecnología, para entender rápidamente la manera en la que aprovechar las capacidades de las mismas. Solo así se consigue una solución que aporte valor de negocio y buen *time to market* que proporcione una ventaja competitiva de mercado. Las áreas de negocio y las tecnológicas evolucionaron así su relación de cliente-proveedor a una relación de *partnership*, trabajando de manera mucho más estrecha.

En paralelo, la eclosión de tecnologías digitales donde la experiencia de usuario es fundamental (Web, *Mobile, Conversational*, Realidades extendidas, etc.) provocó que a los perfiles de negocio y tecnológicos se uniesen nuevos compañeros. A su mesa se unieron perfiles necesarios para la creación de estas nuevas soluciones digitales: diseñadores, sociólogos, psicólogos, filósofos, lingüistas, etc., etc. La mesa cada vez es más diversa.

Con la inteligencia artificial, a todos estos perfiles se unen al menos dos perfiles más: el eticista y el jurista. Con esta tecnología lo normal será verlos juntos buscando un entendimiento para llegar a un lenguaje común comprensible por todos. La multidisciplinariedad es cada vez más necesaria, y todos tienen que saber al menos un poco de la disciplina de su compañero.

Este informe GuIA de buenas prácticas para una inteligencia artificial ética y normativa está dirigido a todos ellos. La GuIA Llega a un detalle importante, pero con la intención de que sus contenidos sean en gran parte entendibles por todos. Esperamos haberlo conseguido, al menos en parte.

Agradecimientos

Gracias a PwC España por acompañarnos en esta iniciativa prestando además sus servicios de manera pro-bono.

Gracias a Google, Microsoft, IBM y Telefónica por su colaboración. Ellos entienden perfectamente la necesidad de una IA ética, y de la oportunidad que esta tecnología puede suponer para nuestro país. Por tanto, parte de su misión es común a la de OdiseIA, y es lo que les hace sentarse en una misma de trabajo pese a tener muchos negocios solapados. Muchas gracias por ello.

Gracias a PwC España, Google y Microsoft por patrocinar la iniciativa GuIA.

Gracias a la Secretaría de Estado de Digitalización e Inteligencia Artificial (SEDIA) por su apoyo.

Gracias a las más de 30 personas que han puesto directamente su granito de arena en la elaboración de esta GuIA. Este grupo de personas, cuyo nombre y apellidos están al final de este documento, son la semilla de una comunidad que deseamos crezca exponencialmente en los próximos tiempos.

Gracias anticipadas a ti lector, por estudiarlo y llevarlo a la práctica.

Esperamos que hagas tuya la misión de OdiseIA, y que ponemos a disposición de tod@s:

> Trabajar activamente por un uso ético y responsable de una inteligencia artificial que tenga en su centro el bienestar del ser humano

Juan Manuel Belloto Gómez.

Director de OdiseIA responsable de GuIA.

PD: Gracias a Idoia Salazar y Richard Benjamins. Sin ellos, GuIA no hubiera sido posible.

1

Introducción

- Qué es la IA
- Qué es la IA ética
- La importancia de la normativa. Cómo se complementa con la ética
- El papel fundamental de la SEDIA en España

1. Introducción

Qué es la IA

La inteligencia artificial, de ahora en adelante IA, es un concepto ampliamente mencionado, comentado y debatido en nuestra sociedad. Y ese hecho no es baladí, pues la importancia que tiene en nuestras vidas esta tecnología, o ciencia como la definirán algunos, impacta de muchas más formas de lo que la sociedad, en promedio, puede alcanzar a comprender.

John McCarthy introduce la expresión "inteligencia artificial" en 1956 durante la conferencia de Dartmouth, siendo el precursor de esta y uno de sus padres, definiéndola como "la ciencia y la ingeniería de crear máquinas inteligentes".

Una definición más reciente, realizada por el *High-Level Expert Group on Artificial Intelligence* (en adelante, AI – HLEG), creado por la Comisión Europea en 2018 en el marco de Estrategia Europea en inteligencia artificial con el objetivo de establecer las bases de una IA fiable, definen los sistemas de IA, en su "Orientaciones Éticas (1) para una IA fiable" publicadas en 2019, como "sistemas que manifiestan un comportamiento inteligente, pues son capaces de analizar su entorno y pasar a la acción —con cierto grado de autonomía— con el fin de alcanzar objetivos específicos".

La IA tiene el potencial de alterar muchos sectores de la economía y, en definitiva, de aumentar la eficiencia con la que se hacen las cosas y mejorar los procesos de toma de decisiones, analizando y aprovechando el potencial del *big data (2)*, pudiendo también conducir a la creación de nuevos servicios, productos, mercados e industrias, impulsando así la demanda de los consumidores, generando nuevas fuentes de ingresos y mejorando el bienestar de la sociedad a través de nuevas soluciones a sus problemas.

Sin embargo, las aplicaciones de la IA también pueden plantear retos y preocupaciones, por ejemplo, en relación con la privacidad, la responsabilidad, la transparencia y la rendición de cuentas, por citar algunos. Así, un estudio europeo titulado "*Opportunities of Artificial Intelligence*" (3) concluye que se pueden distinguir diferentes tipos de aplicaciones de la IA:

- De un lado, se refiere a la mejora del rendimiento y la eficiencia de los procesos industriales a través de la supervisión inteligente, así como a las aplicaciones de optimización o control con toma de decisiones automática y capacidades cognitivas (por ejemplo, a través del aprendizaje en línea).
- Y de otro lado, una categoría más amplia que la anterior, se refiere a la colaboración entre personas y máquinas, que puede incluir la optimización de la interfaz hombre-máquina, la automatización de la gestión del personal y las aplicaciones de realidad virtual/aumentada (por ejemplo, para la formación a distancia y en el puesto de trabajo).

No obstante, aunque los sistemas actuales de IA están lejos de la *AI singularity*, definiéndose esta como la capacidad de las máquinas de ser conscientes de sí mismas y, de esa forma, ser 'inteligentes' abordando las mimas tareas que pueden realizar las personas, es innegable que los grandes resultados obtenidos por la IA en determinados campos, en los cuales las capacidades cognitivas, de análisis y de toma de decisiones de estos sistemas pueden llegar a superar a las del ser humano, han provocado que la IA se haya convertido en la gran catalizadora de la siguiente gran revolución industrial, en la cual la tecnología, en su definición más extensa, sentará las bases de nuevas formas de vida, de trabajo, de investigación y de progreso.

(1) *Ethics Guidelines for Trustworthy AI*, abril de 2019, p. 36.

(2) SZCZEPANSKI, M., *Economic impacts of AI*, Bruselas: *EPRS European Parliamentary Research Service*, 2019, p.

(3) EAGER, J., WHITTLE, M., SMIT, J., CACCIAGUERRA, G., LALE-DEMOZ, E. et al. (2020). *Opportunities of Artificial Intelligence. Policy Department for Economic, Scientific and Quality of Life Policies*.

Con todo esto la IA es aún, hoy en día, una gran desconocida para la gran mayoría, tanto para perfiles técnicos como no técnicos. Los sistemas de IA son modelos, actualmente en mayoría matemático-estadísticos que, a partir de un gran conjunto de datos, son capaces de aprender a realizar una cierta acción, tales como predicciones, estimaciones, simulaciones o acciones puramente físicas, en base a lo que los datos le enseñan. Este aprendizaje, que puede ser o no supervisado por humanos, permite a los modelos IA realizar esas acciones con una efectividad y eficiencia igual a la que un ser humano haría en un mismo supuesto y, en determinados casos, superar ampliamente las capacidades humanas y elevar por completo la calidad de los resultados obtenidos.

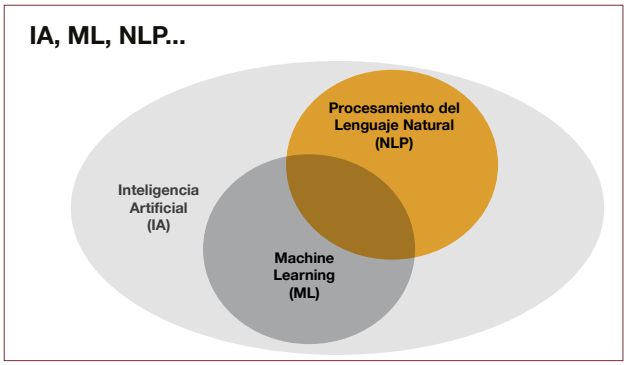
Como se puede intuir hasta ahora, y se evidenciará a lo largo de todo el documento, el uso ético y responsable de la inteligencia artificial es consustancial a la misma.

Machine Learning e inteligencia artificial

Hasta ahora hemos utilizado de manera el término "inteligencia artificial" para todo tipo de casos. Y así lo seguiremos haciendo a lo largo de todo el documento. Sin embargo, existe una diferenciación entre dicho concepto y el conocido como *Machine Learning* que se puede resumir en tres sencillas afirmaciones:

- El ML busca hacer predicciones y tomar decisiones, a la vez que realiza un proceso de aprendizaje automático dirigido por humanos en mayor o menor medida (aprendizajes supervisados y no supervisados respectivamente).
- La IA emula las capacidades cognitivas humanas de forma artificial, orientando muchas de sus aplicaciones prácticas a la interacción directa con humanos.
- El ML es una de las herramientas que puede usar la IA para imitar dichas capacidades cognitivas.

Por ejemplo, las tecnologías de voz son IA, ya que es una de las principales capacidades cognitivas inteligentes del ser humano, y en parte se apoyan en ML.



Concretamente, el ML es la ciencia que permite a estos sistemas de IA aprender de forma autónoma, basándose en algoritmos matemáticos y estadísticos que, a partir de un gran conjunto de datos, pueden aprender a realizar acciones, tal y como se ha descrito anteriormente. Según el campo de aplicación se utilizan algoritmos distintos, a fin de minimizar la probabilidad de error de la acción realizada.

Entre todos los tipos de algoritmos de ML, destacan los algoritmos de aprendizaje profundo, conocidos como "redes neuronales", o *neural networks* en inglés, debido a que son los que han obtenido resultados más sobresalientes en multitud de campos de aplicación, cada uno con mayor impacto sobre nuestras vidas que el anterior. Sin embargo, pese a ser los algoritmos que han generado los mayores avances de la IA, son a su vez los algoritmos más complejos de entender y, sobre todo, de explicar. Esto, unido a la incertidumbre regulatoria actual, son los dos principales escollos para la adopción masiva de la IA.

Impacto y riesgos de la IA

El impacto socioeconómico de la IA es evidente. Gran parte de las empresas con las mayores valoraciones en bolsa a nivel mundial basan su modelo de negocio, parcial o totalmente, en modelos de IA.

Empresas como Facebook o Twitter, que al ser redes sociales generan un gran impacto social, tienen una dependencia total con sus modelos de IA, los cuales les permiten determinar el perfil de consumo, político e ideológico de los usuarios, permitiendo a estas empresas prever patrones de compra, de voto o de opinión, en muchas ocasiones sin que el usuario sea consciente de ello.

Adicionalmente, otras grandes empresas basan gran parte de su valor añadido en sus propios modelos IA, como, por ejemplo: Tesla, con sistemas de IA de conducción autónoma; Netflix, con recomendaciones de contenido audiovisual; Amazon, con predicciones de demanda y optimización de stock en almacenes; o Apple, con la mejora sustancial del rendimiento de sus 'chips'.

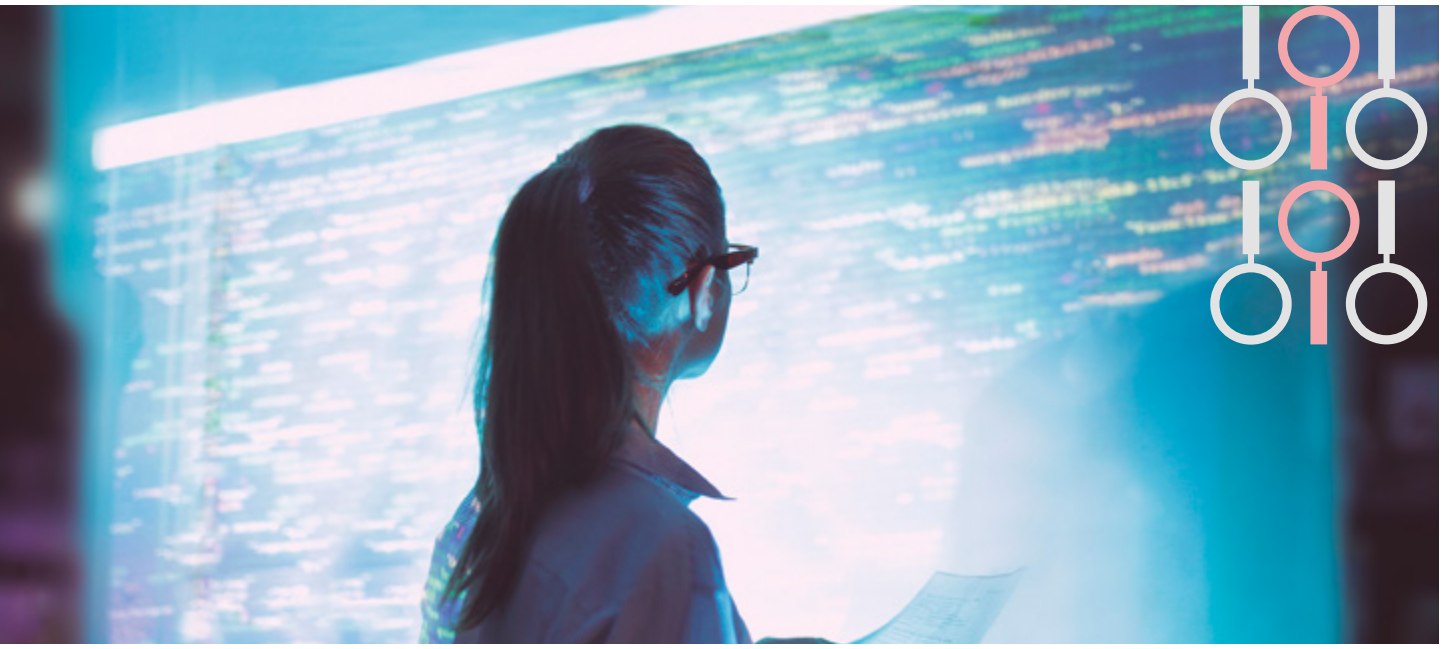
Pero también para empresas "tradicionales" de sectores como *retail*, energía, telecomunicaciones, servicios financieros, seguros, etc. la IA puede generar un gran valor en términos de eficientar los procesos de negocio, generar más ingresos y nuevos productos y servicios.

Los continuos avances en materia de investigación y aplicación de la IA han puesto de manifiesto que su adopción ha de ser prioritaria para la totalidad de las empresas si desean seguir siendo competitivas durante la próxima década. No obstante, la adopción de esta tecnología tiene un gran reto que las empresas han de saber gestionar; el alto impacto que las aplicaciones de la IA tienen ya sobre nuestras vidas ponen en relevancia el hecho de que, nunca en la historia, ha sido tan prioritario asegurar un uso responsable y ético de una tecnología.

Un mal uso de esta tecnología puede tener consecuencias graves para una empresa, no solo a nivel reputacional, sino también para sus propios consumidores que puedan sufrir consecuencias negativas, aunque no sean intencionadas. Algunas empresas tecnológicas, conscientes del potencial de la IA y la importancia de su uso responsable, lideran el desarrollo de estos modelos y su democratización.

Empresas como IBM, Google o Microsoft inviertan grandes cantidades de dinero en generar investigaciones, análisis y, más importantemente, herramientas que facilitan a las empresas la adopción de la IA al permitirles generar modelos de IA que garanticen un tratamiento al usuario justo y una toma de decisiones equitativa, segura, explicable y transparente.

Sin embargo, las aplicaciones de la IA también pueden plantear retos y riesgos, que será necesario evaluar. El siguiente apartado pretende contextualizar la inteligencia artificial Responsable como elemento central del presente documento.



1. Introducción

Qué es la IA ética

Hablamos de una inteligencia artificial ética cuando su uso no tiene impactos negativos sociales o éticos. No es la tecnología *per se* la que es ética o no, si no es el uso que hagamos de ella.

Según la Real Academia Española:

> La ética es: un conjunto de normas morales que rigen la conducta de la persona en cualquier ámbito de la vida

> La moral es: doctrina del obrar humano que pretende regular el comportamiento individual y colectivo en relación con el **bien** y el **mal** y los deberes que implican

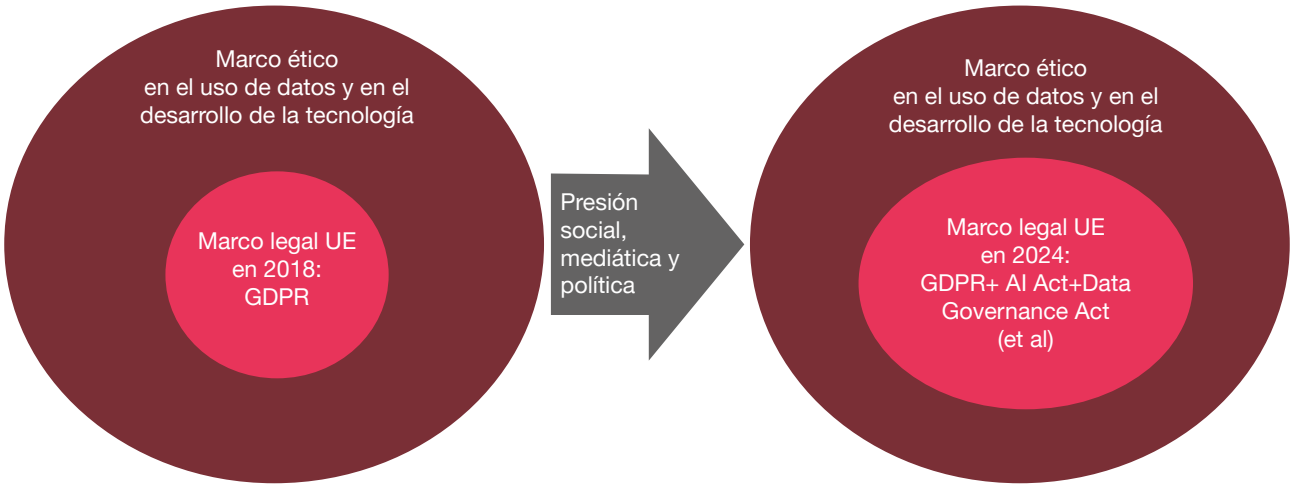
Podemos deducir de esto que un uso ético de la inteligencia artificial implica que su resultado fomenta el bien y evita el mal. Es decir, que el propósito de la inteligencia oficial es hacer bien para las personas a quien aplica, o es evitar posibles consecuencias malas o negativas para las personas o para la sociedad en su conjunto más allá de lo señalado por la ley. Así, puede haber comportamientos corporativos que sean legales, que que puedan ser percibidos por la sociedad como poco éticos.

Como ya hemos visto, la inteligencia artificial es una tecnología transformadora con muchas oportunidades para mejorar nuestra vida, nuestra economía y nuestra sociedad. Pero también sabemos que su uso puede tener consecuencias negativas, aunque muchas veces no intencionadas. ¿Quién no conoce el caso del sistema de IA que discrimina a las personas de color en el ámbito jurídico en Estados Unidos? ¿O el sistema inteligente que ayudaba a la contratación de personal y que discriminaba por género? ¿O la influencia en procesos democráticos por algoritmos de recomendación que fomentan la polarización de la sociedad? ¿Escándalos de privacidad? ¿O los algoritmos opacos que buscan el fraude en prestaciones sociales? ¿Sesgos de género en la traducción automática? Y existe un largo etc.

Estas consecuencias negativas (no intencionadas, si no futo de la carencia de controles en su proceso de conceptualización y posterior desarrollo) ocurren porque un sistema de inteligencia artificial aprende de muchísimos datos, que pueden contener sesgos; su exactitud nunca es un 100 %, es decir, siempre comete errores; y los algoritmos modernos son estructuras muy complejas difícilmente entendibles por las personas. Las consecuencias no deseadas de estos desarrollos desafortunados enturbian la reputación del resto de casos de uso, permean en la opinión pública e impulsan el debate en torno a la necesidad de nuevos controles y regulación al respecto siguiendo el siguiente ciclo:

Fuente: Odiseia

Llegados a este punto cabe hacer una distinción entre comportamiento ético y comportamiento acorde con el marco normativo. A nuestro entender la ética va siempre más allá del cumplimiento de la legalidad vigente. El de la ética es un listón que está más alto que el de la ley, si bien cabe enfatizar que el marco legal es a su vez móvil y evolutivo, y que depende ante todo de que la percepción ciudadana sobre los riesgos asociados a una innovación tecnológica trascienda del debate público y mediático al político, activándose con ello la acción legisladora.



Fuente: Odiseia

Así, en febrero de 2020 la Comisión Europea publico conjuntamente un White Paper on AI y una consulta pública al respecto de esta tecnología, cuyas respuestas -más de 1.200 por parte tanto de ciudadanos como de entidades- arrojaron, entre otros los siguientes resultados:

- El 90% de las personas que respondieron creen que la IA puede atentar contra los derechos fundamentales.
- El 87% opinan que la IA puede llevar a resultados discriminatorios.
- El 78% cree que la IA puede desencadenar decisiones y acciones cuyos motivos no pueden ser explicados.
- EL 70% afirma que sus resultados no son siempre acertados.

Adicionalmente el 42% de las personas y entidades que respondieron afirmaron que se hace necesaria una nueva legislación que regule el desarrollo y las aplicaciones de la IA. Dados estos resultados, y sobre la base de multitud de reflexiones previas -fundamentalmente las recogidas en las estrategias sobre IA nacionales y en las Directrices éticas para una IA fiable, publicadas en 2019- se comenzó a trabajar la propuesta de AI Act hecha pública en 2021 y en actual tramitación parlamentaria.

El objetivo de este documento GuIA es explicar de una manera práctica, a las empresas y organizaciones que quieran aplicar la inteligencia artificial para mejorar su actividad, cómo hacerlo de tal forma que evite las consecuencias negativas (no éticas). Aunque muchas normativas pretenden formalizar la ética y la conducta moral, en el caso de la inteligencia artificial aún no existe un marco normativo vigente. Por tanto, cuando hablamos de ética, nos referimos a normas no sancionables, no jurídicas. Es decir, la aplicación de la ética es un proceso voluntario. Si no se cumple, jurídicamente no existe una infracción. Por eso, la normativa es complementaria a la ética, como veremos en la siguiente sección.

Aunque la aplicación de la ética a la inteligencia artificial no es una ciencia, no quiere decir que no se pueda sistematizar y gestionar. De hecho, existen metodologías y procesos que minimizan el uso no ético. Estas metodologías tratan de concretizar los pasos importantes cuando uno quiere aplicar la inteligencia artificial, incluyendo:

- Seleccionar los principios de inteligencia artificial más adecuado para el sector en cuestión. Los riesgos éticos no son iguales en el sector industrial (donde la robustez y seguridad es muy importante) y el sector sanitario (donde la explicabilidad es importante).
- Estimar el daño en caso de que algo vaya mal con la aplicación del sistema. No es lo mismo equivocarse en la recomendación de una película que en el diagnóstico de una enfermedad grave como el cáncer. El daño de un error de la inteligencia artificial se puede estimar teniendo en cuenta su gravedad, su escala y la probabilidad de que ocurra.
- Aplicar la metodología "inteligencia artificial ética desde el diseño". Esta metodología considera que es necesario unos principios de IA, una formación y concienciación a los empleados, un cuestionario con las preguntas adecuadas además de unas recomendaciones, herramientas técnicas, y un modelo de gobernanza que regule todo el ciclo de vida de un sistema de IA. Si una organización sigue esta metodología, es probable que detecte y pueda mitigar los posibles riesgos.

Una vez conocidos los riesgos y las posibles medidas correctivas, la organización tiene que decidir cómo actuar. Lo correcto (lo ético) sería que en caso de que no se pueda corregir y evitar los riesgos considerados inaceptables, no usar la inteligencia artificial para este caso de uso.

En este momento, "termina" la parte ética y podría entrar la parte normativa. Una normativa podría exigir un proceso riguroso para aplicaciones de alto a riesgo antes de que se pueda poner en producción. Y si no se cumple este proceso, imponer sanciones. Esto es justamente el objetivo de la nueva regulación de IA (AI Act) propuesta por la Comisión Europea y a fecha de febrero 2022 en debate en el Parlamento Europeo y los Estados Miembros.

Allí también entra una posible certificación que certificaría que, en casos de alto riesgo, se haya hecho todo lo posible para evitar riesgos inaceptables del uso de la inteligencia artificial.

1. Introducción

La importancia de la normativa. Cómo se complementa con la ética

En el marco de la inteligencia artificial, la normativa y la ética se complementan de manera natural. ¿Puede existir una sin la otra? ¿Se retroalimentan o son independientes? La ética marca las pautas esenciales a seguir en cada una de las fases del ciclo de vida de la inteligencia artificial, mientras que la normativa establece instrucciones concretas que determinan lo que se puede (norma dispositiva) y lo que se tiene (norma imperativa) que hacer, así como lo que está prohibido llevar a cabo, según la sociedad haya acordado.

Las normas que rigen en un lugar forman el ordenamiento jurídico. Estas normas pueden tener un mayor o menor impacto en la toma ética de decisiones en general y en el marco de la inteligencia artificial en particular. Las normas que tienen un especial impacto positivo en la adopción de prácticas éticas forman parte de la llamada ética-normativa. Un claro ejemplo de ética-normativa lo encontramos en el Considerando 58 del Reglamento General de Protección de Datos, de la Unión Europea, a través del cual se establece que los sujetos obligados tienen que cumplir con el principio de transparencia, que exige que "toda información dirigida al público o al interesado sea concisa, fácilmente accesible y fácil de entender, y que se utilice un lenguaje claro y sencillo, y, además, en su caso, se visualice", lo que conlleva que el principio ético de transparencia en el tratamiento de datos se vea reforzado con una norma imperativa fiscalizable por los poderes públicos y que, de incumplirse, se impongan las correspondientes sanciones.

La ética y la ética-normativa están destinadas a convivir pacíficamente y desarrollarse de forma activa en un mundo en continua evolución. Estos dos conceptos han visto prosperar su contenido a lo largo de la Historia en la que la humanidad ha empleado su intelecto y su ingenio para idear, construir, desarrollar y mejorar soluciones para la convivencia social. Sin embargo, ahora el ser humano puede programar una solución de inteligencia artificial para la toma de decisiones con impacto sobre los negocios, la naturaleza y la propia vida humana. La inteligencia artificial no es y no puede ser ajena a los principios y normas en evolución que rigen el comportamiento humano y mantienen o mejoran la estabilidad social.

La creación de ética-normativa supone uno de los mayores retos de la labor legislativa moderna. A través de la positivización de la ética, el legislador debe procurar convertir en obligatoria la práctica de la equidad y la inclusión en normas que sean aplicables al ciclo de vida de la IA, de forma que se sancione la reproducción o la creación de estereotipos sociales, como la discriminación de minorías, razas o géneros.

En el plano de la explicabilidad, la labor legislativa es aún más compleja, dado que tendrá que encontrar la forma de obligar a través de normas imperativas a que los modelos inteligentes sean explicables internamente o ante un auditor, con la dificultad de aplicar baremos medibles a métricas subjetivas, en particular, cuando puedan resultar en consecuencias en el día a día de un consumidor final. La aplicación de la privacidad podría ser más sencilla de aplicar dado que la norma actual ya sanciona la falta de responsabilidad proactiva a la hora de garantizar que los datos tratados mediante IA sean manipulados, sustraídos, suprimidos o accedidos por terceros no autorizados. Por su parte, en materia de robustez y seguridad sobre la IA se están dando pasos legislativos interesantes mediante la creación de Reglamentos con estándares y sellos de calidad informática sujetos a auditorías constantes, aunque aún parece faltar una capa normativa coercitiva que obligue a securizar y robustecer estos sistemas con objeto de mantener una IA confiable y ética a lo largo de todo su ciclo de vida.

La tendencia actual en relación con la ética-normativa es que regule o, en su caso, limite las acciones que pueden ser llevadas a cabo mediante inteligencia artificial, en lugar de regular la tecnología en sí misma. La ética-normativa aplicada a la IA no trata de prohibir o permitir una u otra tecnología sino regular las consecuencias inmediatas, mediatas o causales cuando no pueden preverse de la inteligencia artificial, incluso cuando son provocadas por el factor de interdependencia, es decir, cuando los resultados de una IA afectan al ciclo de vida y los resultados de otra IA o, explicado a través de un ejemplo, una IA ética en el sector salud ideada y testada contra los sesgos podría verse afectada por otra IA sesgada que le provee de datos para la toma de decisiones. Una ética-normativa coercitiva podría ayudar a evitar incluso estos riesgos de interdependencia entre diferentes IA para anular o mitigar estos efectos éticamente negativos.

¿La humanidad está preparada para normativizar la ética? ¿Es el momento? La ética-normativa evoluciona y se complementa con la ética para una mejor inteligencia artificial en beneficio de la humanidad. Regular sobre IA ética en la actual fase embrionaria de la inteligencia artificial es, sin duda, un reto sometido a una larga travesía de iteraciones, que requiere estudio y análisis y por el que la humanidad debe transitar en busca de una IA ética y confiable.

1. Introducción

El papel fundamental de la SEDIA en España

La Secretaría de Estado de Digitalización e Inteligencia Artificial (SEDIA) del Gobierno de España es un órgano superior del Ministerio de Asuntos Económicos y Transformación Digital. Las competencias de esta Secretaría de Estado son las relativas a la política de impulso a la digitalización de la sociedad y economía, a través del fomento y regulación de los servicios digitales y de la economía y sociedad digitales, la interlocución con los sectores profesionales, industriales y académicos, así como el impulso de la digitalización del sector público. Fue fundada en Enero de 2020, y su responsable es Carme Artigas.

En los 2 años que han transcurrido desde su creación a día de hoy (Febrero de 2022) hemos asistido a un importante número de iniciativas de calado procedentes de la SEDIA posiblemente nunca antes visto en España, y que tienen como objetivo la modernización de nuestra economía y sociedad. Solamente por citar las más relevantes:

Plan España Digital 2025

Plan Nacional de Competencias Digitales

Carta de derechos digitales

Estrategia Nacional de Inteligencia Artificial (ENIA)

Su relación con la inteligencia artificial ética y normativa

La SEDIA se encarga de las políticas de impulso a la digitalización de la sociedad y economía de forma respetuosa con los derechos individuales y colectivos, así como con los valores del ordenamiento jurídico español, y es aquí donde comenzamos a ver su relación con la Inteligencia artificial ética y normativa.

De entre las iniciativas anteriormente señaladas, remarcamos dos contribuciones importantes en el ámbito de la IA ética y normativa: la estrategia nacional de Inteligencia Artificial y la Carta de derechos digitales.

Estrategia Nacional de Inteligencia Artificial (ENIA)

Dicha estrategia, publicada en noviembre de 2020 y enmarcada en el programa España Digital 2025, dedica por entero uno de sus seis ejes fundamentales a la inteligencia artificial ética y normativa, proponiendo su materialización a través de las siguientes 5 acciones estratégicas:

6.

Establecer un marco ético y normativo que refuerce la protección de los derechos individuales y colectivos, a efectos de garantizar la inclusión y el bienestar social

26. Desarrollo de un **sello nacional de calidad IA.**
27. Poner en marcha **observatorios para evaluar el impacto social de los algoritmos.**
28. Desarrollar la **Carta de Derechos Digitales.**
29. Puesta en marcha de un modelo de gobernanza nacional de la ética en la IA (Consejo Asesor IA).
30. Promoción de foros de diálogo, sensibilización y participación nacionales e internacionales en la relacion a la IA.

Fuente: resumen ejecutivo ENIA

Carta de derechos digitales

La Carta de Derechos Digitales fue presentada en julio de 2021, y dedica su punto 5.XXV a los Derechos ante la inteligencia artificial, señalando que:

1. La inteligencia artificial deberá asegurar un enfoque centrado en la persona y su inalienable dignidad, perseguirá el bien común y asegurará cumplir con el principio de no maleficencia.
2. En el desarrollo y ciclo de vida de los sistemas de inteligencia artificial:
   - Se deberá garantizar el derecho a la no discriminación cualquiera que fuera su origen, causa o naturaleza, en relación con las decisiones, uso de datos y procesos basados en inteligencia artificial.
   - Se establecerán condiciones de transparencia, auditabilidad, explicabilidad, trazabilidad, supervisión humana y gobernanza. En todo caso, la información facilitada deberá ser accesible y comprensible.
   - Deberán garantizarse la accesibilidad, usabilidad y fiabilidad.
3. Las personas tienen derecho a solicitar una supervisión e intervención humana y a impugnar las decisiones automatizadas tomadas por sistemas de inteligencia artificial que produzcan efectos en su esfera personal y patrimonial.

El hecho de que la SEDIA haya elaborado estos documentos en un periodo tan corto de tiempo y siguiendo procesos participativos que han involucrado a multitud de expertos, pone de manifiesto el nivel de prioridad que la SEDIA otorga a estas cuestiones, que son complementarias con el esfuerzo paralelo por impulsar el desarrollo de la IA y su adopción por parte del tejido empresarial de nuestro país, pues una IA ética será una IA mejor aceptada por nuestra sociedad.

2

Contexto

- Análisis de 27 iniciativas globales. Resumen ejecutivo
- El porqué de la necesidad de GuIA

2. Contexto

Análisis de 27 iniciativas globales. Resumen ejecutivo

Para tener contexto detallado del estado de la IA ética y normativa alrededor de la misma, no reescribir planteamientos y tener una base estructurada y "científica" sobre la cual profundizar de manera aterrizada (objetivo de este informe GuIA), en Julio de 2021 identificamos y analizamos globalmente 150 iniciativas publicadas en el mundo hasta dicha fecha. De todas ellas, hemos analizado con detalle 27, las que hemos considerado más notables y diferenciales.

Es posible que de entre todas las iniciativas que no analizamos con detalle existan incluso algunas relevantes. El ejercicio de selección tuvo como objetivo disponer de la mayor diversidad posible de planteamientos éticos y normativos de entre las 150 inicialmente identificadas.

También buscamos tener diversidad en cuanto a su origen, representando la perspectiva que los principales actores que conforman la sociedad tienen en relación con la IA ética, tanto en perspectiva actual como futura:

- Organismos Gubernamentales
- Asociaciones independientes
- Entidades privadas
- Académicos y profesionales independientes

Como resultado de dicho análisis generamos un extenso documento, del que en este apartado hacemos un breve resumen ejecutivo.

Dentro de cada uno de estos grupos hemos detallado su enfoque troncal, los principios más destacados que tiene en común y en algunas ocasiones hemos señalado los principales riesgos o las políticas nacionales y de cooperación internacional que, con carácter general, se han identificado en cada una de las categorías.

Organismos Gubernamentales

A. Documentos analizados de esta categoría

- *Opinion of the Data Ethics Commission*, por la *Daten Ethik Kommission*
- *G20 Ministerial Statement on Trade and Digital Economy* por el G20
- Principles on AI, por la OECD
- *AUS- Artificial Intelligence - Australia's Ethics Framework*, por el Departamento de la Industria, Ciencia, Energía y Recursos del Gobierno de Australia.
- Anteproyecto de recomendación sobre la ética de la inteligencia artificial, por la UNESCO
- *Algorithm charter for Aotearoa New Zealand*, por el Gobierno de Nueva Zelanda
- *Data ethics*, por el Gobierno de Reino Unido

B. Enfoque troncal

Con carácter general, las recomendaciones se centran en la elaboración de estándares éticos en el tratamiento de los datos y los sistemas algorítmicos, abordan la inteligencia artificial centrada en el ser humano, cómo se pueden diseñar e implementar unas políticas digitales para maximizar los beneficios y minimizar los desafíos del desarrollo de la economía digital, y superar los desafíos con especial atención a los países en desarrollo y las poblaciones subrepresentadas. También destacan que es de suma importancia promover el desarrollo de la IA asegurando el respeto de los derechos humanos y los valores democráticos para fomentar la consecución de una IA fiable.

C. Principios destacados

Los principios más destacados para los Organismos Gubernamentales se resumen a continuación:

- **Crecimiento inclusivo y bienestar**, A este respecto, se expresa la necesidad de que las partes interesadas participen de manera proactiva en la administración responsable de una inteligencia artificial confiable en la búsqueda de resultados beneficiosos para las personas y el planeta.
- **Valores centrados en el ser humano y equidad**, los actores de la IA deben respetar los derechos humanos y los valores democráticos durante todo el ciclo de vida del sistema de IA, para ello, los actores de la IA deben implementar mecanismos y salvaguardas, como la capacidad de determinación humana, que sean adecuados al contexto y consistentes con el estado del arte.
- **Equidad y no discriminación**. La equidad supone compartir los beneficios de las tecnologías de la IA en los planos local, nacional e internacional, teniendo en cuenta las necesidades específicas de los diferentes grupos de edad, sistemas culturales, diferentes grupos lingüísticos, personas con discapacidad, niñas y mujeres, y poblaciones desfavorecidas, marginadas y vulnerables. Es necesario también abordar la discriminación, las brechas digitales y de conocimientos y las desigualdades mundiales.
- **Sostenibilidad**. La aparición de las tecnologías de la IA puede beneficiar los objetivos de sostenibilidad o dificultar su implementación, dependiendo de la forma en que se apliquen en países con diferentes niveles de desarrollo.
- **Privacidad**. La privacidad, que constituye un derecho esencial para la protección de la dignidad, la autonomía y la capacidad de actuar de los seres humanos, debe ser respetada, protegida y promovida a lo largo del ciclo de vida de los sistemas de IA, tanto a nivel personal como colectivo. Es fundamental que los datos de la IA se recopilen, utilicen, compartan, archiven y supriman de forma coherente con los valores y principios.
- **Transparencia y explicabilidad**. Los actores de IA deben comprometerse con la transparencia y la divulgación responsable con respecto a los sistemas de IA. Particularmente, los afectados (favorable y desfavorablemente) por la IA deben ser conscientes de las interacciones con la misma y entender sus resultados.
- **Robustez, seguridad y protección**. Los sistemas de IA deben ser robustos y seguros durante todo su ciclo de vida; los actores de la IA deben garantizar la trazabilidad, para permitir el análisis de los resultados del sistema de IA y las respuestas a las consultas; los actores de la IA deben aplicar un enfoque sistemático de gestión de riesgos a cada fase del ciclo de vida del sistema de IA de forma continua para abordar los riesgos relacionados con los sistemas de IA, incluida la privacidad, la seguridad, protección y prejuicio.
- **Rendición de cuentas del funcionamiento adecuado de los sistemas de IA**. Todos los actores implicados deben ser responsables de que la IA sea aplicada de forma ética y respetando los principios enumerados. En este sentido, deben invertir todos los recursos posibles en contribuir a la claridad en la responsabilidad sobre los sistemas de IA para que los mecanismos de responsabilidad sean lo más fluido posible.
- **Justicia**. Incorporar este elemento a la IA es una tarea muy complicada, ya que esta significa no solo tratar de igual forma situaciones similares, sino también, dependiendo el caso, hacer excepciones atendiendo a situaciones extraordinarias, siendo una labor muy complicada para los desarrolladores.
- **Sensibilización y alfabetización**. La sensibilización y la comprensión del público respecto de las tecnologías de la IA y el valor de los datos deberían promoverse mediante una educación abierta y accesible, la participación cívica, las competencias digitales y la capacitación en materia de ética de la IA, la alfabetización mediática e informacional y la capacitación dirigida conjuntamente por actores privados y públicos.

- **Gobernanza y colaboración adaptativas y de múltiples partes interesadas**. La soberanía de los datos significa que los Estados, en cumplimiento del derecho internacional, regulan los datos generados dentro de sus territorios o que pasan por ellos y adoptan medidas para la regulación efectiva de los datos sobre la base del respeto del derecho a la privacidad y otros derechos humanos.
- **Proporcionalidad e inocuidad**. Debería reconocerse que las tecnologías de la IA no garantizan necesariamente, por sí mismas, la prosperidad de los seres humanos ni del medio ambiente y los ecosistemas. La elección de un método de IA debería justificarse de las siguientes formas:
  - El método de IA elegido debería ser conveniente y proporcional para lograr un objetivo legítimo determinado.
  - El método de IA elegido no debería repercutir negativamente en los valores fundamentales enunciados en el documento.
  - El método de IA debería ser adecuado al contexto y basarse en fundamentos científicos rigurosos.

D. Políticas nacionales y cooperación internacional para una IA fiable

Inversión en la investigación y desarrollo de IA con fondos públicos en el largo plazo, así como favorecer la inversión privada. La posición de los gobiernos es privilegiada para coordinar los esfuerzos interdisciplinarios de la multitud de actores implicados y favorecer el desarrollo de la IA.

Favorecimiento de un ecosistema digital a través de la infraestructura necesaria y mecanismos para compartir conocimiento sobre IA. En el mismo sentido que el anterior, un Gobierno implicado en la materia puede ser clave para apoyar una transición ágil a la aplicación generalizada de la IA.

Confeccionar políticas favorables para la IA para la transición de las fases de investigación y desarrollo a las fases de despliegue y operativa de IA fiable para lo cual se indica que la cooperación internacional resultará fundamental. Para esta adaptación de los marcos regulatorios, será clave la experimentación en entornos controlados para poner a prueba los sistemas de IA en fases previas a la de despliegue.

Formación en capital humano y transformación del mercado de trabajo. Pueden tener una labor clave de promover una cultura de trabajo responsable en torno a la IA para la aplicación posterior por parte de los actores privados (además de las propias iniciativas que puedan surgir del sector público).

Asociaciones independientes

A. Documentos de esta categoría

- *AI Principles*, por Asilomar
- *Benefits & Risks of Artificial Intelligence*, por Asilomar
- *Universal Guidelines for Artificial Intelligence*, por la *Public Voice coalition*
- TOP 10 *principles for ethical artificial intelligence*, por *UNI Global Union*
- *Ethical Guidelines*, por la *Japanese Society for Artificial Intelligence* (JSAI)

B. Enfoque troncal

Las asociaciones que han elaborado los informes analizados tienen como objetivo despertar interés por los grandes retos que suscita la IA y los efectos que puede tener para toda la sociedad y emiten por ello una serie de recomendaciones para promover la transparencia, la rendición de cuentas y el control humano de estos sistemas.

Tan relevante es el impacto esperado como graves son los riesgos desconocidos vinculados a la IA, por lo que deben establecerse mecanismos robustos y límites estrictos para que la IA sea confiable durante toda su vida útil y que, específicamente, no aprenda demasiado, evolucione o se multiplique tanto que acabe dominando a la humanidad o que provoque que sus fallos o sus acciones, incluso sobre la Justicia, pongan en riesgo los valores humanos. La inteligencia artificial debe anteponer a las personas y al planeta, es por eso que las discusiones éticas sobre IA a escala global son esenciales.

También se refuerza la idea de que es necesario actuar para salvaguardar los intereses de los trabajadores y mantener un equilibrio de poder saludable en los lugares de trabajo.

C. Principios destacados

Los principios más destacados para las Asociaciones son los siguientes:

- **Rendición de cuentas**. El derecho a la determinación de los resultados por parte de los seres humanos se considera fundamental ya que, en última instancia, estos son los responsables de las decisiones automáticas. Paliar la información asimétrica, ya que los sistemas de IA manejan información ingente de los individuos que, por su parte, no conocen apenas estos sistemas. Asimismo, se debe prohibir la elaboración de perfiles de forma secreta para no incidir en la asimetría de la información.
- **Seguridad**. Las asociaciones suelen apostar por la prohibición de la realización por parte de autoridades públicas de evaluaciones unitarias y el establecimiento de clasificaciones en que los individuos tengan determinados puntos con base en el análisis de una gran variedad de datos sobre los que se realice un perfil determinado La inteligencia artificial debe ser responsable, segura y útil, donde las máquinas mantienen el estatus legal de herramientas y las personas jurídicas mantienen el control y la responsabilidad de estas máquinas en todo momento.
- **Transparencia**. Los trabajadores deben tener derecho a exigir transparencia en las decisiones y los resultados de los sistemas de inteligencia artificial, así como en los algoritmos subyacentes. Esto incluye el derecho a apelar las decisiones tomadas por IA / algoritmos y a que un ser humano las revise.
- **No discriminación, equidad e inclusividad**. Cualquier sesgo, ya sea de género, raza, orientación sexual, edad, etc., debe ser identificado y no ser propagado por el sistema.
- **Asegurar una transición justa y garantizar el apoyo a las libertades y derechos fundamentales**. Es vital que se implementen políticas corporativas que aseguren la responsabilidad corporativa en relación con el desplazamiento del trabajador y las tareas laborales, como programas de reciclaje y posibilidades de cambio de trabajo. También se requieren medidas gubernamentales para ayudar a los trabajadores desplazados a volver a capacitarse y a encontrar un nuevo empleo.
- **Establecer mecanismos de gobernanza global**. Los organismos deben incluir diseñadores de IA, fabricantes, propietarios, desarrolladores, investigadores, empleadores, abogados, OSC y sindicatos.

Entidades privadas

A. Documentos de esta categoría

- *AI at Google*: *our principles*, por Google
- *Policy Principles for AI*, por *Connected Health*
- *Microsoft AI Principles*, por *Microsoft*
- *Responsible AI,* por PwC

B. Enfoque troncal

Estas entidades están altamente comprometidas con el desarrollo de la IA para que sea beneficiosa para el conjunto de la sociedad, por lo que destina importantes inversiones a su desarrollo en diversidad de proyectos destinados, entre otros, a solucionar algunas de las problemáticas más extendidas entre los consumidores finales. En el caso de las tecnológicas podemos echar en falta a IBM. Lógicamente fue analizada (de hecho, el resultado de dicho análisis está incluido de manera muy detallada en el Framework GuIA). No está incluida en este resumen por mantener proporcionalidad de las tecnológicas respecto de otro tipo de entidades privadas.

C. Principios destacados

Los principios más destacados para las Entidades Privadas se resumen a continuación:

- **No discriminación y equidad**. Evitar crear o reforzar sesgos que generen consecuencias adversas sobre la población. Los sistemas de IA deben tratar a todas las personas de forma justa, sin aplicar sesgos. La IA debería detectar y eliminar los sesgos en la mayoría de los algoritmos de decisión.
- **Seguridad**: La seguridad y la fiabilidad deben tenerse en cuenta no sólo en las circunstancias previstas, sino también en condiciones inesperadas, como cuando los sistemas son atacados, por lo que los sistemas de IA deben ser probados exhaustivamente, actualizados en función de las opiniones de los usuarios humanos y supervisados para su rendimiento continuo.
- **Excelencia científica**. El desarrollo de IA tiene que seguir el método científico y comprometerse con el rigor intelectual, integridad y colaboración, especialmente con el impacto que puede tener en sectores como el médico o medioambiental.
- **Privacidad y seguridad**: La IA y el aprendizaje automático añaden una nueva complejidad y una mayor dependencia del uso de datos para el entrenamiento de esos sistemas, para lo cual las entidades privadas destacan la necesidad de que las autoridades públicas regulen de manera clara las exigencias a este respecto.
- **Inclusión**: Todo el mundo debería beneficiarse de la tecnología inteligente. Al igual que la equidad, los sistemas de IA deben capacitar a todo el mundo e involucrar a las personas independientemente de su edad, sexo, raza o capacidades físicas o mentales. Este principio es fundamental para un sistema de IA sostenible.
- **Rendición de cuentas**: La rendición de cuentas ayuda a garantizar que los sistemas de IA no sean la autoridad final en cualquier decisión que afecte a la vida de las personas y que los seres humanos mantengan un control significativo sobre los sistemas de IA.

Académicos y profesionales independientes

A. Documentos de esta categoría

- *AI Now 2019 Report*, por la AINOW
- *Principled Artificial Intelligence: Mapping Consensus in Ethical and Rights-based Approaches to Principles for AI*, por *Harvard*
- *ECPAIS*, por el IEEE
- *Ethically Aligned Design v2*, por el IEEE
- *From What to How: An Initial Review of Publicly Available AI Ethics Tools, Methods and Research to Translate Principles into Practices*, por Varios autores (I)-
- *Responsible AI—Two Frameworks for Ethical Design Practice,* por Varios autores (II)
- *Ethics guidelines for trustworthy*, por la *European Commission*
- *A 5-step guide to scale responsible AI*, por la *World Economic Forum*
- *AI and Data Protection Risk Mitigation and Management Toolkit*, por la ICO
- *Report for the Data Governance Working Group*, por el *Global Partnership of AI*
- *Le Livre Blanc: un manifeste autour de l'engagement collectif,* por *Impact IA*

B. Enfoque troncal

Estos documentos analizan la forma en que la IA está aumentando las asimetrías de poder existentes y, desde esta perspectiva, examinan lo que los investigadores, defensores y legisladores pueden hacer para abordar de manera significativa este desequilibrio. Esto se debe a que los sistemas de inteligencia artificial continúan desplegándose rápidamente en dominios de considerable importancia social. La fiabilidad de la IA se apoya en tres componentes que deben satisfacerse a lo largo de todo el ciclo de vida del sistema: a) la IA debe ser lícita, es decir, cumplir todas las leyes y reglamentos aplicables; b) ha de ser ética, de modo que se garantice el respeto de los principios y valores éticos; y c) debe ser robusta, tanto desde el punto de vista técnico como social. Los principios éticos o declaraciones al respecto rara vez se centran en cómo se puede implementar la ética de la IA y si son efectivas.

Uno de sus objetivos principales es aportar una visión común sobre la proliferación de intentos por inspirar los "principios de la IA" y así ayudar a todos los actores, desde los reguladores a los investigadores, a realizar sus actividades desde una perspectiva ética, así como una metodología práctica para cumplir con los criterios de implementación ética de los certificados de sistemas de IA destinados al rastreo automático de personas (*Contact Tracing Applications y Contact Tracing Technologies* (CTA/CTT)).

Para alcanzar una IA ética hay que empezar formando a la sociedad tanto en materias vinculadas con el diseño, el desarrollo y el uso de la IA como en valores éticos. La sociedad, conocedora de las virtudes y ventajas que promete la IA debe ser parte en cada una de las fases de la IA. Esto supone que ya no sólo se pensará en marcar casillas de cumplimiento legal y ético, sino que se mantendrá una atención constante a cómo funciona la IA, por qué lo hace, para qué lo hace, cómo se debe mejorar, por qué se debe mejorar, cuáles son las demás funciones que debe tener, etc. La ética debe ser contemplada transversalmente y no solo en la vertical de quien la idea, la desarrolla y la usa, ya sea el que la utiliza o el que se beneficia directamente de ella.

También se establecen herramientas de gestión y mitigación de riesgos de protección de datos e inteligencia artificial que puedan surgir en los sistemas IA.

C. Principios destacados

Los principios más destacados para los autores de este grupo son los siguientes:

- **Rendición de cuentas**. El estudio detecta que, con carácter general, este tipo de publicaciones suelen argumentar que dado que el término "inteligencia artificial" sugiere equivalencia con la inteligencia humana, surge la duda de quién estará al mando de revisar la IA y de responder por ella en un futuro en que su presencia sea máxima.
- La preocupación por la **justicia y la no discriminación** es el tema más representado de los documentos estudiados. Los sistemas de IA pueden contener ciertos sesgos derivados del uso de algoritmos, como la sobre o infra representación de ciertos sectores de población.
- **Promoción de los valores humanos** para aportar al bienestar generalCon el potencial de los sistemas de IA, asegurar el respeto a los valores humanos es clave para la ética de esta tecnología, debiendo siempre priorizar enfoques humanitarios.
- **Crecimiento inclusivo, desarrollo sostenible y bienestar**. Las partes interesadas deben participar de manera proactiva en la administración responsable de la IA confiable en la búsqueda de resultados beneficiosos para las personas y el planeta.
- **Responsabilidad**. Los actores de la IA deben ser responsables del correcto funcionamiento de los sistemas de IA y del respeto de los principios anteriores, en función de sus funciones, el contexto y de forma coherente con el estado de la técnica.

D. Riesgos

- **Gobernanza**. La falta de conocimiento de la administración sobre las implicaciones de protección de datos de los sistemas de inteligencia artificial hace que no puedan demostrar su responsabilidad para comprender y abordar los problemas, y el incumplimiento del principio de responsabilidad.
- **Compensación**. Los análisis / decisiones de compensación inadecuados o inapropiados conducen a sistemas de IA que priorizan incorrectamente un criterio sobre otro criterio más importante.
- **Exactitud estadística**. La falta de procesos de prueba estructurados conduce a que las pruebas previas a la implementación no se realicen o completen de manera efectiva.
- **Discriminación**. Si no se revisa el diseño del sistema, se generan sesgos o se produce discriminación en el sistema.
- **Minimización de datos**. La falta de revisiones en cada etapa del ciclo de vida de la IA conlleva el riesgo de retención inadecuada de datos y el incumplimiento del artículo 5, apartado 1, letra c).
- **Derechos individuales**. La falta de información clara sobre cómo las personas pueden ejercer sus derechos conduce a que las personas no puedan ejercer sus derechos y al incumplimiento de los artículos 12 a 22.
- **Revisión humana**. La falta de revisores humanos con los niveles apropiados de independencia operativa y de entrenamiento conduce a que no se realicen revisiones humanas, o que las revisiones no sean completas o efectivas.

E. Políticas nacionales y cooperación internacional necesarias para la consecución de una IA confiable

- Invertir en investigación y desarrollo de IA.
- Fomento de un ecosistema digital para la IA.
- Dar forma a un entorno político propicio para la IA.
- Fortalecer la capacidad humana y preparación para la transformación del mercado laboral.
- Cooperación internacional para una IA confiable.

2. Contexto

El porqué de la necesidad de GuIA

La conclusión a día de hoy (Febrero de 2022), tras el análisis que hemos realizado de estas 27 iniciativas, sigue siendo la misma que hizo la Universidad de *Harvard* en su informe publicado en 2020 *Principled Artificial Intelligence: Mapping Consensus in Ethical and Rights-based Approaches to Principles for AI*, donde se analizaban 36 iniciativas globales acerca de una IA ética.

> Existe una amplia brecha y complicada articulación entre conceptos de alto nivel y su aplicación en el mundo real

Es por tanto necesario profundizar en los planteamientos que permitan a las empresas asimilar la Inteligencia Artificial ética y normativa de manera pragmática en su día a día. Aterrizar los conceptos éticos mediante recomendaciones, guidelines y modelos de gobierno que ayuden a las empresas a saber cómo hacerlo. Incluso aterrizar los conceptos éticos mediante tecnologías que permiten automatizar su gestión.

A continuación desarrollamos un marco ético-jurídico-tecnológico. Nuestro *FrameWork* GuIA.

Un *framework* que permite aterrizar, permite conocer cómo gestionar el principio ético y su normativa asociada a través de las tecnologías, herramientas y *guidelines* necesarias para su gestión. Llegando incluso hasta, en muchos casos, al algoritmo necesario para implementarlo, quedándonos a las puertas del *coding*.

3

Framework GuIA

- Principios éticos aplicables a la IA y su normativa asociada
- Cómo aterrizar cada principio ético
  - El enfoque de Google
  - El enfoque de Microsoft
  - El enfoque de IBM
  - Caso de éxito de Telefónica

3. Framework GuIA

Principios éticos aplicables a la IA y su normativa asociada

Introducción

Si algo se deduce de lo que hasta ahora se ha avanzado en la GuIA es lo imprescindible de garantizar la ética de la inteligencia artificial. Solo de esta forma se podrá garantizar que su expansión sea sostenible. Para conseguir este objetivo, es necesario que en todas las facetas y etapas de las soluciones de IA se respeten y actúen como referencia una serie de principios que orienten el desarrollo y uso de esta tecnología.

La sostenibilidad, en su dimensión social, económica y laboral, es un término muy amplio que se refiere a encontrar un justo equilibrio en la balanza que permita y garantice un mejor futuro, tanto para las empresas como para las personas y los gobiernos. Para garantizar que su desarrollo sea sostenible y se contribuya a la justicia intergeneracional, la evaluación exhaustiva del potencial impacto de las soluciones de IA debe ser tenida en cuenta desde el diseño y durante toda la vida de la solución. Solo así se podrá generar confianza en las personas y asegurar la prosperidad de la humanidad en su entorno.

Los documentos analizados para elaborar GuIA coinciden en que las partes interesadas deben participar de manera proactiva en la administración responsable de una inteligencia artificial confiable, cuya máxima sea la búsqueda de resultados beneficiosos para las personas y el planeta. Para ello, es preciso que los análisis y evaluaciones de impacto y de riesgos trasciendan de los intereses y derechos individuales y que se realicen en perspectiva social y colectiva.

Como se establece en la "Recomendación 31" del Proyecto de recomendación sobre la ética de la inteligencia artificial de la UNESCO (desarrollado por Grupo Especial de Expertos (GEE)) (1) , las soluciones IA pueden ser beneficiosas para la sostenibilidad o perjudiciales, dependiendo de cómo sea su aplicación en los distintos países del mundo. Por este motivo es necesario que las soluciones IA se utilicen de forma plenamente consciente de las repercusiones que estas pueden ocasionar en la sostenibilidad. En este sentido cabe destacar los Objetivos de Desarrollo Sostenible (ODS) de las Naciones Unidas como un conjunto de metas en constante evolución.

La Asamblea General de las Naciones Unidas el 25 de septiembre de 2015 (2) aprobó los 17 Objetivos de Desarrollo Sostenible (ODS), una colección de objetivos globales conectados entre sí para lograr un futuro mejor y sostenible entre los que se encuentran objetivos relacionados con aspectos democráticos, económicos, sociales, laborales, educativos y medioambientales.

Para afrontar estos retos se ha creado la Agenda 2030 concretando los objetivos en 169 metas específicas. Sin duda, la IA puede ser un instrumento importante para su logro. De este modo se determinan los objetivos vinculados a los valores y derechos humanos, justicia y no discriminación y de sostenibilidad. En este sentido, algunos estudios señalan que la IA impacta positivamente o facilita el 79% de las metas, mientras puede inhibir o impactar negativamente en el 35% (3). En general no cabe duda de que la explotación de datos y el uso de IA puede permitir "conseguir en tiempo real un nuevo conocimiento sobre el bienestar de las personas, así como para ayudar mejor a los grupos más vulnerables" (4) . Existen proyectos destacados vinculados a este nexo de IA y ODS (AI4SDGs) (5), en general vinculados a los proyectos IA for good . Así, entre otros, se han desarrollado plataformas interactivas para analizar las interconexiones de sistemas, mejorar la coherencia de las políticas, identificar riesgos, etc. También se ofrecen conjuntos de datos relacionados con ODS, recursos abiertos de formación sobre IA y ODS.

(1) UNESCO. (20 de septiembre de 2020). ANTEPROYECTO DE RECOMENDACIÓN SOBRE LA ÉTICA DE LA INTELIGENCIA ARTIFICIAL. 2020, de UNESCO Sitio web: https://unesdoc.unesco.org/ark:/48223/pf0000373434_spa

(2) Asamblea General Naciones Unidas. (21 de octubre de 2015). Transformar nuestro mundo: la Agenda 2030 para el Desarrollo Sostenible. 2015, de Naciones Unidas Sitio web: https://unctad.org/system/files/official-document/ares70d1_es.pdf

(3) Vinuesa, R. y otros "The role of artificial intelligence in achieving the Sustainable Development Goals". Nat Commun 11, 233 (2020). https://doi.org/10.1038/s41467-019-14108-y

(4) Macrtodatos y los Ods, Naciones Unidas, 2018 https://www.un.org/es/global-issues/big-data-for-sustainable- development).

(5) AI for Sustainable Development Goals (AI4SDGs), https://ai-for-sdgs.academy/ a través de la Academia de Inteligencia Artificial de Beijing (BAAI).

(6) Entre otros, ver "AI + Sustainable Development Goals", https://ai4good.org/ai-for-sdgs/

Tan importante es el impacto esperado y tan graves son los riesgos desconocidos vinculados a la IA a los que nos enfrentamos, que se deben establecer mecanismos robustos y límites estrictos para que la IA sea confiable durante toda su vida útil y que, específicamente, no aprenda demasiado, evolucione o se multiplique tanto que acabe dominando a la humanidad o que provoque que sus fallos o sus acciones, incluso sobre la Justicia, pongan en riesgo los valores humanos. Igualmente es esencial tener presente el principio de responsabilidad y evitar riesgos irreversibles. Así, "Las generaciones actuales tienen la responsabilidad de garantizar la plena salvaguardia de las necesidades y los intereses de las generaciones presentes y futuras." (art. 1 Declaración sobre las responsabilidades de las generaciones actuales para con las generaciones futuras, UNESCO 1997) y "deben esforzarse por asegurar el mantenimiento y la perpetuación de la humanidad".

En definitiva, la inteligencia artificial debe ser de utilidad para las personas y el planeta, por lo que deben incluirse todas las medidas necesarias para garantizar que, a través del respeto del conjunto de principios que se expondrán a continuación, sea una tecnología sostenible que contribuye al interés público. Como máxima, los beneficios deben ser siempre mayores que los potenciales costes (1).

Toda la sociedad, con especial énfasis en los grupos minoritarios, y el medio ambiente deben ser partes interesadas en todo el ciclo de vida de la IA. La sostenibilidad y la responsabilidad ecológica deben fomentarse. Es necesario impulsar la investigación para que en la promoción del desarrollo sostenible la inteligencia artificial tenga un papel protagonista.

En la medida en que los sistemas de IA respeten la sostenibilidad y contribuyan a su promoción se conseguirá proteger el futuro de las generaciones venideras. De esta forma, se exponen a continuación los principios de GuIA que, tanto por separado como de forma conjunta, deben contribuir a la garantía de la sostenibilidad en las soluciones de IA.

(1) Dawson D. and Schleiger E., Horton J., McLaughlin J., Robinson C., Quezada G, Scowcroft J, and Hajkowicz S, Artificial Intelligence: Australia's Ethics Framework, 2019.

Privacidad y gobierno de datos

Definición descriptiva del principio

Salvaguardar la privacidad de las personas durante todo el ciclo de vida de las soluciones de inteligencia artificial es una cuestión básica ante la proliferación de este tipo de tecnología ya que su desarrollo, entrenamiento y explotación se caracterizan principalmente por el uso intensivo de los datos y, por supuesto, de datos personales. Este principio queda especialmente afectado por las soluciones IA y, al estar destinado a la prevención del daño, debe abarcar tanto la calidad e integridad de los datos como su pertinencia para el proyecto en cuestión (1).

En este sentido, este principio y derecho fundamental está ya reconocido como una tradición constitucional de los estados y el Derecho de la Unión Europea. El derecho a la privacidad ha sido sólidamente reconocido por la jurisprudencia y tiene como base jurídica los artículos 7 y 8 de la **Carta de los Derechos Fundamentales de la UE** relativos a la protección de las personas físicas en lo que respecta al respeto de la vida privada y familiar. Además, como principal referencia, del **Reglamento General de Protección de Datos**, en adelante "RGPD", se derivan derechos y obligaciones concretas, fundamentalmente las expuestas a continuación:

- Transparencia y acceso a información sobre los datos, cómo se usarán y para qué finalidades.
- Obtención del consentimiento.
- Ejercicio de derechos de los interesados.
- Garantizar la seguridad de los datos personales en todo momento.
- Derecho a la rectificación y eliminación.

Habida cuenta de los grandes efectos adversos que puede tener para las personas la utilización de soluciones IA por su poder de análisis y utilización de los datos personales, estas deben tener el derecho de acceder, manejar, controlar y eliminar de una forma accesible y sencilla sus datos. Además, el consentimiento debe ser prestado de forma voluntaria, es decir en condiciones de libertad tanto real como percibida (2).

El impacto potencial de este tipo de tecnología sobre la privacidad de las personas hace necesario incorporar a los proyectos que utilicen IA el principio de privacidad desde el diseño y por defecto (también conocido por su expresión inglesa *Privacy by design and by default*). Esto significa que la privacidad debe vigilarse durante todo el ciclo de vida de las soluciones basadas en inteligencia artificial: desde la captación de los datos hasta su utilización.

Bajo esta premisa, la privacidad por desde el diseño (o *Privacy by design*) debe dar cabida a la realización de notificaciones y solicitudes de consentimientos, fomentar la creación de estructuras y procesos en los proyectos que cuenten con salvaguardas de privacidad y, por supuesto, facilitar mecanismos adecuados de transparencia y control sobre el uso de los datos (3).

Uno de los principales objetivos de la Unión Europea con la norma citada es proteger la privacidad de las personas, darles más control sobre sus datos personales y establecer una nueva relación entre el usuario y el sistema. En este sentido, el tratamiento lícito de datos personales y el uso no vigilado de las soluciones que incorporan IA podrían entrar en conflicto. La Unión Europea ha desarrollado y mantiene en vigor normativa tanto de carácter general como específico para nuevas tecnologías, que permite resolver las principales cuestiones derivadas de la privacidad y el gobierno del dato. Sin embargo, no dispone de normativas específicas para el uso de soluciones basadas en inteligencia artificial.

(1) Comisión Europea - Grupo Independiente de Expertos de Alto nivel sobre inteligencia artificial, Directrices éti cas para una IA fiable, 2019.

(2) Future of Life Institute. (2020). ASILOMAR AI PRINCIPLES. 2020, de Future of Life Institute Sitio web: https://futureoflife.org/2017/08/11/ai-principles/

(3) Sundar Pichai. (7 de junio de 2018). AI at Google: our principles. 2018, de Google Sitio web: https://www.blog.google/technology/ai/ai-principles/

La utilización de IA atrae casi por defecto la aplicación del régimen de protección de datos. La normativa de protección de datos es aplicable a la IA en tanto que en muchos casos estas soluciones implican la elaboración de perfiles de personas físicas o la toma de decisiones automatizadas sobre estas. De igual modo es aplicable el régimen de protección de datos cuando los variados macrodatos que alimentan la IA sean datos de personas identificadas, identificables o reidentificables. En los casos de aplicación, resulta necesario garantizar, entre otros requisitos, los principios, la legitimación del tratamiento, los derechos, la responsabilidad proactiva y la privacidad en el diseño, así como el cumplimiento de la norma en el régimen de las transferencias internacionales de datos. Igualmente, se exigirá el estudio de impacto por defecto.

Así, los principios y derechos de protección de datos han pasado a ser la estructura básica de los derechos de las personas ante la IA y la fuente de obligaciones de quienes la desarrollan. El principio esencial de cumplimiento responsable y proactivo y la llamada ética en el diseño proceden directamente del Derecho de protección de datos. Asimismo, buena parte de los contenidos de los principios de la IA están claramente conectados: transparencia y explicabilidad, seguridad, privacidad, gestión, gobernanza y calidad de los datos, control humano, responsabilidad, etc. La proyección mundial del Derecho de la UE trae causa, entre otros, del sistema del RGPD, que también se sigue en la propuesta del esperado Reglamento de IA, aún en fase de propuesta en la Unión Europea.

Desarrollo normativo

A continuación, analizamos la evolución normativa del principio de privacidad y gobierno de los datos y su relación con otras normativas europeas.

El derecho a la protección de datos personales es regulado por primera vez de forma expresa en el **Convenio para la protección de las personas físicas en el tratamiento automático de datos personales** (Convenio núm. 108). En este acuerdo, los "datos personales" mencionados en el artículo 2.a se refieren a cualquier información relacionada con una persona física identificada o identificable. Por lo tanto, se puede vincular no solo a datos personales y familiares, sino también a cualquier tipo de información. Este Convenio ha sido reformado precisamente para reconocer el derecho frente a las decisiones automatizadas (art. 9. 1º, mayo 2018).

Además de su reconocimiento jurisprudencial, la "Carta de los Derechos **Fundamentales de la Unión Europea**" hace de la protección de datos personales un derecho básico en su artículo 8.

Entre el Derecho derivado, durante décadas destacó la **Directiva 95/46 CE del Parlamento Europeo y del Consejo** de 24 de octubre de 1995, relativa a la protección de las personas físicas en cuanto al tratamiento de datos personales y la libre circulación de estos datos. A partir de esta se ha desarrollado en toda la UE un sólido y completo Derecho de protección de datos que ha culminado con la aprobación y entrada en vigor en 2016 del Reglamento General de Protección de Datos.

Así, a partir del 25 de mayo de 2018 es de aplicación el **Reglamento General de Protección de Datos**. Cabe destacar que, sin perjuicio de la importancia de la necesidad de legitimación, por lo general a través del consentimiento, así como el reconocimiento de Derechos, el RGPD fortalece el modelo de responsabilidad proactiva que impone la referida privacidad en el diseño y por defecto. Asimismo, el principio de proporcionalidad y especialmente de minimización obliga a tratar los mínimos datos posibles o, por ejemplo, deriva la medida de seguridad de anonimizar o seudonimizar los datos siempre que sea posible para evitar o mitigar riesgos.

En el ámbito de la IA resulta especialmente destacable el "derecho" a no ser sometido a decisiones automatizadas reconocido en el artículo 22 RGPD (o el artículo 11 de la Directiva (UE) 2016/680 o art. 9. 1º Convenio 108). En este se reafirma como derecho subjetivo lo que es un conjunto de especialidades y garantías. La clara intención de este "derecho" es que las decisiones solo automatizadas y de impacto para las personas, por su sensibilidad o particularidad, tienen que ser compensadas con garantías especiales. Así, se imponen particulares deberes de transparencia e información significativa sobre los datos utilizados, su calidad, gestión adecuada, pertinencia, procesamiento y la lógica aplicada. Igualmente se garantiza la posibilidad de recurrir y la intervención humana en el tratamiento. Asimismo, se pueden exigir medidas específicas como controles y auditorías, seguridad reforzada, garantías contractuales, legislación específica que permita los tratamientos con datos especialmente protegidos, anonimización y seudonimización, comités de ética, entre otros.

En relación con las principales obligaciones para las organizaciones: En concreto las implicaciones, de obligado cumplimiento, que tienen que vigilar las compañías en materia de protección de datos son:

(i) designación de un delegado de protección de datos cuando se tratan datos a gran escala,

(ii) aquellas compañías que tienen sede fuera de la UE tienen la obligación de aplicar las mismas normas al ofrecer servicios o productos,

(iii) integrar salvaguardias desde las primeras etapas del desarrollo de los servicios y productos,

(iv) utilizar técnicas que respeten la privacidad de los usuarios, como el cifrado o seudonimización,

(v) llevar a cabo evaluaciones del impacto cuando el tratamiento de datos pueda ocasionar un mayor riesgo para los derechos y libertades de las personas, y,

(vi) mantener un registro de las actividades del tratamiento llevadas a cabo por la Compañía.

Estas obligaciones son cada vez más relevantes ya que la tecnología ha transformado la economía como la vida social y cada vez más es necesario facilitar la libre circulación de datos personales garantizando un nivel de protección adecuado. Por tanto, los desarrolladores de soluciones basadas en inteligencia artificial deben tomar medidas razonables, teniendo en cuenta la tecnología y los medios a su disposición.

Para el tratamiento de datos personales en el ámbito penal y policial cabe destacar la Directiva (UE) 2016/680, de 27 de abril, relativa a la protección de las personas físicas en lo que respecta al tratamiento de datos personales por parte de las autoridades competentes para fines de prevención, investigación, detección o enjuiciamiento de infracciones penales o de ejecución de sanciones penales, y a la libre circulación de dichos datos.

Esta directiva tiene como objetivo proteger mejor los datos de las personas cuando estos son tratados por las autoridades policiales y de justicia penal. Por tanto, exige a las autoridades policiales que los datos recogidos sean:

(i) tratados de forma lícita y legal,

(ii) recogidos con fines determinados, explícitos y legítimos y tratados en consonancia con dichos fines,

(iii) adecuados, pertinentes y no excesivos en relación con el fin para el que son tratados, (iv) exactos y actualizados si fuera necesario,

(iv) conservados de forma que permita identificar a la persona durante un período no superior al necesario para el fin del tratamiento, y

(v) protegidos adecuadamente, incluida la protección frente al tratamiento no autorizado o ilícito.

En concreto, cuando hace uso de nuevas tecnología, como es el caso de soluciones basadas en inteligencia artificial, por su naturaleza, alcance, contexto o fines, supongan un alto riesgo para los derechos y libertades de las personas físicas, el responsable del tratamiento debe llevar a cabo, con carácter previo, una evaluación del impacto de las operaciones de tratamiento previstas en la protección de datos personales, que permita determinar los riesgos derivados de dicho tratamiento para los derechos y libertades y establecer medidas mitigadoras. Además, si estas medidas no permiten limitar el impacto o reducir la probabilidad de materialización, el responsable debe consultar con la autoridad de control antes de proceder al tratamiento de los datos personales.

En relación con la **protección de datos aplicable al registro de nombres de pasajeros**, contamos con la Directiva (UE) 2016/681 del Parlamento Europeo y del Consejo, de 27 de abril de 2016, relativa a la utilización de datos del registro de nombres de los pasajeros (PNR) para la prevención, detección, investigación y enjuiciamiento de los delitos de terrorismo y de la delincuencia grave.

Esta directiva tiene como núcleo central regular la transferencia de datos del registro de nombres de los pasajeros y el tratamiento de dichos datos por las autoridades competentes. Y limitar el tratamiento de los datos recabados a las finalidades de prevención, detección, investigación y enjuiciamiento de los delitos de terrorismo o delitos graves.

En concreto, solo deben tratarse los datos para:

(i) realizar una evaluación antes de la llegada de los pasajeros;

(ii) determinadas investigaciones o enjuiciamientos,

(iii) contribuir al desarrollo de los criterios de evaluación del riesgo.

Además, se establecen obligaciones al responsable respecto al almacenamiento como el plazo de conservación de estos y en especial, transcurridos seis meses los datos transmitidos deben ser "despersonalizados" para ocultar determinada información. De hecho, la divulgación de datos completos únicamente está permitida si es razonablemente necesario para atender a las solicitudes de datos realizadas por las autoridades competentes o si ha sido autorizado por una autoridad judicial u otra autoridad nacional competente para verificar si se cumplen las condiciones para la divulgación.

A diferencia de las distintas normativas enumeradas *ut supra*, esta directiva no contiene obligaciones concretas cuando en el tratamiento de datos personales se utilizan tecnologías nuevas como es la inteligencia artificial. Por ello, podemos concluir que se trata de obligaciones generalistas en relación con la salvaguarda de datos personales.

Es destacable también en **materia de privacidad y comunicaciones electrónicas** la Directiva 2002/58/CE del Parlamento Europeo y del Consejo, de 12 de julio de 2002, relativa al tratamiento de los datos personales y a la protección de la intimidad en el sector de las comunicaciones electrónicas (1), establece normas para garantizar la seguridad de la información en el marco de comunicaciones electrónicas como internet y la telefonía móvil y fija, así como de sus redes de apoyo. Esta norma tiene un carácter específico para salvaguardar la privacidad y confidencialidad.

Siendo de obligado cumplimiento para los responsables la protección de sus servicios, estos deben:

(i) garantizar que solo acceden a los datos personales las personas autorizadas;

(ii) proteger los datos personales frente a su pérdida o alteración accidental y a otras formas de tratamiento ilícitas o no autorizadas; y

(iii) garantizar la aplicación de una política de seguridad relativa al tratamiento de datos personales.

Sobre todo, el responsable del tratamiento debe garantizar la confidencialidad de las comunicaciones realizadas y la supresión o anonimización cuando los datos personales que son objeto de tratamiento dejen de ser necesarios. El consentimiento del usuario es, en definitiva, otro punto central de la mencionada directiva, siendo esencial para el tratamiento de datos personales en determinadas situaciones como acción comercial o comunicaciones no solicitadas.

En cuanto a la libre circulación de datos no personales en la Unión Europea, es de suma importancia el Reglamento (UE) 2018/1807, relativo a un marco para la **libre circulación de datos no personales en la Unión Europea**. Tiene como finalidad garantizar que los datos no personales electrónicos puedan ser tratados en toda la Unión Europea y su principal característica consiste en establecer una prohibición a las restricciones en cuanto al lugar donde pueden tratarse o almacenarse los datos.

(1) Este apartado estará sujeto a modificaciones con la entrada en vigor de la regulación relativa a ePrivacy.

Esta directiva pretende dar respuesta a los problemas jurídicos que han generado las nuevas tecnologías, cómo la inteligencia artificial, en cuanto al acceso a los datos y su reutilización, la responsabilidad, la solidaridad y la ética de dichos problemas. Definiendo un marco jurídico previsible, exhaustivo y claro para el tratamiento de datos no personales en el mercado de la Unión Europea, tiene un enfoque basado en principios y es lo suficientemente flexible para tener en cuenta nuevas circunstancias que se vayan generando con el paso del tiempo.

Obliga su aplicación en un sentido amplio sin excluir ningún tipo de sistema informático e incluye todos los datos con independencia de su localización, a no ser que por motivos de seguridad se aconseje. Para garantizar la aplicación efectiva del principio de libre circulación de los datos no personales, los Estados miembros tienen la obligación de comunicar inmediatamente a la Comisión cualquier proyecto de acto que introduzca un nuevo requisito de localización de datos o modifique un requisito existente.

Conclusiones y perspectivas de futuro

Como hemos podido observar, las normativas en relación con la privacidad y el gobierno del dato tienen un papel importante en el desarrollo y crecimiento de soluciones basadas en inteligencia artificial, al ser una tecnología basada en datos. No obstante, resulta poco probable que dicha normativa presente un remedio completo a los daños derivados de un mal uso de soluciones de inteligencia artificial.

Para una parte de la doctrina, la norma existente es restrictiva, poco clara o incluso paradójica en lo que respecta a las implicaciones en materia de protección de datos en relación con el ciclo de vida de una solución de inteligencia artificial. De hecho, para una correcta determinación objetiva de la adecuación de los tratamientos, la existencia de avales y medidas para gestionar sus riesgos es necesario un nivel de madurez alto de las soluciones que hacen uso de inteligencia artificial.

Sin embargo, otra parte de la doctrina considera que la normativa actual en materia de protección de datos personales vigente en la Unión Europea, complementada adecuadamente con la correcta aplicación constante de los principios éticos, plantea un marco favorable de acción para la ideación, el desarrollo y, en general, la evolución de soluciones éticas de inteligencia artificial en todo su ciclo de vida con pleno respeto al principio ético de privacidad y gobierno del dato.

Seguridad y protección. Fiabilidad, robustez y precisión

Definición descriptiva del principio

Como recientemente ha recordado ENISA (*European Union Agency for Cibersecurity* o Agencia de la Unión Europea para la Ciberseguridad), más allá de las ciberamenazas generales, se dan una serie de amenazas más particulares para los entornos de IA (1), como el envenenamiento de datos (el atacante altera los datos o el modelo para modificar el comportamiento del algoritmo en una dirección elegida). Los ataques también corrompen las etiquetas de los datos de entrenamiento, exfiltran o fugan datos o los modelos y algoritmos. Igualmente pueden quedar comprometidos componentes o herramientas de desarrollo de la solución de IA. También los fallos de la solución de IA y los errores humanos son amenazas habituales, así como problemas de denegación de servicio debido a datos inconsistentes, o que no se comuniquen a tiempo incidentes de ciberseguridad.

Frente a estas amenazas, los controles de seguridad convencionales deben complementarse con controles de seguridad específicos, como los que apunta ENISA (2). Ahora bien, se trata de una materia novedosa y los controles aún no están validados ni estandarizados en cuanto a la forma de aplicarse. Respecto de la IA, la Unión Europea deriva hacia un "concepto ampliado de seguridad a fin de proteger a los consumidores y usuarios", así, "los daños pueden ser tanto materiales (para la seguridad y la salud de las personas, con consecuencias como la muerte, y menoscabos al patrimonio) como inmateriales (pérdida de privacidad, limitaciones del derecho de libertad de expresión, dignidad humana, discriminación en el acceso al empleo, etc.) y pueden estar vinculados a una gran variedad de riesgos." A partir de tal concepto amplio, se pretende construir un "Ecosistema de confianza" en la UE, especialmente a través de la "seguridad en el diseño" y a través del cumplimiento de las normas de la UE. El marco regulador debe centrarse en cómo minimizar los distintos riesgos reduciendo la materialización de estos, especialmente los más significativos (3).

El principio de seguridad, incluida la seguridad en sentido estricto, la protección, la fiabilidad, la robustez y la precisión, debe ser entendido en su doble vertiente: por un lado, la seguridad en el funcionamiento interno de soluciones IA y su fiabilidad para no comprometer su entorno; y, por otro, la capacidad de la IA de resistir ante amenazas y vulnerabilidades externas.

El efectivo cumplimiento del principio de seguridad y protección exige una aproximación preventiva basada en el riesgo tanto con carácter previo al desarrollo de soluciones IA como durante su ciclo de vida, de forma que estos sistemas se comporten como inicialmente fueron previstos, se minimice cualquier daño no previsto y no intencionado y se prevea cualquier daño que no resulte aceptable, entendiendo daño como la garantía de la integridad física y mental. Solo así podrá garantizarse que la solución sea segura, fiable y robusta durante todo su ciclo de vida.

(1) ENISA, Malatras, Apostolos, Agrafiotis, Ioannis y Adamczyk, Monika (eds.), Securing machine learning algorithms, ENISA, diciembre 2021, pp. 13-18 DOI: 10.2824/874249

(2) Ibídem, pp. 18-26.

(3) COMISIÓN EUROPEA (2020). Libro Blanco. Sobre la inteligencia artificial... cit. p. 13. Ver su anexo, COMISIÓN EUROPEA Report on the safety and liability implications of Artificial Intelligence, the Internet of Things and robotics. Brussels, 19.2.2020. Report from the Commission to the European Parliament, the Council and the European Economic and Social Committee. Anexo a la COM (2020) 64 final, https://ec.europa.eu/info/publications/commission-report-safety-and-liability-implications-ai-internet-things-and-robotics-0_en p. 3.

El principio de seguridad, incluida la seguridad en sentido estricto, la protección, la fiabilidad, la robustez y la precisión, debe ser entendido en su doble vertiente: por un lado, la seguridad en el funcionamiento interno de soluciones IA y su fiabilidad para no comprometer su entorno; y, por otro, la capacidad de la IA de resistir ante amenazas y vulnerabilidades externas.

El efectivo cumplimiento del principio de seguridad y protección exige una aproximación preventiva basada en el riesgo tanto con carácter previo al desarrollo de soluciones IA como durante su ciclo de vida, de forma que estos sistemas se comporten como inicialmente fueron previstos, se minimice cualquier daño no previsto y no intencionado y se prevea cualquier daño que no resulte aceptable, entendiendo daño como la garantía de la integridad física y mental. Solo así podrá garantizarse que la solución sea segura, fiable y robusta durante todo su ciclo de vida.

El principio de seguridad comprende la seguridad en sentido de evitación (*safety*, en inglés), y la seguridad en sentido de protección (*security*, en inglés), así como la precisión (accuracy, en inglés) y la confiabilidad y reproducibilidad de los resultados de la IA (*reliability and reproducibility*, en inglés):

Evitación (*Safety*). Aplicada a la protección o prevención del daño del propio funcionamiento del sistema hacia los usuarios y su entorno, debiendo aplicar fuertes medidas para evitar la producción de daños tanto exógenos como endógenos.

Protección y resiliencia (*Security and resilience*). Hace referencia a la protección para prevenir el daño, sobre todo, ante ataques y vulnerabilidades externas, intencionadas y no intencionadas, previstas y no previstas. Esta engloba, por tanto, la robustez y resiliencia del sistema.

Precisión (*Accuracy*). La precisión hace referencia a la necesidad de los sistemas de IA de realizar actos (clasificar, organizar, etc.), predicciones, recomendaciones o tomar decisiones de forma correcta, basados en los datos introducidos, para lo cual se requiere un proceso de desarrollo y evaluación robusto que pueda apoyar, mitigar y corregir aquellos riesgos que se deriven de la falta de precisión de las acciones llevadas a cabo por la IA.

Confiabilidad y reproducibilidad (*Reliability and reproducibility*). Hace referencia a la necesidad de que los resultados de los sistemas de IA sean confiables y reproducibles en cualquier tipo de situación.

Por último, este principio está íntimamente ligado con el principio de "Responsabilidad y rendición de cuentas (1)" desarrollado en esta Guía.

(1) Esta relación ha sido subrayada por las instituciones a nivel europeo desde el inicio (por ejemplo, en las "*Ethics Guidelines for Trustworthy AI*", del Grupo de expertos de alto nivel en AI, publicado en abril 2019).

Desarrollo normativo

El ordenamiento jurídico incluye normas que obligan a tener en cuenta este principio en las diferentes fases del proceso creativo de una solución de IA. En algunos casos hablamos de normas que obligan de manera imperativa y, en otros, de meras guías, directrices o recomendaciones, pero en cualquier caso son parte del cuerpo normativo que impulsa la creación y el uso de una IA segura, precisa, fiable y robusta.

Una de las principales características de este principio es la **seguridad de las redes y sistemas de información**, esencial en soluciones IA puesto que los sistemas (eminentemente, los programas de ordenador o software) son la parte esencial de la solución. En este sentido, cabe destacar la Directiva (UE) 2016/1148, de 6 de julio de 2016, relativa a las medidas destinadas a garantizar un elevado nivel común de seguridad de las redes y sistemas de información (conocida como "Directiva NIS I") y que impulsa la creación, en los diferentes estados de la Unión Europea, de una red CSIRT (*Computer Security Incident Response Team* o Equipo de Respuesta ante Incidencias de Seguridad Informáticas).

Se entiende por "seguridad de las redes y sistemas de información": la capacidad de las redes y sistemas de información de resistir, con un nivel determinado de fiabilidad, toda acción que comprometa la disponibilidad, autenticidad, integridad o confidencialidad de los datos almacenados, transmitidos o tratados, o los servicios correspondientes ofrecidos por tales redes y sistemas de información o accesibles a través de ellos;" (art. 4. 2º) Las soluciones IA deben ir acordes a esta definición. Los sistemas de IA no solo deberán atender a la seguridad de los sistemas de información, sino también deberá considerar la normativa general y específica en materia de seguridad (vertiente *safety*), como es la Directiva relativa a la seguridad general de los productos, o la Directiva sobre máquinas (1), ligadas al principio de responsabilidad y a los problemas de responsabilidad vinculados a estos sistemas.

Por su parte, el Reglamento (UE) 2019/881 considera **ciberseguridad** al conjunto de actividades necesarias para la protección de las redes y los sistemas de información, los usuarios de estas y otros afectados por las ciberamenazas; y define la **ciberamenaza** como una circunstancia, un evento o una acción potencial capaz de dañar, interrumpir o afectar de forma adversa a las redes y a los sistemas de información, así como a sus usuarios y otras partes interesadas.

Es fundamental tener en cuenta tanto la Directiva NIS directiva como la normativa nacional de transposición (2) a la hora de desarrollar soluciones IA. Estas normas delimitan las entidades que prestan servicios esenciales para la comunidad y dependen de las redes y sistemas de información y se identifican los principales operadores que prestan dichos servicios. Estas entidades deben adoptar medidas proporcionadas a los niveles de riesgo basadas en una evaluación previa de los mismos. La notificación de incidentes es clave, aunque no hayan tenido un efecto real, también para crear cultura de gestión de riesgos. Hay una plataforma común de notificación que también podrá ser empleada para la notificación de vulneraciones de la seguridad de datos personales. El sistema citado de notificación es confidencial y a través de él se protege a la entidad notificante y al personal que informe sobre incidentes ocurridos. Las autoridades competentes ejercen, por su parte, las funciones de vigilancia y promueven el desarrollo de las obligaciones.

En 2019 se adoptó el Reglamento (UE) 2019/881 del Parlamento Europeo y del Consejo, relativo a ENISA y a la certificación de la ciberseguridad de las tecnologías de la información y la comunicación. ENISA emite de forma recurrente guías e informes, y según mencionado, en diciembre de 2021 lanzó la primera específica para IA. Por otra parte, se pretende aumentar la confianza y la seguridad en los productos, servicios y procesos TIC evitando la multiplicación de esquemas de certificaciones nacionales contradictorios, redundantes y, en cualquier caso, con una eficacia limitada al territorio de validez del certificado, ya que el objetivo principal de este marco jurídico es instaurar esquemas europeos de certificación. Ello puede tener especial interés en el ámbito IA en el que van a desarrollarse modelos de certificación.

Recientemente, se ha adoptado el Reglamento (UE) 2021/887, de 20 de mayo de 2021, por el que se establecen el Centro Europeo de Competencia Industrial, Tecnológica y de Investigación en Ciberseguridad y la Red de Centros Nacionales de Coordinación. También hay que hacer seguimiento de la *Joint Cyber Unit*, Unidad Cibernética Conjunta.

(1) Comisión Europea. (19 de febrero de 2020). Informe sobre las repercusiones en materia de seguridad y responsabilidad civil de la inteligencia artificial, el internet de las cosas y la robótica. 2020, de Eur-lex Sitio web: https://eur-lex.europa.eu/legal-content/ES/TXT/?uri=CELEX:52020DC0064

(2) En España, traspuesto por el Real Decreto-ley 12/2018 de 7 de septiembre de seguridad de las redes y sistemas de información y su normativa de desarrollo, el Real Decreto 43/2021, de 26 de enero, por el que se desarrolla el Real Decreto-ley 12/2018, de 7 de septiembre, de seguridad de las redes y sistemas de información (también conocido como "Normativa NIS").

En este apartado, también es importante destacar la Directiva 2013/40/UE de 12 de agosto de 2013, relativa a **los ataques contra los sistemas de información**, así como su normativa nacional de transposición. Esta directiva establece las normas mínimas relativas a la definición de las infracciones penales y a las sanciones aplicables en el ámbito de los ataques contra los sistemas de información. También tiene por objeto facilitar la prevención de dichas infracciones y la mejora de la cooperación entre las autoridades judiciales y otras autoridades europeas competentes. En esta línea, la IA podría utilizarse de forma preventiva ante determinados delitos.

Otro elemento a tener en cuenta, principalmente en el ámbito de las soluciones de IA aplicadas en el sector financiero de la identificación electrónica, es la **seguridad relacionada con las transacciones electrónicas en el mercado interior, relativo a la identificación de las personas**. Con el objetivo de garantizar el correcto funcionamiento del mercado interior, aspirando al mismo tiempo a un nivel de seguridad adecuado de los medios de identificación electrónica y los servicios de confianza, se debe considerar el Reglamento (UE) nº 910/2014 de 23 de julio de 2014 relativo a la identificación electrónica y los servicios de confianza para las transacciones electrónicas en el mercado interior. Este reglamento establece las condiciones en que los estados miembros de la Unión Europea deberán reconocer los medios de identificación electrónica de las personas físicas y jurídicas pertenecientes a un sistema de identificación electrónica notificado de otro estado miembro, así como las normas para el establecimiento de servicios de confianza, en particular para las transacciones electrónicas, y un marco jurídico para el reconocimiento de las firmas electrónicas, los sellos electrónicos, los sellos de tiempo electrónicos, los documentos electrónicos, los servicios de entrega electrónica certificada y los servicios de certificados para la autenticación de sitios web.

En el marco del uso de soluciones de inteligencia artificial para la identificación electrónica y los servicios de confianza para las transacciones electrónicas, se establece la obligación normativa de desplegar un entorno lógico y una serie de herramientas, con las firmas electrónicas y los sellos de tiempo, que aporten seguridad jurídica a los diferentes actores, de forma que la IA sea una solución confiable y tendente a dotar de garantía a las operaciones que se lleven a cabo en el mercado interior de la Unión Europea. Esta definición de entorno y garantías con enfoque territorial opera como catalizadora de la implantación de estándares internacionales de seguridad para las personas, las empresas y los gobiernos, en línea con el principio de seguridad y protección desde un punto de vista ético-normativo.

En relación con el principio **de privacidad y protección de datos**, cabe recordar que el artículo 32 impone que el responsable y el encargado del tratamiento de datos aplicarán medidas técnicas y organizativas apropiadas para garantizar un nivel de seguridad adecuado al riesgo para los derechos y libertades de los interesados. Tales medidas deben adecuarse a las amenazas y riesgos y se han de contextualizar a los costes, tipo de tratamiento. Como recuerda la AEPD (Agencia Española de Protección de Datos), no hay una solución estándar para todos los tratamientos, y mucho menos para aquellos que incluyan un componente de IA (1). De igual modo, hay que tener en cuenta la Directiva 2002/58/CE del Parlamento Europeo y del Consejo, de 12 de julio de 2002, relativa al tratamiento de los datos personales y a la protección de la intimidad en el sector de las comunicaciones electrónicas, así como su normativa de desarrollo e implementación (respectivamente).

Y de especial importancia es la Propuesta de Reglamento del Parlamento Europeo y del Consejo de abril de 2021 por el que se establecen normas armonizadas en materia de inteligencia artificial (ley de inteligencia artificial).

Bajo el enfoque del RGPD y el Libro blanco de IA, se establece mayor imposición de obligaciones y garantías cuanto mayor riesgo implique el tratamiento de datos o la solución de IA. Y ello sucede respecto de las obligaciones de ciberseguridad.

En virtud de esta propuesta, las soluciones de IA de alto riesgo – esta normativa solo recoge esta obligación respecto a aquellos sistemas de IA que sean calificados de alto riesgo, siendo voluntario para el resto de sistemas de IA - deben funcionar de manera consistente durante todo su ciclo de vida y presentar un nivel adecuado de precisión, solidez y ciberseguridad con arreglo al estado de la técnica generalmente reconocido (art. 15), debiéndose comunicar a los usuarios el nivel de precisión y los parámetros empleados para su medición. Según esta propuesta, la solidez técnica y la resiliencia a los riesgos asociados a las limitaciones del sistema (i.e. fallos, incoherencias, errores, imprevistos, etc.), así como la ciberseguridad, son requisitos claves para garantizar la resistencia de los sistemas de IA, debiéndose adoptar medidas adecuadas para resistir a los errores, fallos e incoherencias que pueden surgir en los propios sistemas o en el entorno donde operan, en particular a causa de su interacción con personas físicas u otros sistemas, su prevención, actuación temprana, mitigación y corrección. Esta propuesta se integrará en la legislación sectorial vigente en materia de seguridad.

(1) AEPD, Adecuación al RGPD de tratamientos que incorporan inteligencia artificial. Una introducción, 2020, p. 42.

Este listado no es exhaustivo y, dependiendo del uso de la solución de IA o de su integración en un producto final sujeto a determinada regulación armonizada (i.e. máquinas, juguetes, ascensores, equipo y sistemas de protección para uso en atmósferas potencialmente explosivas, equipos radioeléctricos, equipos a presión, equipo de embarcaciones de recreo, instalaciones de transporte por cable, aparatos que queman combustibles gaseosos, productos sanitarios y productos sanitarios para diagnóstico in vitro) deberá considerarse otros requisitos específicos de seguridad previstos en dichas normativas. Por ejemplo, cuando los sistemas de IA sean componentes de seguridad de otros productos o sistemas, deberán atender también a la normativa específica de cada sector (1).

(1) En particular, al Reglamento (CE) n.º 300/2008 del Parlamento Europeo y del Consejo, de 11 de marzo de 2008, sobre normas comunes para la seguridad de la aviación civil, el Reglamento (UE) n.º 167/2013 del Parlamento Europeo y del Consejo, de 5 de febrero de 2013, relativo a la homologación de los vehículos agrícolas o forestales, y a la vigilancia del mercado de dichos vehículos, el Reglamento (UE) n.º 168/2013 del Parlamento Europeo y del Consejo, de 15 de enero de 2013, relativo a la homologación de los vehículos de dos o tres ruedas y los cuatriciclos, y a la vigilancia del mercado de dichos vehículos, la Directiva 2014/90/UE del Parlamento Europeo y del Consejo, de 23 de julio de 2014, sobre equipos marinos, la Directiva (UE) 2016/797 del Parlamento Europeo y del Consejo, de 11 de mayo de 2016, sobre la interoperabilidad del sistema ferroviario dentro de la Unión Europea, el Reglamento (UE) 2018/858 del Parlamento Europeo y del Consejo, de 30 de mayo de 2018, sobre la homologación y la vigilancia del mercado de los vehículos de motor y sus remolques y de los sistemas, los componentes y las unidades técnicas independientes destinados a dichos vehículos, el Reglamento (UE) 2018/1139 del Parlamento Europeo y del Consejo, de 4 de julio de 2018, sobre normas comunes en el ámbito de la aviación civil y por el que se crea una Agencia de la Unión Europea para la Seguridad Aérea, y el Reglamento (UE) 2019/2144 del Parlamento Europeo y del Consejo, de 27 de noviembre de 2019, relativo a los requisitos de homologación de tipo de los vehículos de motor y de sus remolques, así como los sistemas, componentes y unidades técnicas independientes destinados a esos vehículos, en lo que respecta a su seguridad general y a la protección de los ocupantes de los vehículos y de los usuarios vulnerables de la vía pública.

Conclusiones y perspectivas de futuro.

Cabe concluir los siguientes puntos clave:

I. Las soluciones de IA deben ser robustas y seguras durante todo su ciclo de vida, tanto respecto a la prevención del daño del propio funcionamiento interno del sistema y de la falta de su precisión, como respecto a potenciales ataques y vulnerabilidades externas, intencionadas y no intencionadas, previstas y no previstas.

II. Los actores de la IA deben garantizar la trazabilidad, confiabilidad y reproducibilidad de los resultados, para permitir el análisis de los resultados del sistema de IA y las respuestas a las consultas.

III. Los actores de la IA deben aplicar un enfoque sistemático de gestión de riesgos a cada fase del ciclo de vida de la solución IA de forma continua para abordar los riesgos relacionados con las soluciones IA, incluida la privacidad, la seguridad, la protección y el prejuicio. Para ello, es primordial garantizar la trazabilidad de los datos, procesos y decisiones tomadas por la IA.

Responsabilidad y rendición de cuentas

Definición descriptiva del principio

La responsabilidad y la rendición de cuentas son requisitos fundamentales para garantizar que tanto el desarrollo como la utilización de las soluciones de IA no puedan causar directa o indirectamente algún tipo de daño o perjuicio a un tercero y evitar que se den fallos. Asimismo, ante cualquier posible fallo, este principio permite contar con una visión clara y transparente de la trazabilidad de estos, permitiendo identificar al responsable correspondiente.

Desde un inicio hay que subrayar que este modelo se inspira en los principios reguladores de la protección de datos establecidos en el RGPD, en concreto: la responsabilidad proactiva, la privacidad por defecto y en el diseño. El modelo es el de "más vale prevenir que curar", cuyo objetivo es prever el origen de los posibles riesgos, así como la intensidad de estos. Con carácter previo al desarrollo de soluciones IA, es necesario establecer una serie de controles destinados a mitigar posibles riesgos técnicos y humanos que se puedan dar a nivel organizativo en el desarrollo de este tipo de soluciones. Igualmente, durante el desarrollo de soluciones IA el rol asumido por los diseñadores y desarrolladores, así como los miembros directivos, es particularmente relevante. A este respecto, en aras de garantizar que todos los intervinientes den cumplimiento a este principio, es necesaria tanto la asignación de responsabilidades y la formalización de ámbitos de actuación de los intervinientes en el desarrollo, como la definición de metodologías de trabajo. De tal manera, mediante la definición y despliegue de este tipo de medidas, las organizaciones tendrán la capacidad de identificar posibles incumplimientos y/o áreas de mejora.

En este nuevo modelo pasa a ser esencial el cumplimiento ético y normativo en el diseño, la responsabilidad proactiva, la privacidad (o ética) en el diseño y por defecto, la responsabilidad demostrada (que se pueda probar que se han ido adoptando todas las medidas y cumpliendo todas las obligaciones) y las evaluaciones de impacto. Igualmente, es relevante la figura de los delegados de protección de datos (o futuros responsables de cumplimiento ético y normativo de IA) en el sector público y grandes empresas.

Desarrollo normativo

A **nivel regulatorio**, el Derecho de la Unión Europea, así como el de sus Estados Miembros, velan por el cumplimiento de los principios de responsabilidad y rendición de cuentas ante cualquier producto o servicio que pueda ocasionar un perjuicio en las personas físicas de forma general, **sin que, hasta el momento, exista ningún tipo de exigencia normativa específica en relación con soluciones IA**. Las políticas de protección y defensa del consumidor a nivel comunitario tienen su base jurídica en los artículos 4.2.f), 12, 114 y 169 del Tratado de Funcionamiento de la Unión Europea (TFUE), y en el artículo 38 de la Carta de los Derechos Fundamentales de la Unión Europea.

A este respecto, el objetivo de la Unión Europea y de sus Estados Miembros, pasa siempre por garantizar un alto nivel de protección a los consumidores y usuarios. Las medidas europeas de protección de este colectivo tienen por objeto proteger la salud, la seguridad y los intereses económicos y jurídicos de los consumidores europeos, así como promover su derecho a la información, a la educación y salvaguardar sus intereses con independencia del lugar en el que residan. La normativa europea pretende regular tanto las transacciones físicas como electrónicas e incluye tanto obligaciones de aplicación general, como disposiciones dirigidas a productos específicos.

Si bien, en el Espacio Económico Europeo existe normativa que regule la responsabilidad y la rendición de cuentas con carácter general, estas normas no han sido desarrolladas en el campo de la IA. La incertidumbre generada por esta falta de regulación se ve reforzada por la inexistencia de desarrollo doctrinal o jurisprudencial relevante al respecto. Consecuentemente, en la actualidad los desarrolladores operan en el marco de la IA bajo una patente inseguridad jurídica debido a la inexistencia de unos principios reguladores que definan de manera clara las obligaciones y los límites a tener en cuenta. Y en buena medida esta falta de regulación la está supliendo la aplicación de la normativa de protección de datos en los casos en los que la IA implica un tratamiento de datos personales. Asimismo, y como se ha señalado, la futura normativa de IA se basa en el modelo regulatorio de la protección de datos.

En relación con la **responsabilidad por los daños causados** por productos defectuosos, los diseñadores o desarrolladores son responsables de los daños causados por los defectos de sus productos. Por ello, las personas y organizaciones responsables de la creación y aplicación de las soluciones IA deben ser siempre identificables, siendo capaces de rendir cuentas ante los impactos negativos de ese algoritmo, incluso si son involuntarios.

Esta responsabilidad, se encuentra recogida en la Directiva (UE) 85/374/CEE del Consejo de 25 de julio de 1985, relativa a la aproximación de las disposiciones legales, reglamentarias y administrativas de los Estados Miembros en materia de responsabilidad por los daños causados por productos defectuosos. A este respecto, debemos considerar que se trata de una normativa de 1985 y, por tanto, sus conceptos se encuentran actualmente desfasados, lo que ocasiona grandes dificultades a la hora de aplicarla a los productos de la economía digital en general.

Las **legislaciones actuales** en materia de responsabilidad por productos defectuosos descansan sobre los patrones clásicos de responsabilidad por culpa o negligencia, es decir, el fabricante solo responde cuando es posible demostrar algún grado de culpa o negligencia en el proceso de fabricación y el daño producido. El problema se plantea cuando una solución de IA, como consecuencia del aprendizaje automático (o *machine learning*), toma decisiones propias que no provengan de una orden expresa humana. Ante esta situación, ¿quién sería responsable, el desarrollador o el software?, ¿puede imputarse responsabilidad a un ente sin conciencia o capacidad de distinguir entre el bien y el mal?

Desde las instituciones europeas se está tratando de dar solución a esta problemática; y, sin embargo, todavía no contamos con legislación vigente al respecto. Es claro que en nuestro actual marco jurídico las soluciones de IA no pueden ser responsables por actos u omisiones que causen daños a terceros. Por ello la UE en su Resolución, de fecha 16 de febrero de 2017, que contiene recomendaciones destinadas a la Comisión sobre normas de Derecho Civil sobre robótica 2015/2103 (INL), recomienda establecer un **régimen de seguro obligatorio de responsabilidad civil** para los fabricantes o desarrolladores ante los posibles daños o perjuicios que pueda ocasionar el sistema de IA.

En octubre de 2020, el Parlamento Europeo dictó una Resolución en la que se contienen recomendaciones a la Comisión sobre el régimen de responsabilidad civil en materia de IA (2020/2014/INL), en este caso el Parlamento considera responsable ante un posible daño o perjuicio que cause el sistema o la actividad que lleve a cabo por su cuenta a través del aprendizaje, a la persona que cree, mantenga, controle o explote la solución de IA.

La mencionada Resolución diferencia entre la responsabilidad objetiva de las soluciones de IA de alto riesgo y la responsabilidad subjetiva del resto de soluciones de IA. El operador de una solución de IA de alto riesgo será objetivamente responsable de cualquier daño o perjuicio causado por una actividad física o virtual, un dispositivo o un proceso gobernado por dicho sistema de IA; sin embargo, el operador de una solución de IA que no esté catalogada como de alto riesgo responderá de manera subjetiva y no será responsable si puede demostrar que no tuvo culpa en el daño o perjuicio causado basándose en alguno de los siguientes motivos:

- La solución de IA se activó sin su conocimiento y se tomaron todas las medidas razonables y necesarias para evitar dicha activación.
- Se observó la diligencia debida a través de la selección de una solución de IA adecuada para las tareas y las capacidades pertinentes, la correcta puesta en funcionamiento de la solución, el control de las actividades y el mantenimiento de la fiabilidad a través de la instalación periódica de todas las actualizaciones disponibles.

Por otro lado, en materia de **protección de datos personales**, el principio de responsabilidad proactiva, recogido en los artículos 5 y 24 del Reglamento General de Protección de Datos, impone al responsable del tratamiento la obligación de tratar los datos personales de conformidad con las previsiones recogidas en la normativa europea y, además, debe ser capaz de demostrarlo en todo momento. De acuerdo con el principio anteriormente descrito, en el marco del desarrollo de una IA debe tener en cuenta la naturaleza, el ámbito, el contexto y los fines del tratamiento, así como los riesgos de diversa probabilidad y gravedad para los derechos y libertades de las personas físicas. En este sentido, el responsable debe haber implementado sobre la solución las medidas técnicas y organizativas apropiadas para garantizar y demostrar que el tratamiento es conforme con dicha normativa.

En este orden, resulta relevante el continuo crecimiento en el despliegue de **sistemas de reconocimiento facial** por parte de autoridades gubernamentales, sujetas a los requisitos de licitud y lealtad previstos en la Directiva (UE) 2016/680, de 27 de abril, relativa a la protección de las personas físicas en lo que respecta al tratamiento de datos personales por parte de las autoridades competentes para fines de prevención, investigación, detección o enjuiciamiento de infracciones penales o de ejecución de sanciones penales, y a la libre circulación de dichos datos. Así, los Estados miembros deben garantizar que los datos se traten con fines determinados, explícitos y legítimos, que sean adecuados, pertinentes y no excesivos, a que sean exactos, que se conserven durante un período no superior al necesario y que sean tratados de tal manera que se garantice una seguridad adecuada.

Las obligaciones anteriores, se ven reforzadas por el reciente marco ético aprobado por la UNESCO en noviembre de 2021, así como Resolución del Parlamento Europeo, de 6 de octubre de 2021, sobre la inteligencia artificial en el Derecho penal y su utilización por las autoridades policiales y judiciales en asuntos penales. En ambos documentos se establecen algunas recomendaciones para los Estados como, por ejemplo, que las autoridades sean transparentes e informen cuándo están trabajando con agencias de reconocimiento facial, o que estos sistemas no se utilicen con fines de vigilancia masiva o rendición de cuentas sociales.

Ante esta situación, la adopción de las normas de responsabilidad a la era digital y a la IA, han sido recientemente sometidas a consulta. La consulta (periodo de consulta: 18 de octubre de 2021 y el 10 de enero de 2022) tuvo como objetivo analizar con detalle los problemas relacionados con determinados tipos de IA, que dificultan la identificación de los responsables, la verificación de su culpabilidad o la demostración del defecto de un producto y la relación causal con el daño.

En la misma línea, el Informe de la Comisión al Parlamento Europeo, al Consejo y al Comité Económico y Social Europeo sobre las repercusiones en materia de seguridad y responsabilidad civil de la IA, el internet de las cosas y la robótica destaca que si bien, en principio, la normativa en vigor de la Unión y sus Estados Miembros en materia de responsabilidad puede hacer frente a las vicisitudes jurídicas que plantean las tecnologías emergentes, las particularidades de las soluciones IA requieren de una regulación adicional a nivel local.

Conclusiones y perspectivas de futuro

I. Es importante que las legislaciones futuras junto con la doctrina y la jurisprudencia a nivel comunitario aclaren las cuestiones de responsabilidad, culpabilidad, obligación y rendición de cuentas de las soluciones de IA.

II. Deben crearse sistemas de registro que garanticen la trazabilidad desde el diseño y durante el funcionamiento de la solución, permitiendo la delimitación de responsabilidades entre los distintos actores intervinientes (desarrolladores, operadores, propietarios y usuarios).

III. Deben establecerse sistemas eficaces de mitigación de daños.

IV. Deben proporcionarse oportunidades apropiadas para la retroalimentación, las explicaciones pertinentes y la apelación.

V. Resulta fundamental comprender el alcance de la responsabilidad de la empresa o de la persona sobre la IA.

Transparencia y explicabilidad

Definición descriptiva del principio

El principio de transparencia y explicabilidad aplicado a la inteligencia artificial se define como la cualidad de una solución de IA de **poder ser enteramente comprendida por un ser humano**, desde la forma en la captación y tratamiento los datos, sean o no personales, hasta la manera en la que se toman las decisiones, así como las consecuencias de estas y sus interdependencias. Este principio garantiza que la persona tenga y mantenga el **control informado** global sobre la existencia y la actividad de una solución de inteligencia artificial.

La transparencia incluye el acceso al funcionamiento y resultados, explicabilidad, trazabilidad y auditabilidad del uso público y privado de la IA. Se puede traducir también en datos y algoritmos de código abierto, contratación pública abierta e información de cuándo se interactúa con una IA o cuándo adopta decisiones sobre las personas.

La transparencia, explicabilidad e interpretabilidad de los algoritmos debe **vigilarse en todo el ciclo de vida de las soluciones de IA**, lo que indudablemente incluye tanto el modelo en sí como los datos utilizados. Si una solución de IA causa algún tipo de daño, la persona debe poder conocer de manera transparente y sencilla qué ha pasado y por qué ha sucedido. Por lo tanto, estos principios deben garantizarse tanto desde la concepción de la propia solución en sus fases iniciales y de desarrollo, como en las de despliegue y aplicación en el sector de que se trate, de forma que tanto diseñadores, desarrolladores, implementadores, supervisores y responsables de estas soluciones (en cualquiera de sus fases y momento del ciclo de vida) deben estar obligados a velar por ellos.

La **transparencia** debe ser inherente a las soluciones de IA debido a las potenciales **decisiones** que esta pueda tomar (por ejemplo, sobre la concesión de beneficios o subvenciones, sobre el acceso a recursos públicos o privados o sobre decisiones judiciales o administrativas, entre otras). Por esta razón, toda participación de una tecnología de decisión autónoma, en sus diferentes niveles de autonomía, debe poder ser **explicada y auditada** por parte de una persona humana. Por ello se deben idear y generar controles para medir de forma empírica el nivel de transparencia de la tecnología y asegurar que la solución cuenta con mecanismos adecuados para explicar su toma de decisiones.

La **explicabilidad** de la tecnología hace referencia a la capacidad de explicar en términos humanos la toma de decisiones de la inteligencia artificial. Del conjunto de códigos que resulten de todo el proceso de toma de decisión de este tipo de soluciones, personas expertas en la materia deben ser capaces de inferir su significado para poder **"traducirlo" a un lenguaje humano**. Es importante señalar que no cualquier persona estará capacitada para comprender las decisiones de una concreta solución IA, ya sea por la naturaleza de sus decisiones o por la complejidad técnica de sus algoritmos. Pero, en cualquier caso, una persona humana con conocimientos expertos sí debe ser capaz de llegar a comprender y dominar lo que una solución de IA realice, vaya a realizar o pueda llegar a realizar, en el más amplio sentido. Esto es determinante no solo en términos de transparencia, sino también a efectos del propio desarrollo y promoción de soluciones de IA, ya que la transparencia y explicabilidad permitirán a los ingenieros y científicos especializados en la materia mantener un **diálogo constructivo** orientado a la **mejora continua de la tecnología** y la creación de una cultura de cooperación y confianza.

Desarrollo normativo

El principio de transparencia es uno de los principios para el que hoy encontramos respuesta en la **normativa** vigente aplicable a las soluciones de IA. La normativa de protección de datos **sienta las bases del principio de transparencia y explicabilidad**. Garantizando la transparencia de las soluciones de IA es posible la aplicación de principios como la rendición de cuentas, la explicabilidad, la robustez y fiabilidad y la vigilancia humana. Así, este principio está recogido en la norma europea como uno de los derechos del interesado y, como se recoge en el artículo 12 RGPD, obliga al responsable del tratamiento a facilitar la información sobre la forma en que los datos personales serán usados.

La especial transparencia y explicabilidad aplicable a las decisiones solo algorítmicas se detalla en los artículos 22 y en los particulares deberes de transparencia. Así, se da la concreta obligación de facilitar "información significativa sobre la lógica aplicada, así como la importancia y las consecuencias previstas de dicho tratamiento para el interesado" (art. 13. 2º f y 14. 2º g) RGPD). De igual modo, el derecho de acceso permite solicitar dicha información (artículo 15. 1º. h RGPD). Además, el G29-UE (2018: 35) (1) detalla la información que debe facilitarse: las categorías de datos que se han utilizado o se utilizarán en la elaboración de perfiles o el proceso de toma de decisiones; por qué estas categorías se consideran pertinentes; cómo se elaboran los perfiles utilizados en el proceso de decisiones automatizadas, incluidas las estadísticas utilizadas en el análisis; por qué este perfil es pertinente para el proceso de decisiones automatizadas; y cómo se utiliza para una decisión relativa al interesado. El G29-UE, recomienda también informar en general respecto de toda decisión automatizada, aunque no sean las protegidas por el artículo 22 y este derecho G29-UE).

Adicionalmente, en la toma de decisiones únicamente algorítmicas de relevancia, la AEPD ha concretado las obligaciones de transparencia aplicables (2). Estas obligaciones incluirían: el detalle de los datos empleados para la toma de decisión, más allá de la categoría, y en particular información sobre los plazos de uso de los datos (su antigüedad y plazo de conservación); la importancia relativa que cada uno de ellos tiene en la toma de decisión; la calidad de los datos de entrenamiento y el tipo de patrones utilizados; los perfilados realizados y sus implicaciones; valores de precisión o error según la métrica adecuada para medir la bondad de la inferencia; la existencia o no de supervisión humana cualificada; la referencia a auditorías, especialmente sobre las posibles desviaciones de los resultados de las inferencias, así como la certificación o certificaciones realizadas sobre el sistema de IA. En el caso de encontrarse ante sistemas adaptativos o evolutivos, es necesario también aportar los datos de la última auditoría realizada. Finalmente, en el caso de que el sistema IA contenga información de terceros identificables, existe la prohibición de tratar esa información sin legitimación, detectando las consecuencias de realizarlo.

Por otro lado, tal como ha sido expuesto en el apartado primero, la Directiva (UE) 2016/680 también contiene previsiones sobre transparencia en lo que respecta al tratamiento de los datos personales por parte de las autoridades, relacionado con el **ámbito penal y policial y la libre circulación de esos datos**. Por el potencial uso que se le puede dar a las soluciones de IA en el ámbito de la justicia y el impacto que puede tener para la libertad de los interesados, la utilización de esta tecnología y qué tratamiento de los datos hacen debe ser transparente, lo que afecta a las autoridades, pero también a los desarrolladores, implementadores o cualesquiera corresponsables o encargados que se designen para las diferentes fases de vida de la IA.

Igualmente, en la Propuesta de Reglamento del Parlamento Europeo y del Consejo 2021/0106 (conocida popularmente como "**Ley de Inteligencia Artificial**") también se percibe la preocupación del legislador europeo por la transparencia de las soluciones de IA en todo su ciclo de vida. Todos los agentes implicados deben tenerla en cuenta tanto en el ejercicio de su actividad como a la hora de poder llevar a cabo las auditorías necesarias que garanticen su control y rendición de cuentas. Así, en esta propuesta podemos encontrar una serie de iniciativas que podrían **acercar la realidad normativa al ideal del principio de transparencia**.

(1) G29-UE. (2018). Directrices sobre decisiones individuales automatizadas y elaboración de perfiles a los efectos del Reglamento 2016/679, 3 de octubre de 2017, versión final 6 de febrero de 2018, Doc WP251rev.01

(2) AEPD, Adecuación al RGPD de tratamientos que incorporan inteligencia artificial. Una introducción. Febrero de 2020, pág. 24.

En primer lugar, se prevé una importante **obligación de información a los interesados y registro** de toda la documentación relativa a las soluciones de IA, entre lo que se incluye la información técnica, obligaciones especialmente enfatizadas y ampliadas para soluciones destinadas a interactuar con personas físicas o que generen contenido (o lo manipulen) que pueda asemejarse a personas, objetos o lugares y puedan inducir a error sobre la naturaleza de la interacción. Además, el registro de toda esta documentación implicaría un **control reforzado por parte de las autoridades**, ya que como se recoge en el artículo 64 de Ley de inteligencia artificial, se prevé el acceso por parte de las mismas a estos registros con un amplio alcance a fin de que puedan controlar estas soluciones.

Por otro lado, también relacionado con la **rendición de cuentas**, pero en relación con la obligación de transparencia, la Ley de inteligencia artificial propone la obligación de informar de inmediato cuando se presenten riesgos en determinada solución de IA. Este deber de información incluye la identificación de las medidas que se tomarán para paliarlos, especialmente cuando tenga afectación sobre los derechos fundamentales.

En lo referente a la **explicabilidad**, la futura Ley de Inteligencia Artificial, aún en fase de propuesta, recoge este principio, pero no de una forma tan concreta como sería deseable. La explicabilidad es brevemente mencionada en el considerando 38, y el artículo 14 de la propuesta recoge una serie de requisitos que, aunque no expresamente sobre explicabilidad y sí sobre la vigilancia humana, se exigirían para garantizar que los usuarios, técnicos o responsables de este tipo de soluciones sean capaces de interpretar su información de salida.

En países en los que grandes proveedores de servicios digitales poseen un poder de mercado significativo y un gran control sobre el mismo, como sucede en Estados Unidos, en los últimos años se están **produciendo movimientos tendentes a regular las implicaciones éticas de la IA**. No existe todavía una regulación específica en este sentido, pero sí que se aprecian esfuerzos que podrían resultar próximamente en novedades legislativas. En particular, la FTC (Agencia Federal de Protección al Consumidor) de Estados Unidos ha elaborado y publicado una serie de recomendaciones y aclaraciones sobre la utilización de la IA entre las que se encuentran la preocupación por la transparencia y la explicabilidad de este tipo de soluciones, ya que considera que **solo a través de ellas podrá realizarse la evaluación efectiva de los sesgos de las soluciones de IA** (ya que se entienden como condiciones que puedan garantizar la justicia y no discriminación de esta tecnología).

La FTC recoge que se tendrán que establecer mecanismos de transparencia para abrir las soluciones a la inspección externa. Unido a ello, también prevé que este tipo de tecnología debe incorporar unas instrucciones de explicación a los consumidores cuando la intervención de la tecnología haya sido determinante para la denegación de productos o servicios, la elaboración de ránquines o la modificación de condiciones o contratos en general. Estas explicaciones deben ser específicas y verdaderamente aclaratorias.

Conclusiones y perspectivas de futuro

El principio de transparencia y explicabilidad es la base sobre la que se construye la confianza en la inteligencia artificial. El avance tecnológico es afortunadamente inevitable y la normalización de este principio en relación con las soluciones de IA se espera que también lo sea, quizá en un principio impulsado por normas que obliguen a ello, pero, en cualquier caso, se acabará imponiendo por ser la única vía admisible para que la IA siga cumpliendo su función bajo el **control informado del ser humano**.

El principio de justicia

Definición descriptiva del principio

El principio de justicia, que debe garantizar la no discriminación, la diversidad y la inclusión, es el más mencionado en los documentos sobre Ética de la Inteligencia Artificial de los distintos organismos e instituciones a nivel europeo e internacional. En la totalidad de las Guías y Recomendaciones sobre una IA ética y confiable, aparecen los principios de equidad y no discriminación (1).

El principio se compone de varios elementos en torno a la justicia y la equidad. Su contenido (2) es muy amplio e inclusivo de muchos otros principios. Bajo el principio de justicia, se debe asegurar que el uso de la IA crea beneficios que se comparten (o al menos se pueden compartir) y que la IA previene la creación de nuevas injusticias, como el socavamiento de las estructuras sociales existentes.

El principio de justicia también implica altos estándares de responsabilidad y la reparación efectiva o el remedio si se produce un daño, así como la evaluación comparativa del rendimiento. Al mismo tiempo, el desarrollo, despliegue y utilización de soluciones de IA debe ser equitativo para que esta sea fiable (Directrices Éticas para una IA fiable) (3), puesto que la equidad es sinónimo de estabilidad social y de justicia (*Ethics of AI* de la Universidad de Helsinki) (4). El principio de justicia se relaciona con la mejor distribución de recursos, en algunos usos de la IA como el de salud, implica que haya opciones de tratamiento nuevas y experimentales o simplemente la disponibilidad general de la atención médica.

Por otro lado, la justicia y la equidad evitan la **discriminación** y pueden promover la **inclusión**. Cuando se utiliza la inteligencia artificial se debe evitar que los algoritmos supongan sesgos (ni en discrepancia de una estadística, ni en ciencia cognitiva, ni en justicia social) y estos a su vez, impliquen el incumplimiento de estos principios. Las soluciones de IA deben ser diseñadas para maximizar la equidad y promover la inclusión (Principios de Harvard).

La justicia en inteligencia artificial implica que todas las personas deben ser tratadas como iguales y/o con equidad (5), de acuerdo con el principio de igualdad de trato que implica la ausencia de discriminación. Alcanza al uso de la IA para la distribución de recursos y acceso equitativo; para “eliminar todos los tipos de discriminación” (Montreal, 2017 (6). Así, se debe usar la IA para corregir errores pasados, como eliminar la discriminación injusta y las discriminaciones históricas, evitar sesgos, estigmatización y discriminación. Pero también se trata de que se generen “beneficios compartidos” y “prosperidad compartida” (según el texto mencionado con anterioridad, *Asilomar Principles* (7)). También implica evitar nuevas discriminaciones sociales. Asimismo, la necesidad de diversidad implica que se tomen en consideración las diferencias culturales, lingüísticas, raciales, de género o de educación existentes para favorecer la inclusión y para no ejercer la discriminación.

Asimismo, se busca garantizar un acceso equitativo a la IA y a sus resultados, frente al riesgo de que puedan aumentar las diferentes desigualdades sociales, territoriales y regionales en el mundo. La misma riqueza generada por la IA debe beneficiar a la sociedad en su conjunto, también de modo equitativo entre Estados y regiones evitando la generación de brechas. El acceso a la IA debe implicar también la formación para la adaptación a los cambios de todos los colectivos, obviamente entre ellos los más afectados por la brecha digital, así como los habitualmente excluidos o marginados. Se subraya también la importancia de promover la educación para comprender, utilizar y desarrollar la IA, así como la accesibilidad para colectivos especiales como pueden ser las personas con diferentes capacidades.

(1) Fjeld, Jessica; et. al. Principled Artificial Intelligence: Mapping Consensus in Ethical and Rights-based Approaches to Principles for AI. Berkman Klein Center for Internet & Society Research at Harvard University. January, 2020.

(2) (AI-HLEG: 10; Floridi et al, 2018: 698).

(3) Independent High-Level Expert Group on Artificial Intelligence. Ethics Guidelines for Trustworthy AI. April, 2019.

(4) University of Helsinki. Ethics of AI. Chapter 6: Fairness.

(5) Aunque en términos semánticos no siempre igualdad y equidad son sinónimos, en algunos documentos legales analizados sí aparecen como tales.

(6) The Forum on the Socially Responsible Development of Artificial Intelligence, Montreal Declaration, 2017.

(7) Future of Life Institute. (2020). ASILOMAR AI PRINCIPLES. 2020, de Future of Life Institute Sitio web: https://futureoflife.org/2017/08/11/ai-principles/

Desarrollo normativo

La proyección a los sistemas IA de la normativa antidiscriminatoria y de protección de datos

Pese a la enorme mejora en el último siglo, no podemos desconocer que vivimos en sociedades sesgadas y discriminatorias. Y los algoritmos y los datos de los que se alimentan y aprenden las soluciones de IA son productos humanos que necesariamente integran los sesgos y discriminaciones de colectivos y grupos tradicionalmente discriminados. De este modo se perpetúan e incluso agravan situaciones de desventaja y de discriminación. Asimismo, la discriminación y el sesgo pueden fácilmente producirse por la naturaleza y características habituales en las que se desarrollan las decisiones automatizadas: la mala calidad y gestión de los datos o la opacidad del sistema dificultan más si cabe la resolución de estos problemas. Es por ello que todos los mandatos generales y particulares de no discriminación deben ser proyectados al diseño, desarrollo, distribución y uso de soluciones de IA.

El artículo 14 y el protocolo nº 12 del Convenio Europeo de Derechos Humanos (CEDH) sobre no discriminación, y a partir de ellos, toda la jurisprudencia del Tribunal Europeo de Derechos Humanos (TEDH) se establecen como una viga maestra del Derecho europeo de la igualdad y la no discriminación. Así, los principios de igualdad y justicia están reconocidos como tradiciones constitucionales comunes de todos los Estados de la Unión Europea que, junto con los derechos fundamentales, son principios generales del Derecho de la Unión. En el artículo 2 del Tratado de Funcionamiento la Unión Europea (TFUE) se sientan los fundamentos para una protección de la dignidad humana y de la igualdad en un conjunto de Estados en donde prevalecerá la no discriminación, el pluralismo y la tolerancia, entre otros valores ligados al de la justicia. El Tratado establece asimismo en su artículo 3 que la UE combatirá la exclusión social, fomentando la justicia e igualdad, siempre respetando la riqueza de la diversidad cultural y lingüística que le es inherente, y en su artículo 9, garantiza que, con fundamento en el principio de igualdad, los ciudadanos se beneficiarán por igual de la atención de sus instituciones y organismos. Finalmente, en el artículo 10, el Tratado, establece que la Unión llevará a cabo todos sus políticas y actos bajo el principio de no discriminación.

La Carta de Derechos Fundamentales de la Unión en sus artículos 20 a 26 hace mención explícita a la protección y garantía de: i) la igualdad ante la ley, ii) la prohibición de toda discriminación por orígenes, características, nacionalidad o preferencias, iii) el respeto a la diversidad cultural, religiosa y lingüística, iv) la igualdad entre hombres y mujeres, v) los derechos del menor, vi) los derechos de las personas mayores, y; vii) la integración de las personas discapacitadas.

Como reflejo de estos principios generales y derechos fundamentales del Derecho de la UE, hay numerosa normativa antidiscriminación más general, particular o sectorial. Entre otras, en el ámbito de la igualdad de trato en el empleo, la ocupación, la formación (Directiva 200/78/CE del Consejo, de 27 de noviembre de 2000), igualdad respecto del origen racial o étnico (Directiva 2000/43/CE del Consejo, de 29 de junio de 2000), por cuanto al acceso a bienes y servicios (Directiva 2004/113/CE del Consejo, de 13 de diciembre), en materia de seguridad social (Directiva 79/7/CEE del Consejo, de 19 de diciembre de 1978) o, entre otras, respecto del permiso parental y conciliación de la vida laboral con la personal y familiar (Directiva 2010/18/UE del Consejo, de 8 de marzo).

Todos estos y otros instrumentos son plenamente aplicables al uso de las soluciones de inteligencia artificial y a los datos de los que se alimentan o los datos o decisiones que resultan de su uso.

El RGPD y normativa conexa (Directiva (UE) 2016/680, de 27 de abril policial y penal Directiva (UE) 2016/681 y terrorismo y de la delincuencia grave), fortalecen el marco jurídico para lograr la justicia y la no discriminación algorítmica. Esencialmente hay que recordar que los principios de protección de datos (artículo 5 RGPD) imponen que los datos que alimentan las soluciones de IA sean exactos, no sesgados y deben "de forma lícita, justa y transparente". Ello permite derivar obligaciones importantes para evitar la discriminación y el sesgo. No pocas de las 150 cuestiones del checklist en las Directrices para la ética en el diseño en la UE o las Guías de la AEPD sobre IA, establecen concretos lineamientos para para lograr la exactitud y fiabilidad, integridad, calidad de los datos, mecanismos de supervisión, gobernanza y gestión de datos, medidas de control, representatividad de datos con inclusión poblaciones específicas, consideración de afectados indirectos, consultas a colectivos específicos, medición de la imparcialidad, accesibilidad y diseño universal, si el equipo de desarrolladores es representativo.

En cuanto a **los elementos del tratamiento jurídico**. Desde el punto de vista del Derecho antidiscriminatorio, las discriminaciones directas voluntarias y pretendidas no son habituales y resultan además muy difíciles de probar si no se cuentan con todos los elementos del sistema algorítmico. Las STJUE casos Danfoss y Galina Meister orientan negativamente la posibilidad de acceder al código fuente de sistemas algorítmicos que son sospechosos de producir discriminación (1). Sin embargo, habría que ver el caso concreto. Pese a que no se facilitase acceso completo al sistema, será posible experimentar con el sistema para comprobar sus resultados bajo diferentes perfiles. No obstante, será muy difícil poder probar que el sistema está diseñado teniendo en cuenta datos especialmente sospechosos de discriminación.

La solución de IA y algoritmos puede estar diseñado para tener en cuenta circunstancias especialmente prohibidas (sexo, raza, religión, salud, etc.) y solo este hecho ya implica la necesidad de contar con garantías jurídicas reforzadas. Las decisiones algorítmicas que supongan tratamientos diferentes basados en circunstancias prohibidas son especialmente sospechosas de discriminación. Se presume que no son admisibles y quedan sometidas a un escrutinio o test de admisibilidad más riguroso. Desde la normativa de protección de datos, en muchos casos, supondrá un tratamiento de datos especialmente protegidos (art. 9 RGPD), lo cual implica particulares restricciones, prohibiciones y garantías. Y, además, al régimen con especiales garantías de tratamiento de datos especialmente protegidos se podrán superponer las particulares garantías de las decisiones solo automatizadas significativas (art. 22. 3º RGPD-UE).

El artículo 11. 3º de la Directiva (UE) 2016/680 sobre tratamientos automatizados en materia penal y de justicia expresamente prohíbe "La elaboración de perfiles que dé lugar a una discriminación de las personas físicas basándose en las categorías especiales de datos personales establecidas en el artículo".

A este respecto, debe prestarse especial atención a los sistemas que pretenden eludir las prohibiciones o limitaciones de utilizar datos especialmente protegidos en las soluciones IA utilizando *proxies* o datos afines. Así, datos de comportamiento, preferencias, geográficos o similares, pueden fácilmente revelar sexo, raza, religión, orientación sexual, etc. En muchas ocasiones, el tratamiento de datos aparentemente neutros por la solución de IA generará resultados diferenciadores en colectivos especialmente protegidos, o diferencias no razonables ni proporcionales. En este sentido, habrá que analizar posibles "enmascaramientos" y la elección intencional de factores que están cerca de los prohibidos.

No es fácil detectar una discriminación indirecta o encubierta y hay mucha jurisprudencia que varía el porcentaje de personas que deben quedar afectadas negativamente, por lo que aquí interesa, por una solución de IA. No obstante, el TJUE señala que la diferencia de trato perjudicial debe estar sobre entre el 80 y el 90% (2). En todo caso, los datos pueden parecer menos importantes, pero puede ser relevante que se mantengan en el tiempo y como señala Soriano (3), una vez que un algoritmo incorpora un sesgo que perjudica a las personas pertenecientes a un grupo desaventajado y se puede perpetuar, por lo que podrán exigirse diferencias de trato inferiores al referido porcentaje.

(1) Sentencia TJUE de 17 de octubre de 1989, asunto 109/88, Handels- og Kontorfunktionærernes Forbund i Danmark y Dansk Arbejdsgiverforening, en nombre de Danfoss. Sentencia TJUE de 19 de abril de 2012, asunto 415/10, Galina Meister y Speech Design Carrier Systems GmbH.

(2) STJUE 9 de febrero de 1997, C-167/97, Regina contra Secretary of State for Employment, ex parte Nicole Seymour-Smith y Laura Perez, párrafo 61.

(3) Soriano Arnanz, Alba, "Decisiones automatizadas y discriminación: aproximación y propuestas generales", RGDA, nº 56, 2021.

La resolución del Parlamento UE sobre macrodatos “insta” a “minimizar la discriminación y el sesgo algorítmicos” (nº 20) y afirma también la necesaria “mitigación algorítmica” (nº 21, ver también 32). Subraya especialmente que se incluyan mecanismos de transparencia y rendición de cuentas y la posibilidad de corrección de datos y de recurrir las decisiones algorítmicas. La minimización, no obstante, debe garantizar que el sistema inteligente no pierda su eficacia, su propia naturaleza o lleve a generar resultados absurdos. Las medidas de corrección de posibles sesgos han de estar bien justificadas en su necesidad, así como en su razonabilidad y proporcionalidad. De lo contrario pueden a su vez constituir un tratamiento discriminatorio.

Cabe proyectar técnicas de ***discrimination impact assessments*** y de evaluación de impacto de género bien conocidas en la UE. En todo caso, hay que extender el modelo de la responsabilidad proactiva en protección de datos, la no discriminación en el diseño y por defecto, así como medidas concretas en los estudios de impacto.

Igualmente, resulta preciso analizar los resultados del sistema algorítmico, así como las decisiones que se adoptan a partir de los resultados.

Cabe tener también especialmente en cuenta los supuestos de diseño de soluciones de IA que pretenden dar ventaja a los colectivos discriminados tradicionalmente. Se trata de las obligaciones de las acciones positivas en las Directivas (1) o iniciativas voluntarias de compensar algorítmicamente las situaciones discriminatorias. El algoritmo precisamente podrá estar diseñado para “discriminar” positivamente. En estos supuestos habrá que analizar específicamente el diseño de la solución y la adecuación de su funcionamiento, resultados y proporcionalidad.

Por último, en relación con la justicia y equidad, cabe evaluar, por ejemplo, el riesgo de pérdida de puestos de trabajo o de descualificación de la mano de obra y, sobre todo, el impacto social global asociado al uso de las soluciones de IA más allá del usuario final, así como la previsión de mecanismos adecuados de compensación.

(1) Art. 5 de la Directiva de igualdad racial, art. 7 de la Directiva de igualdad en el empleo, art. 3 de la Directiva de igualdad de mujeres y hombres en el empleo y art. 6 de la Directiva de igualdad de mujeres y hombres en el acceso a bienes y servicios.

Conclusiones y perspectivas de futuro

Las soluciones de inteligencia artificial deben evitar ser entrenadas con datos no representativos, defectuosos o sesgados (Principios de Harvard). Si un conjunto de datos no es representativo, el reflejo de la realidad y el mundo no es fiel (*Ethics of IA*). Todo esto debe ser aplicable al ciclo completo de la solución de IA (desde su concepción hasta su retirada del mercado) y para su puesta en marcha se debe tomar en cuenta a todos los agentes involucrados (Ethical *Principles and (Non-) Existing Legal Rules for AI* (1)).

Se debe conseguir que las soluciones de IA, sus algoritmos y redes neuronales puedan sortear los problemas técnicos y así **eviten los sesgos**, pero deben también evitar replicar las prácticas sociales que pueden llevar después a sesgos. En este sentido, es aconsejable que, si una solución de IA prueba su eficiencia en términos de justicia, inclusión, diversidad y no discriminación, las personas hagan caso de sus conclusiones y no apliquen criterios humanos que sí puedan llevar a cometer injusticias y exclusiones.

Igualmente, es imprescindible que la **representatividad** de las muestras se corresponda con el universo a representar, en donde el tamaño y el tipo de muestra deben ser lo más fiables posible para cumplir con el principio de justicia, igualdad, diversidad e inclusión.

Los beneficios de la inteligencia artificial deben **extenderse a todos las personas y regiones** y también deben procurar compartir los bienes y recursos derivados con los países menos adelantados (PMA), los países en desarrollo sin litoral (PDSL) y los pequeños estados insulares en desarrollo (PEID) (Recomendación de la UNESCO) (2).

Las soluciones de inteligencia artificial deben evitar engañar a los usuarios finales o limitar su capacidad de elección. Asimismo, se debe garantizar la capacidad de impugnar y apelar de manera efectiva las decisiones tomadas por las soluciones de IA y por las personas que las operan. Para que esto sea posible, la entidad responsable de la decisión debe ser identificable y el proceso de toma de decisiones debe ser explicable (Principios Éticos del *Knowledge Centre Data & Society*).

Los Estados miembros deben adoptar las medidas adecuadas para garantizar a las personas con discapacidad el acceso, en igualdad de condiciones con las demás, al entorno físico, al transporte, a la información y comunicaciones, incluidos las soluciones de IA, de acuerdo con el artículo 9 de la Convención de las Naciones Unidas sobre los derechos de las personas con discapacidad (Principios Éticos del *Knowledge Centre Data & Society*).

El Derecho europeo antidiscriminatorio y de protección de datos y su interpretación por tribunales y autoridades es muy avanzado. En todo caso, aún no una respuesta completa para lograr el cumplimiento del principio de justicia y no discriminación. Se necesita reforzar algunos elementos para forzar al cumplimiento en el diseño y hacen falta garantías para la mejor detección de las posibles discriminaciones y la forma de actuar frente a las mismas con acceso a los sistemas y mecanismos de prueba.

(1) *Knowledge Centre Data & Society* ((*Kenniscentrum Data & Maatschappij*). *Ethical principles and (non-) existing legal rules for AI*. *October*, 2021.

(2) *United Nations Educational, Scientific and Cultural Organization*. *Recommendation On The Ethics Of Artificial Intelligence*. Noviembre, 2021.

Foco en el ser humano: control y vigilancia humana

Definición descriptiva del principio

El principio de control y vigilancia humana en los sistemas de inteligencia artificial se subdivide en tres vertientes, las cuales, en la medida de lo posible deberían estar presentes en toda solución de IA:

En primer lugar, debe existir presencia humana en algún momento del proceso de toma de decisiones automatizadas de la solución de IA.

Las soluciones han de ser diseñadas para que las personas usuarias puedan supervisarlas, tener el control de estas e interpretar los resultados vertidos por dichas soluciones una vez que comiencen a desplegar sus efectos en un entorno real.

A su vez, debe existir una evaluación humana de los resultados emitidos por la so-lución. Si la revisión se realiza antes de que se adopte la decisión, estaremos ante una evaluación ex ante (decisión parcial automatizada). Si la revisión del resultado se realiza una vez que el resultado se convierte en una decisión que afecta a una persona, estare-mos ante una evaluación *ex post* (decisión totalmente automatizada). Existen ámbitos especialmente sensibles donde se prohíbe la total automatización del proceso decisorio (reserva de humanidad).

Además, el personal encargado de supervisar, controlar y evaluar los resultados de la solución de IA ha de estar lo suficientemente formado, tener la suficiente competencia para adoptar las decisiones y quedar identificado.

En segundo lugar, la persona sobre la que se toman las decisiones ha de ser consciente de que está interactuando con un sistema de inteligencia artificial.

Las soluciones de IA destinadas a interactuar con personas físicas han de estar diseñadas y desarrolladas de forma que dichas personas estén informadas de que están interactuando con una solución de IA, excepto en las situaciones en las que esto resulte evidente debido a las circunstancias y al contexto de utilización. Ejemplos de ello lo encontramos en robots, *chatbots*, sistemas de reconocimiento de emociones o sistemas que generan o manipulan contenido de imagen, sonido, *fake news*, noticias generadas por sistemas inteligentes, etc.

En tercer lugar, las personas sobre las que se toman las decisiones o interactúan con soluciones de IA deben tener el control sobre estas.

Las personas que interactúen con soluciones de IA deben poder mantener una autonomía plena y efectiva sobre sí mismas. Las soluciones de IA no deberían subordinar, coaccionar, engañar, manipular, condicionar o dirigir a los seres humanos de manera in-justificada. En lugar de ello, las soluciones de IA deberían diseñarse de forma que aumenten, complementen y potencien las aptitudes cognitivas, sociales y culturales de las personas. Ello se manifiesta en la posibilidad de que estas personas puedan interactuar con las soluciones o incluso poder evaluarlas o alterarlas. Este principio es una manifestación de la libertad individual en sus diferentes variantes; libertad de conciencia, de pensamiento, política, religiosa, etc.

Por razones de eficacia los seres humanos pueden decidir depender de las soluciones de IA, sin embargo, esta cesión de control ha de ser siempre humana y en cualquier caso no puede implicar una negación de responsabilidad y de rendición de cuentas. Se deben establecer controles para medir de la forma más ajustada posible la contribución de las soluciones IA al bienestar humano, como se declara en *IEEE USE CASE - Criteria for addressing ethical challenges in transparency, accountability and privacy of contact tracing- draft* (1). Según el texto *Asilomar AI Principles*, el desarrollo de la IA debe promover el bien común y el bienestar de la humanidad en todas sus consideraciones sociales y económicas, buscando siempre que sus efectos y contribución pesen más que sus potenciales efectos adversos. Los beneficios deben ser siempre mayores que los potenciales costes.

(1) *Scott L. David, Jean-Claude Goldenstein, Ali G. Hessami, Patricia Shaw, Eleanor (Nell) Watson, Gerlinde Weger. (*14 de octubre 2020*). IEEE USE CASE— CRITERIA FOR ADDRESSING ETHICAL CHALLENGES IN TRANSPARENCY, ACCOUNTABILITY, AND PRIVACY OF CONTACT TRACING— DRAFT. 14* de octubre 2020*, de IEEE SA Standards Association* Sitio web*: https://engagestandards.ieee.org/rs/211-FYL-955/images/ECPAIS_USECASE_10132020_DRAFT.pdf*

Future of Life Institute. (2020). ASILOMAR AI PRINCIPLES. 2020, de Future of Life Institute Sitio web: https://futureoflife.org/2017/08/11/ai-principles/

Desarrollo normativo

En primer lugar, el futuro Reglamento europeo de inteligencia artificial establece en su artículo 14.1 que los sistemas de inteligencia artificial deberán ser diseñados de manera que estos puedan ser supervisados y vigilados una vez que se pongan en funcionamiento. En este sentido, aquellas organizaciones que utilicen sistemas de IA deberán vigilar el funcionamiento de estos basándo e establecer todo tipo de funcionalidades que permitan a las personas que están supervisando los sistemas de IA poder intervenir o interrumpir el funcionamiento de estos (Artículo 14.4.e), y en su caso, invalidar o revertir la información de salida que el sistema genere (Artículo 14.4.d).

Además, esta misma norma obliga a las organizaciones que utilizan los sistemas de inteligencia artificial a identificar al personal que revisa las decisiones en concretos sectores. Así, el Artículo 12.4 establece que, en los supuestos en los que se utilicen sistemas de identificación biométrica, las organizaciones deberán identificar a las personas encargadas de verificar los resultados emitidos por estos sistemas. Estas personas, antes de que se adopte la decisión basada en los resultados emitidos por el sistema, deberán verificar y confirmar los resultados emitidos. Concretamente, la propuesta europea obliga a las organizaciones a que al menos dos personas verifiquen estos resultados (Artículo 14.5).

A su vez, y con el objetivo de que la presencia humana sea real y efectiva, la propuesta normativa mencionada establece la necesidad de diseñar sistemas con un nivel de transparencia suficiente que permita a las personas que analizarán los resultados interpretarlos adecuadamente (Artículo 13.1 y 14.4.c).

Finalmente, cuando las organizaciones diseñen sistemas de IA destinados a interactuar con humanos, estos, se deberán desarrollar de forma que las personas sean conscientes que efectivamente están interactuando con un sistema de IA (Artículo 52.1)

En segundo lugar, la **Recomendación de la UNESCO sobre la ética de la inteligencia artificial** de noviembre de 2021, al analizar el principio de vigilancia y control humano, señala la necesidad de que durante todo el ciclo de vida de los sistemas de IA sea posible atribuir responsabilidad ética y jurídica a personas físicas o a entidades jurídicas existentes (Apartado 35). Por otro lado, aunque la Unesco es consciente que por razones de eficacia en muchos ámbitos el proceso decisorio puede llegar a automatizarse por completo, también considera que determinados sectores, como pueden ser las decisiones de vida o muerte, no deberían cederse por completo a los sistemas de IA (Apartado 36).

En tercer lugar, algunas normas están conectadas con la vigilancia y control humano. En este sentido, tanto el artículo 3.1.g) del **Reglamento (UE) 2021/1232 de 14 de julio de 2021** sobre lucha contra los abusos sexuales de menores en línea (1), como el artículo 5.3 del **Reglamento (UE) 2021/784 de 29 de abril de 2021** sobre difusión de contenido terrorista en línea (2), obligan a las plataformas en línea a que, cuando utilicen medios automatizados para controlar el contenido potencialmente ilícito, establezcan los cauces necesarios para que los resultados emitido por los sistemas de IA se analicen por al menos una persona física antes de que se adopte una decisión que afecta a un particular. Es decir, se obliga a las organizaciones a realizar una evaluación ex ante de los resultados antes de que finalmente se adopte una decisión que afecta a un particular.

Por otro lado, el **artículo 22.3 del RGPD** reconoce a los particulares sometidos a decisiones totalmente automatizadas relevantes adoptadas por sistemas de inteligencia artificial el derecho a solicitar que la decisión que ha adoptado el sistema pueda ser revisada por una persona. Es decir, el RGPD permite la toma de decisiones plenamente automatizadas, sin embargo, si estas se adoptan, los responsables del tratamiento deben reconocer a los particulares el derecho a solicitar la revisión de los resultados vertidos por los sistemas de IA.

(1) Reglamento (UE) 2021/1232 del Parlamento Europeo y del Consejo de 14 de julio de 2021 por el que se establece una excepción temporal a determinadas disposiciones de la Directiva 2002/58/CE en lo que respecta al uso de tecnologías por proveedores de servicios de comunicaciones interpersonales independientes de la numeración para el tratamiento de datos personales y de otro tipo con fines de lucha contra los abusos sexuales de menores en línea.

(2) Reglamento (UE) 2021/784 del Parlamento Europeo y del Consejo, de 29 de abril de 2021, sobre la lucha contra la difusión de contenidos terroristas en línea.

Finalmente, y con el objetivo de otorgar cierto control a los particulares que interactúan con sistemas de IA o sobre los que se adoptan decisiones, el artículo 29.2 de la Propuesta de Reglamento europeo de servicios digitales proyecta permitir a los usuarios de las grandes plataformas acceder y rectificar los perfiles sobre recomendación de contenido automatizado que estén generando los algoritmos que utilizan estas organizaciones (1). En este mismo se sentido se ha pronunciado el G29-UE a la hora de interpretar el RGPD. Así, esta autoridad de control europea considera que los responsables del tratamiento deberían poder permitir a los titulares de los datos el acceso y rectificación a los perfiles que elaboran los sistemas de IA a través de una interfaz o herramienta de gestión de preferencias (2).

En cuarto lugar, a nivel nacional cabe destacar la Carta de Derecho Digitales española. (texto no vinculante). Su artículo 18.4 reconoce a los ciudadanos el derecho a relacionarse con los poderes públicos de forma no telemática. El artículo 23.1 reconoce a los particulares el derecho a solicitar que la asistencia sanitaria sea presencial. Además, el artículo 18.6.d) prohíbe la toma de decisiones discrecionales administrativas plenamente automatizadas salvo que normativamente se prevea tal automatización. Además, i) El derecho de los particulares a solicitar la supervisión e intervención humana, así como a impugnar las decisiones automatizadas tomadas por sistemas de inteligencia artificial (Artículo 25.3). ii) El derecho a que las personas puedan saber cuándo una noticia informativa u otro contenido basada en el ejercicio de la libertad de información o expresión ha sido creado por sistemas automatizados (Artículo 15.2.a).

Otras normas españolas reconocen la revisión y supervisión humana, como: el artículo 73.11 del **Real Decreto-ley 24/2021 sobre sobre derechos de autor** (3) o el artículo 14.1 de la Ley Orgánica 7/2021 sobre tratamientos de datos con fines de investigación penal.

(1) Propuesta de REGLAMENTO DEL PARLAMENTO EUROPEO Y DEL CONSEJO relativo a un mercado único de servicios digitales (Ley de servicios digitales) y por el que se modifica la Directiva 2000/31/CE. Resolución de 15 de diciembre de 2020.

(2) G29-UE. (2018). Directrices sobre decisiones individuales automatizadas y elaboración de perfiles a los efectos del Reglamento 2016/679, 3 de octubre de 2017, versión final 6 de febrero de 2018, Doc WP251rev.01.

(3) Real Decreto-ley 24/2021, de 2 de noviembre, de transposición de directivas de la Unión Europea en las materias de bonos garantizados, distribución transfronteriza de organismos de inversión colectiva, datos abiertos y reutilización de la información del sector público, ejercicio de derechos de autor y derechos afines aplicables a determinadas transmisiones en línea y a las retransmisiones de programas de radio y televisión, exenciones temporales a determinadas importaciones y suministros, de personas consumidoras y para la promoción de vehículos de transporte por carretera limpios y energéticamente eficientes.

Conclusiones y perspectivas de futuro

El principio de control y vigilancia humana en los sistemas de inteligencia artificial no deja de ser un reflejo de la desconfianza interior que tenemos las personas hacia los sistemas de IA. En este sentido, dejar que un sistema de IA adopte decisiones que anteriormente venía realizando una persona es un proceso al que los humanos nos costará adaptarnos, sin embargo, este proceso es imparable.

Se hace necesario por tanto establecer toda una serie de garantías mínimas que han de estar presentes en aquellos sistemas de inteligencia artificial cuando los mismos generen riesgos relevantes. Hasta la fecha, las normas han focalizado su atención en tres vertientes: i) Se debe garantizar la presencia humana en algún momento del proceso de toma de decisiones automatizadas, ya sea a través del control, supervisión o la revisión de los resultados emitidos por el sistema de IA. ii) La persona sobre la que se toman las decisiones ha de ser consciente de que está interactuando con un sistema de inteligencia artificial. iii) Las personas sobre las que se toman las decisiones o interactúan con soluciones de IA deben tener el control sobre estas.

El despliegue de las garantías mencionadas en algunos sectores resultará obligatorio mientras que en otros variará en función de diversas razones como pueden ser el ámbito donde se implemente el sistema, la funcionalidad o las características técnicas del mismo.

En el futuro, y fruto de la cada vez mayor presencia de las tecnologías en los procesos decisorios, será necesario establecer ámbitos o sectores específicos vetados a la plena automatización. Por otro lado, la supervisión y control por parte de las personas que analizan el funcionamiento y los resultados emitidos no podrá considerarse una garantía efectiva y plena si esas personas no llegan a comprender realmente los sistemas de IA y los efectos que los mismos pueden generar en el entorno donde se desplegarán.

Promoción de los valores y de los derechos humanos

Definición descriptiva del principio.

Como afirmó el Parlamento de la Unión Europea en 2017, la UE no puede renunciar a sus "valores humanistas intrínsecamente europeos y universales que caracterizan la contribución de Europa a la sociedad propios valores humanistas y principios basados en la dignidad y los derechos fundamentales" (1).

El principio de promoción de los valores y derechos humanos integra, de un lado, **la promoción de los valores humanos y la mejora y progreso de la humanidad**. De este modo, el desarrollo y uso de la IA ha de estar dirigido al beneficio y bienestar de la sociedad y de la civilización humana para la mejora de condiciones de vida, salud, trabajo, desarrollo de capacidades físicas y psíquicas. Los efectos y contribuciones positivas de las soluciones de IA deben pesar más que sus potenciales efectos adversos. Las soluciones de IA tienen un gran potencial de generar una prosperidad económica que debe ser distribuida equitativamente, beneficiar y empoderar a toda la humanidad (2).

Los objetivos de la IA implican la vinculación de sus usos con intereses públicos, como en general, la mayor eficacia de los derechos humanos, así como los principios de democracia, justicia y el Estado social y de Derecho, además de la sostenibilidad. Así, frente a visiones que únicamente primen la rentabilidad, la seguridad y un mero uso legal de la IA. El desarrollo y uso de la IA obliga a valorar más allá de los beneficios directos o inmediatos del desarrollador, sino teniendo en cuenta los beneficios sociales y globales sobre los riesgos y ventajas. Debe señalarse en todo caso, que algunos de estos elementos están claramente integrados en el principio de justicia y no discriminación de la IA o sostenibilidad. Como luego se indica, resulta de interés analizar el nexo entre la IA y los Objetivos de Desarrollo Sostenible ("ODS").

Por otro lado, cabe centrarse en **la promoción de los derechos humanos**, elemento básico inspirador del tratamiento ético y jurídico de la IA: inspira los objetivos a perseguir, esencialmente, hacer efectivos estos derechos; limita los instrumentos y herramientas a utilizar y define las garantías de la sociedad civil y los individuos frente al desarrollo y uso de la IA.

El punto de partida y premisa ética de la IA en las numerosas declaraciones y los diferentes documentos no es otro que la dignidad y los derechos fundamentales. Por ejemplo, desde 2018 la marca de la IA confiable y en el diseño *AI Made in Europe*, parten de la dignidad humana y los derechos fundamentales. Para el Grupo de Altos Expertos de la UE, los derechos fundamentales son la base de concepto de "IA confiable" y del propio "fin ético" de la IA, que es precisamente garantizar su cumplimiento. Los derechos son "la base para la formulación", "el trampolín *stepping stone* para identificar principios y valores éticos abstractos y concretos y operativos" (3). Especial acierto tuvo el Supervisor Europeo de Protección de datos cuando afirmó que hay que situar a "La dignidad en el centro de una nueva ética digital", "un mayor respeto de la dignidad humana y una mayor salvaguardia de la misma podrían servir de contrapeso a la vigilancia generalizada y la asimetría de poder a la que se enfrentan las personas" (4). Además, señalar que la UNESCO (5) afirma recientemente que "las personas nunca deberían ser cosificadas, su dignidad no debería ser menoscabada de ninguna otra manera, y sus derechos humanos y libertades fundamentales nunca deberían ser objeto de violación o abusos" (Recomendación nº 15).

(1) Parlamento Europeo Normas de Derecho civil sobre robótica. Resolución del Parlamento Europeo, de 16 de febrero de 2017, con recomendaciones destinadas a la Comisión sobre normas de Derecho civil sobre robótica (2015/2103(INL)) letra U. Acceso en http://www.europarl.europa.eu/sides/getDoc.do?pubRef=-//EP//TEXT+TA+P8-TA-2017-0051+0+DOC+XML+V0//ES

(2) *Future of Life Institute*. (2020). *ASILOMAR AI PRINCIPLES*. 2020, de *Future of Life Institute* Sitio web: https://futureoflife.org/2017/08/11/ai-principles/

(3) Grupo de expertos de alto nivel sobre inteligencia artificial, AI-HLEGHLEG, Directrices éticas para una IA fiable, 20182018, IV, 5-6).

(4) Supervisor Europeo de Protección de Datos (2015 a). Dictamen 4/2015. Hacia una nueva ética digital. Datos, dignidad y tecnología, 11 septiembre, pp. 14 y ss.) acceso en https://edps.europa.eu/data-protection/our-work/publications/opinions/towards-new-digital-ethics-data-dignity-and_en

(5) Conferencia General 41ª reunión - París, 41 C/73, 22 de noviembre de 2021. Anexo. https://unesdoc.unesco.org/ark:/48223/pf0000379920_spa

Una advertencia necesaria es que hay un riesgo general con la referencia general al cumplimiento ético de la IA, que en ocasiones sirve para un blanqueamiento de las soluciones de IA por referencia a genéricas e inconcretas obligaciones. Este riesgo puede darse especialmente respecto de principios más genéricos como este, relativos al desarrollo de valores y derechos humanos. Pues bien, de los derechos humanos derivan tanto en el plano nacional como internacional determinadas y concretas obligaciones que en modo alguno pueden ser eludidas ni por poderes públicos ni por entidades privadas.

Cabe destacar en esta dirección la mencionada Recomendación UNESCO (1) cuando afirma lo siguiente:

> Los valores y principios […] deberían ser respetados por todos los actores durante el ciclo de vida de los sistemas de IA, en primer lugar, y, cuando resulte necesario y conveniente, ser promovidos mediante modificaciones de las leyes, los reglamentos y las directrices empresariales existentes y la elaboración de otros nuevos. Todo ello debe ajustarse al derecho internacional, en particular la Carta de las Naciones Unidas y las obligaciones de los Estados Miembros en materia de derechos humanos, y estar en consonancia con los […] Objetivos de Desarrollo Sostenible (ODS) de las Naciones Unidas. (Recomendación nº 61)

Por lo tanto, la legislación debe actualizarse en esta dirección (Unesco nº 61).

Del principio de derechos humanos cabe derivar un seguimiento estricto de las obligaciones internacionales que los derechos imponen, incluyendo articular sistemas de gobernanza internacional y regional para su seguimiento. En esta dirección, la Declaración UNESCO recuerda que “derecho internacional de los derechos humanos, son esenciales a lo largo del ciclo de vida de los sistemas de IA” (Recomendación nº 13).

Así las cosas, hay que hacer un particular esfuerzo para derivar recomendaciones y obligaciones concretas a partir de este principio y la posibilidad de establecer indicadores y estándares básicos para su cumplimiento. Esta es la clara dirección de los instrumentos de la UE e internacionales y ya pueden derivarse diversas obligaciones a partir de la normativa de protección de datos. El principio de derechos humanos respecto de la IA implica el sometimiento a todas las obligaciones del Derecho supranacional de los derechos humanos, así como impulso y seguimiento de la actualización de las garantías y contenidos de los derechos al ámbito de la IA, el establecimiento de nuevas técnicas de garantías, especialmente los análisis de riesgos, evaluaciones de impacto, auditorías, integración de los derechos y los intereses colectivos en estos instrumentos, la fijación de indicadores de medición y evaluación. Han de darse especiales esfuerzos para la proyección de los derechos en el ámbito privado, especialmente en las grandes compañías mundiales, así como en sectores y profesiones específicas y vinculadas a la IA.

(1) *UNESCO. (22 de noviembre de 2021). INFORME DE LA COMISIÓN DE CIENCIAS SOCIALES Y HUMANAS (SHS). 2021, de UNESCO Sitio web: https://unesdoc.unesco.org/ark:/48223/pf0000379920_spa*

Desarrollo normativo

Desde mitad del siglo XX se viene desarrollando un Derecho internacional y supranacional de los derechos humanos. Así, a nivel mundial (1) y, especialmente, en el ámbito europeo contamos con más de cien instrumentos en el marco del Consejo de Europa con sus diferentes sistemas de garantía. Destaca en este sentido el Convenio Europeo de Derechos Humanos y el Tribunal Europeo de Derechos Humanos.

En el concreto marco de la UE cabe recordar que el artículo 2 del Tratado de la Unión Europea dispone que "La Unión se fundamenta en los valores de respeto de la dignidad humana, libertad, democracia, igualdad, Estado de Derecho y respeto de los derechos humanos, incluidos los derechos de las personas pertenecientes a minorías. Estos valores son comunes a los Estados miembros en una sociedad caracterizada por el pluralismo, la no discriminación, la tolerancia, la justicia, la solidaridad y la igualdad entre mujeres y hombres." Además, su artículo 3.5 se refiere a la promoción de estos valores en relación con el resto del mundo, promoviendo la contribución a "la paz, la seguridad, el desarrollo sostenible, la solidaridad y el respeto mutuo entre los pueblos". La Carta de derechos fundamentales de la UE constituye asimismo un referente ineludible para la UE y sus Estados miembros. Desde hace décadas, pero especialmente en los últimos años, los derechos y libertades fundamentales están incorporando contenidos y garantías específicas para el ámbito digital y en particular ante el fenómeno de la IA. En este sentido destaca como primer paso a tener en cuenta la Carta de Derechos digitales adoptada en julio de 2021 (2) en España.

Como se ha adelantado, es muy importante la introducción de los derechos y libertades en las **técnicas de evaluación de riesgos del desarrollo y uso de la IA**. La evaluación de riesgos para los derechos humanos se afirma en los documentos internacionales, recientemente lo recuerda la UNESCO (3): "debería garantizarse la aplicación de procedimientos de evaluación de riesgos y la adopción de medidas para impedir que ese daño se produzca", (Recomendación nº 25). Y no se trata de una mera afirmación, sino que deben aplicarse los procedimientos de evaluación de riesgos (Recomendación, nº 25) y al decidir el uso de un sistema IA debe justificarse que no supone "una violación o un abuso de los derechos humanos" (Recomendación nº 26 b). Se insiste en que el "aprendizaje sobre el impacto de los sistemas de IA" y "el enfoque y la comprensión de los sistemas de IA deberían basarse en el impacto de estos sistemas en los derechos humanos" (Recomendación nº 45). En estas evaluaciones de impacto hay que tener particularmente en cuenta a personas "en situación de vulnerabilidad, los derechos laborales, el medio ambiente y los ecosistemas, así como las consecuencias éticas y sociales, y facilitar la participación ciudadana" (Recomendación nº 50) y ello, en todas las etapas del ciclo IA (Recomendación nº 51). Se afirma el compromiso de "la elaboración de una metodología de la UNESCO de evaluación del impacto ético de las tecnologías de la IA basada en una investigación científica rigurosa y fundamentada en el derecho internacional de los derechos humanos" (Recomendación nº 131).

De las Directrices éticas para una IA fiable de la UE cabe tener en cuenta apartados relativos a la acción y supervisión humana, exigiendo si "¿ha llevado usted a cabo una evaluación del impacto sobre los derechos fundamentales? ¿Ha identificado y documentado los posibles equilibrios entre los diferentes principios y derechos?".

Como se ha hecho referencia respecto de la sostenibilidad, hay que tener en cuenta las cuestiones sobre el "Bienestar social y ambiental". De igual modo, el análisis de "impacto social" general "más allá del que tenga sobre el usuario (final), como, por ejemplo, las partes interesadas que pueden verse indirectamente afectadas por dicho sistema". Se hace referencia asimismo a la necesidad de "algún mecanismo para identificar los intereses y valores que implica el sistema de IA y los posibles equilibrios entre ellos" y que existan procesos para decidir sobre los equilibrios necesarios y que queden documentados. En particular, ¿se afirma la conveniencia de evaluar el riesgo de pérdida de puestos de trabajo o de descualificación de la mano de obra? ¿Qué pasos se han dado para contrarrestar esos riesgos?

(1) Declaración Universal de Derechos Humanos (1948), el Pacto Internacional de Derechos Civiles y Políticos (1966), el Pacto Internacional de Derechos Económicos, Sociales y Culturales (1966), Estatuto de los Refugiados (1951), el Convenio sobre la Discriminación (Empleo y Ocupación) (1958), la Convención Internacional sobre la Eliminación de Todas las Formas de Discriminación Racial (1965), la Convención sobre la Eliminación de Todas las Formas de Discriminación contra la Mujer (1979), la Convención sobre los Derechos del Niño (1989), la Convención sobre los Derechos de las Personas con Discapacidad (2006), la Convención relativa a la Lucha contra las Discriminaciones en la Esfera de la Enseñanza (1960) y la Convención sobre la Protección y Promoción de la Diversidad de las Expresiones Culturales (2005) y un largo etcétera. Puede seguirse, https://derechoshumanos.net/

(2) Gobierno de España-De la Quadra, Tomás (coord.) (2021). Carta de Derechos digitales, julio de 2021. https://www.lamoncloa.gob.es/presidente/actividades/Documents/2021/140721-Carta_Derechos_Digitales_RedEs.pdf

(39 UNESCO. (22 de noviembre de 2021). INFORME DE LA COMISIÓN DE CIENCIAS SOCIALES Y HUMANAS (SHS). 2021, de UNESCO Sitio web: https://unesdoc.unesco.org/ark:/48223/pf0000379920_spa

Debe recordarse que en la actual **normativa de protección de datos vigente** se impone en general el análisis de riesgos y en particular el estudio de impacto, que aplica en general a las soluciones de IA. Y estos análisis no se limitan solo al impacto que puede tener la solución de IA en el concreto derecho de protección de datos, sino en todos los derechos y libertades. A partir de esta obligación actual y existente, la **AEPD** insiste en el análisis de proporcionalidad y necesidad: "un análisis y gestión del riesgo para los derechos y libertades de los interesados que introduce en el tratamiento el procesamiento de los datos mediante el componente IA" (1). Ello conlleva la obligación en evaluar y documentar que se han tenido en cuenta todas las opciones posibles para el menor impacto en todos los derechos y libertades afectados. Antes de optar por nuevas soluciones de IA, valorar soluciones más probadas que minimizan datos o hacen una solución menos intensivo de explotación de datos. Si hay que abordar nuevos problemas, justificar el uso mismo de IA que pueda afectar derechos. Estas evaluaciones de impacto en derechos deben darse también en la fase de pruebas (y se insiste en la protección de los derechos en una visión amplia), como desde el punto de vista de seguridad (2).

(1) *AEPD,2021. Requisitos para Auditorías de Tratamientos que incluyan IA, 2021, p. 17.*

(2) Ibídem. ps. 28 y 31.

Conclusiones y perspectivas de futuro

El **futuro reglamento de IA** lleva implícito todo el sistema de análisis de riesgos que gira en buena medida sobre los derechos y libertades. Se prohíben directamente algunas soluciones que implican una lesión o un riesgo inaceptable para los derechos. En todo caso, el núcleo de la regulación son las soluciones de alto riesgo de IA que plantean riesgos significativos para la salud y la seguridad o los derechos fundamentales de las personas. Respecto de estos se determinan "los requisitos de datos de alta calidad, documentación y trazabilidad, transparencia, supervisión humana, exactitud y solidez, son estrictamente necesarios para mitigar los riesgos para los derechos fundamentales y la seguridad que plantea la IA y que no están cubiertos por otros marcos jurídicos existentes". No obstante, pese a la importancia de los derechos en este enfoque de riesgos, es especialmente deseable que el futuro reglamento incorpore garantías específicas para los derechos de las personas afectadas y para los colectivos que las representen, algo inexistente en la propuesta actual. El RGPD no es suficiente a este respecto.

Sostenibilidad ambiental

Definición descriptiva del principio

La sostenibilidad medioambiental, además de un principio, debe ser una condición indispensable para las soluciones de IA. Las condiciones de vida de las futuras generaciones deben ser preservadas y, para ello, es necesario que las infraestructuras creadas por y para este tipo de tecnologías economicen los recursos energéticos utilizados y eliminen y reparen el *hardware* y demás recursos involucrados. No solo se debe prevenir el daño, también hacer énfasis en generar efectos positivos para el medioambiente.

En este sentido, desarrollo sostenible podría definirse como el acceso a la cultura, impulsar la agricultura sostenible, implementar en las empresas políticas de cuidado a los mayores y de atención a las personas con capacidades especiales, políticas de igualdad y diversidad, políticas respetuosas con el medio ambiente…

Es de particular importancia, en materia de sostenibilidad la cuestión del medioambiente y las normas locales e internacionales que regulan este aspecto, aportando instrucciones de ética-normativa y guías y directrices éticas de comportamiento y acción con impacto directo en todas las fases del proceso creativo de las soluciones de IA.

Desarrollo normativo

Especialmente vinculado con la sostenibilidad medioambiental, cabe destacar el **Pacto Verde Europeo** (1). Este estudio evalúa el papel potencial de soluciones IA con respecto al Pacto Verde Europeo (en adelante **PVE**). El PVE tiene la intención de promover la equidad y la prosperidad y una economía eficiente, competitiva y más sostenible para Europa. Las soluciones IA tienen un amplio rango de potencial para apoyar una transformación socioecológica, su implementación también puede dar lugar a riesgos ambientales graves y de gran alcance. La política y la regulación ambiental tienen un amplio espectro de instrumentos a su disposición para permitir el desarrollo de tecnologías e innovaciones de manera orientada a objetivos, así como para proteger y tomar precauciones contra riesgos específicos de sostenibilidad.

Para ello se propone reforzar la investigación sobre las soluciones IA y sus aplicaciones para los objetivos del PVE y así desarrollar, promover e implementar metodologías para evaluaciones de impacto ambiental de tecnologías de IA. También se propone establecer un mayor seguimiento y evaluación de las tendencias, potenciar los sistemas y aplicaciones transformadores en los mercados digitales, explorar, desarrollar y promover un modelo europeo de economía de datos que aproveche los datos como recurso para la IA sostenible.

Es preciso que los análisis y evaluaciones de impacto y de riesgos trasciendan de los intereses y derechos solo individuales y que se realicen en perspectiva social y colectiva. Así, resulta necesaria la entrada de agentes sociales y de la sociedad civil en el diseño de políticas y medidas de garantía respecto del uso de la IA. Ello es ya habitual en materia de medio ambiente, en concreto: el Convenio de Aarhus de 25 de junio de 1998 del que forman parte 46 estados y la amplia normativa medioambiental de la UE que facilita la transparencia y la participación. También cabe tener en cuenta el futuro desarrollo del artículo 80 RGPD para que una entidad, organización o asociación sin ánimo de lucro pueda representar a los interesados.

Por otro lado, la **Recomendación sobre la Ética de la inteligencia artificial** (2), de la UNESCO, fija como objetivo "poner los sistemas de IA al servicio de la humanidad, las personas, las sociedades y el medio ambiente y los ecosistemas, así como para prevenir daños." (nº 5). Al respecto del medio ambiente y los ecosistemas se insiste en la obligación de respetar todas las normas aplicables, así como "reducir el impacto ambiental de los sistemas de IA, en particular, aunque no exclusivamente, su huella de carbono" (Recomendación nº 18). Asimismo, la evaluación de riesgos y las medidas a adoptar han de incluir la protección del medio ambiente y los ecosistemas (Recomendaciones nº 25 y en especial 84 a 86) y eliminarse a lo largo de todo el ciclo de vida (Recomendación nº 27).

(1) *Peter Gailhofer, Anke Herold, Jan Peter Schemmel, Cara-Sophie Scherf, Cristina Urrutia, Andreas R. Köhler and Sibylle Braungardt.* (Mayo 2021). *The role of Artificial Intelligence in the European Green Deal. 2021,* de *European Parliament* Sitio web*: https://www.europarl.europa.eu/RegData/etudes/STUD/2021/662906/IPOL_STU(2021)662906_EN.pdf*

(2) Conferencia General 41ª reunión - París*, 41 C/73,* 22 de noviembre de 2021. Anexo*. https://unesdoc.unesco.org/ark:/48223/pf0000379920_spa* y texto definitivo de la Recomendación sobre la Ética de la Inteligencia Artificial *https://unesdoc.unesco.org/ark:/48223/pf0000380455_spa*

Asimismo, uno de los grandes proyectos de ***The Global Partnership on AI*** (1) es crear una estrategia de IA responsable para el medio ambiente. Declaran que la lucha combinada contra el cambio climático y la preservación de la biodiversidad representa uno de los desafíos más urgentes que enfrenta la humanidad. Como respuesta, este proyecto tiene como objetivo desarrollar una estrategia global de adopción de IA responsable para la acción climática y la preservación de la biodiversidad.

(1) GPAI. (Junio de 2020). *The Global Partnership on Artificial Intelligence. 2020, de GPAI.* Sitio web: https://gpai.ai/

Conclusiones y perspectivas de futuro

La sociedad en su conjunto, prestando especial atención a los grupos minoritarios y al medio ambiente, debe ser partes interesada en todo el ciclo de vida de la solución IA. La sostenibilidad y la responsabilidad ecológica debe fomentarse y se debe impulsar la investigación para promover el desarrollo sostenible como base necesaria para la inteligencia artificial ética. Solo así se garantizará el futuro de las próximas generaciones.

Resumen

En conclusión, este es el **marco ético-jurídico del *Framework* GuIA** con las normativas españolas y europeas a tener en cuenta por cada principio ético. Diferenciando entre las que, a día de hoy, son obligatorias y las que son recomendaciones.

1. Privacidad y gobierno de datos

Normativa obligatoria:

- Convenio n. 108 del Consejo de Europa, de 28 de enero de 1981, para la protección de las personas con respecto al tratamiento automatizado de datos de carácter personal.
  https://www.boe.es/buscar/doc.php?id=BOE-A-1985-23447
- Artículos 7 y 8 de la Carta de los Derechos Fundamentales de la Unión Europea. -
  https://www.boe.es/doue/2010/083/Z00389-00403.pdf
- Reglamento General de Protección de Datos.
  https://www.boe.es/doue/2016/119/L00001-00088.pdf
- Directiva (UE) 2016/680, de 27 de abril, relativa a la protección de las personas físicas en lo que respecta al tratamiento de datos personales por parte de las autoridades competentes para fines de prevención, investigación, detección o enjuiciamiento de infracciones penales o de ejecución de sanciones penales, y a la libre circulación de dichos datos.
  https://www.boe.es/doue/2016/119/L00089-00131.pdf
- Directiva (UE) 2016/681 del Parlamento Europeo y del Consejo, de 27 de abril de 2016, relativa a la utilización de datos del registro de nombres de los pasajeros (PNR) para la prevención, detección, investigación y enjuiciamiento de los delitos de terrorismo y de la delincuencia grave.
  https://www.boe.es/doue/2016/119/L00132-00149.pdf
- Directiva 2002/58/CE del Parlamento Europeo y del Consejo, de 12 de julio de 2002, relativa al tratamiento de los datos personales y a la protección de la intimidad en el sector de las comunicaciones electrónicas.
  https://www.boe.es/doue/2002/201/L00037-00047.pdf
- Reglamento (UE) 2018/1807 del Parlamento Europeo y del Consejo, de 14 de noviembre de 2018, relativo a un marco para la libre circulación de datos no personales en la Unión Europea.
  https://www.boe.es/doue/2018/303/L00059-00068.pdf

2. Seguridad y protección. Fiabilidad, robustez y precisión

Normativa obligatoria:

- Directiva (UE) 2016/1148, de 6 de julio de 2016, relativa a las medidas destinadas a garantizar un elevado nivel común de seguridad de las redes y sistemas de información.
  https://www.boe.es/doue/2016/194/L00001-00030.pdf
- Reglamento (UE) 2019/881 del Parlamento Europeo y del Consejo, de 17 de abril de 2019, relativo a ENISA (Agencia de la Unión Europea para la Ciberseguridad) y a la certificación de la ciberseguridad de las tecnologías de la información y la comunicación y por el que se deroga el Reglamento (UE) nº 526/2013 («Reglamento sobre la Ciberseguridad»).
  https://www.boe.es/doue/2019/151/L00015-00069.pdf
- Reglamento (UE) 2021/887 del Parlamento Europeo y del Consejo, de 20 de mayo de 2021, por el que se establecen el Centro Europeo de Competencia Industrial, Tecnológica y de Investigación en Ciberseguridad y la Red de Centros Nacionales de Coordinación.
  https://www.boe.es/doue/2021/202/L00001-00031.pdf
- Directiva 2013/40/UE del Parlamento Europeo y del Consejo, de 12 de agosto de 2013, relativa a los ataques contra los sistemas de información y por la que se sustituye la Decisión marco 2005/222/JAI del Consejo.
  https://www.boe.es/doue/2013/218/L00008-00014.pdf
- Reglamento (UE) 910/2014 del Parlamento Europeo y del Consejo, de 23 de julio de 2014, relativo a la identificación electrónica y los servicios de confianza para las transacciones electrónicas en el mercado interior y por la que se deroga la Directiva 1999/93/CE.
  https://www.boe.es/doue/2014/257/L00073-00114.pdf
- Directiva 2002/58/CE del Parlamento Europeo y del Consejo, de 12 de julio de 2002, relativa al tratamiento de los datos personales y a la protección de la intimidad en el sector de las comunicaciones electrónicas.
  https://www.boe.es/doue/2002/201/L00037-00047.pdf
- Reglamento General de Protección de Datos.
  https://www.boe.es/doue/2016/119/L00001-00088.pdf

Recomendaciones:

Propuesta de Reglamento del Parlamento Europeo y del Consejo por el que se establecen normas armonizadas en materia de inteligencia artificial (Ley de inteligencia artificial) y se modifican determinados actos legislativos de la Unión.

https://eur-lex.europa.eu/resource.html?uri=cellar:e0649735-a372-11eb-9585-01aa75ed71a1.0008.02/DOC_1&format=PDF

3. Responsabilidad y rendición de cuentas

Normativa obligatoria:

- Artículos 4.2.f), 12, 114 y 169 del Tratado de Funcionamiento de la Unión Europea.
  https://www.boe.es/doue/2010/083/Z00047-00199.pdf
- Artículo 38 de la Carta de los Derechos Fundamentales de la Unión Europea.
  https://www.boe.es/doue/2010/083/Z00389-00403.pdf
- Directiva (UE) 85/374/CEE del Consejo de 25 de julio de 1985, relativa a la aproximación de las disposiciones legales, reglamentarias y administrativas de los Estados Miembros en materia de responsabilidad por los daños causados por productos defectuosos.
  https://www.boe.es/doue/1985/210/L00029-00033.pdf
- Artículos 5 y 24 del Reglamento General de Protección de Datos.
  https://www.boe.es/doue/2016/119/L00001-00088.pdf
- Directiva (UE) 2016/680, de 27 de abril, relativa a la protección de las personas físicas en lo que respecta al tratamiento de datos personales por parte de las autoridades competentes para fines de prevención, investigación, detección o enjuiciamiento de infracciones penales o de ejecución de sanciones penales, y a la libre circulación de dichos datos.
  https://www.boe.es/doue/2016/119/L00089-00131.pdf

Recomendaciones:

Resolución, de fecha 16 de febrero de 2017, que contiene recomendaciones destinadas a la Comisión sobre normas de Derecho Civil sobre robótica 2015/2103 (INL).

https://www.europarl.europa.eu/doceo/document/TA-8-2017-0051_ES.html#title1

Resolución en la que se contienen recomendaciones a la Comisión sobre el régimen de responsabilidad civil en materia de IA (2020/2014/INL).

https://www.europarl.europa.eu/doceo/document/TA-9-2020-0276_ES.html#title1

4. Transparencia y explicabilidad

Normativa obligatoria:

- Reglamento General de Protección de Datos.
  https://www.boe.es/doue/2016/119/L00001-00088.pdf
- Directiva (UE) 2016/680, de 27 de abril, relativa a la protección de las personas físicas en lo que respecta al tratamiento de datos personales por parte de las autoridades competentes para fines de prevención, investigación, detección o enjuiciamiento de infracciones penales o de ejecución de sanciones penales, y a la libre circulación de dichos datos.
  https://www.boe.es/doue/2016/119/L00089-00131.pdf

Recomendaciones:

Propuesta de Reglamento del Parlamento Europeo y del Consejo por el que se establecen normas armonizadas en materia de inteligencia artificial (Ley de inteligencia artificial) y se modifican determinados actos legislativos de la Unión.

https://eur-lex.europa.eu/resource.html?uri=cellar:e0649735-a372-11eb-9585-01aa75ed71a1.0008.02/DOC_1&format=PDF

5. El principio de justicia

Normativa obligatoria:

- Artículo 14 y el protocolo nº 12 del Convenio Europeo de Derechos Humanos (CEDH).
  https://www.echr.coe.int/documents/convention_spa.pdf
- Artículos 2, 3, 9 y 10 del Tratado de Funcionamiento de la Unión Europea.
  https://www.boe.es/doue/2010/083/Z00047-00199.pdf
- Artículos 20 a 26 de la Carta de los Derechos Fundamentales de la Unión Europea.
  https://www.boe.es/doue/2010/083/Z00389-00403.pdf
- Directiva 2000/78/CE del Consejo, de 27 de noviembre de 2000, relativa al establecimiento de un marco general para la igualdad de trato en el empleo y la ocupación.
  https://www.boe.es/buscar/doc.php?id=DOUE-L-2000-82357
- Directiva 2000/43/CE del Consejo, de 29 de junio de 2000, relativa a la aplicación del principio de igualdad de trato de las personas independientemente de su origen racial o étnico.
  https://eur-lex.europa.eu/legal-content/ES/TXT/PDF/?uri=CELEX:32000L0043&from=ES
- Directiva del Consejo 2004/113/CE, de 13 de diciembre de 2004, por la que se aplica el principio de igualdad de trato entre hombres y mujeres al acceso a bienes y servicios y su suministro.
  https://www.boe.es/buscar/doc.php?id=DOUE-L-2004-82937
- Directiva del Consejo, de 19 de diciembre de 1978, relativa a la aplicación progresiva del principio de igualdad de trato entre hombres y mujeres en materia de seguridad social.
  https://www.boe.es/doue/1979/006/L00024-00025.pdf
- Directiva 2010/18/UE del Consejo, de 8 de marzo de 2010, por la que se aplica el Acuerdo marco revisado sobre el permiso parental, celebrado por BUSINESSEUROPE, la UEAPME, el CEEP y la CES, y se deroga la Directiva 96/34/CE.
  https://boe.es/doue/2010/068/L00013-00020.pdf
- Reglamento General de Protección de Datos.
  https://www.boe.es/doue/2016/119/L00001-00088.pdf
- Directiva (UE) 2016/680, de 27 de abril, relativa a la protección de las personas físicas en lo que respecta al tratamiento de datos personales por parte de las autoridades competentes para fines de prevención, investigación, detección o enjuiciamiento de infracciones penales o de ejecución de sanciones penales, y a la libre circulación de dichos datos.
  https://www.boe.es/doue/2016/119/L00089-00131.pdf
- Directiva (UE) 2016/681 del Parlamento Europeo y del Consejo, de 27 de abril de 2016, relativa a la utilización de datos del registro de nombres de los pasajeros (PNR) para la prevención, detección, investigación y enjuiciamiento de los delitos de terrorismo y de la delincuencia grave.
  https://www.boe.es/doue/2016/119/L00132-00149.pdf

Recomendaciones:

Resolución del Parlamento Europeo, de 14 de marzo de 2017, sobre las implicaciones de los macrodatos en los derechos fundamentales: privacidad, protección de datos, no discriminación, seguridad y aplicación de la ley.
https://www.europarl.europa.eu/doceo/document/TA-8-2017-0076_ES.html

6.- Foco en el ser humano: control y vigilancia humana

Normativa obligatoria:

- Reglamento (UE) 2021/1232 del Parlamento Europeo y del Consejo de 14 de julio de 2021 por el que se establece una excepción temporal a determinadas disposiciones de la Directiva 2002/58/CE en lo que respecta al uso de tecnologías por proveedores de servicios de comunicaciones interpersonales independientes de la numeración para el tratamiento de datos personales y de otro tipo con fines de lucha contra los abusos sexuales de menores en línea.
  https://www.boe.es/doue/2021/274/L00041-00051.pdf
- Reglamento (UE) 2021/784 del Parlamento Europeo y del Consejo, de 29 de abril de 2021, sobre la lucha contra la difusión de contenidos terroristas en línea.
  https://eur-lex.europa.eu/legal-content/ES/TXT/PDF/?uri=CELEX:32021R0784&from=ES
- Artículo 22.3 del Reglamento General de Protección de Datos.
  https://www.boe.es/doue/2016/119/L00001-00088.pdf
- Real Decreto-ley 24/2021, de 2 de noviembre, de transposición de directivas de la Unión Europea en las materias de bonos garantizados, distribución transfronteriza de organismos de inversión colectiva, datos abiertos y reutilización de la información del sector público, ejercicio de derechos de autor y derechos afines aplicables a determinadas transmisiones en línea y a las retransmisiones de programas de radio y televisión, exenciones temporales a determinadas importaciones y suministros, de personas consumidoras y para la promoción de vehículos de transporte por carretera limpios y energéticamente eficientes.
  https://www.boe.es/buscar/doc.php?id=BOE-A-2021-17910
- Ley Orgánica 7/2021, de 26 de mayo, de protección de datos personales tratados para fines de prevención, detección, investigación y enjuiciamiento de infracciones penales y de ejecución de sanciones penales.
  https://www.boe.es/buscar/act.php?id=BOE-A-2021-8806

Recomendaciones:

Propuesta de Reglamento del Parlamento Europeo y del Consejo por el que se establecen normas armonizadas en materia de inteligencia artificial (Ley de inteligencia artificial) y se modifican determinados actos legislativos de la Unión.

https://eur-lex.europa.eu/resource.html?uri=cellar:e0649735-a372-11eb-9585-01aa75ed71a1.0008.02/DOC_1&format=PDF

7.- Promoción de los valores y de los derechos humanos

Normativa obligatoria:

- Tratado de la Unión Europea.
  https://www.boe.es/doue/2010/083/Z00013-00046.pdf
- Carta de los Derechos Fundamentales de la Unión Europea.
  https://www.boe.es/doue/2010/083/Z00389-00403.pdf
- Convenio Europeo de Derechos Humanos (CEDH).
  https://www.echr.coe.int/documents/convention_spa.pdf
- Reglamento General de Protección de Datos.
  https://www.boe.es/doue/2016/119/L00001-00088.pdf

8.- Sostenibilidad ambiental

Normativa obligatoria:

- Convenio de Aarhus de 25 de junio de 1998.
  https://www.boe.es/boe/dias/2005/02/16/pdfs/A05535-05547.pdf
- Artículo 80 del Reglamento General de Protección de Datos.
  https://www.boe.es/doue/2016/119/L00001-00088.pdf

Recomendaciones:

Recomendaciones 5, 18, 25, 27, 84, 85 y 86 de la Recomendación sobre la Ética de la Inteligencia Artificial.
https://unesdoc.unesco.org/ark:/48223/pf0000380455_spa

Pacto Verde Europeo.
https://ec.europa.eu/info/strategy/priorities-2019-2024/european-green-deal_es

A continuación enlazamos este *framework* ético-normativo con otro tecnológico que permite conocer de qué manera gestionar los principios éticos y su normativa asociada mediante tecnologías, herramientas, recomendaciones y modelos de gobierno.

A continuación enlazamos este *framework* ético-normativo con otro tecnológico que permite conocer de qué manera gestionar los principios éticos y su normativa asociada mediante tecnologías, herramientas, recomendaciones y modelos de gobierno.

El contenido de este capítulo ha sido generado por OdiseIA y PwC a partir de sesiones de trabajo mantenidas con equipos de Google, Microsoft, IBM y Telefónica. A partir de la información recibida en dichas sesiones, nosotros lo hemos extendido analizando:

- La información referenciada por dichas compañías en las sesiones de trabajo.
- La Información pública de las mismas.

Su contenido:

- Es un resumen de dichas fuentes manteniendo los enfoques y contenidos de las mencionadas compañías, sin interpretaciones, y siempre referenciando la fuente original.
- En el caso de Google, Microsoft e IBM tiene una estructura global homogénea entre ellas, para facilitar así la lectura y el entendimiento de los planteamientos de las tres compañías.
  - Su enfoque respecto de la IA ética.
  - Buenas prácticas generales.
  - Principios éticos sobre los que se focalizan.
  - Sus recursos (tecnologías, herramientas, guidelines, etc.) para gestionar los principios éticos a lo largo del ciclo de vida.
- En el caso de Telefónica su estructura es diferente, ya que se expone el caso de éxito de dicha compañía a la hora de adoptar la una manera ética y responsable.

De esta manera iremos profundizando desde una visión global a una aterrizada, permitiendo que cualquier entidad pueda entender planteamientos globales que poder adoptar en su estrategia, su organización y procesos de gestión hasta llegar al detalle de cómo hacerlo en el día a día de sus operaciones, en el Delivery de sus proyectos que tengan a la Inteligencia Artificial como su tecnología principal.

3. Framework GuIA

Cómo aterrizar cada principio ético

El enfoque de Google

El enfoque de Google para gestionar los principios éticos

El contenido de este apartado ha sido generado a partir de sesiones de trabajo mantenidas con el equipo de *Google Research* (Pilar Manchón, Fernanda Viégas y Kathy Meier-Hellstern), gracias a la mediación de Josetxo Soria Checa, cuyo contenido hemos extendido analizando:

- La información referenciada por dicho equipo en las sesiones de trabajo.
- La Información pública de Google relacionada con las temáticas tratadas en dichas sesiones de trabajo.

Este apartado:

- Es un resumen de dichas fuentes manteniendo los enfoques y contenidos de Google, sin interpretaciones, y siempre referenciando la fuente original.
- Tiene una estructura global homogénea respecto de la Microsoft e IBM expuesta en este mismo documento GuIA, para facilitar así la lectura y el entendimiento de los planteamientos de las tres compañías.

En esta sección dedicada a los planteamientos de Google vamos a desarrollar el marco de trabajo de dicha compañía para conseguir una IA ética. Para ello profundizaremos sobre los siguientes contenidos:

- Enfoque de Google respecto de la IA ética.
- Buenas prácticas generales.
- Principios éticos sobre los que Google se focaliza (definición, buenas prácticas recomendadas, investigación, y recursos).
- Los recursos (tecnologías, herramientas, guías, etc.) de Google para gestionar los principios a lo largo del ciclo de vida.

De esta manera iremos profundizando desde una visión global a una aterrizada, permitiendo que cualquier entidad pueda entender planteamientos globales que poder adoptar en su estrategia, su organización y procesos de gestión hasta llegar al detalle de cómo hacerlo en el día a día de sus operaciones, en el *Delivery* de sus proyectos que tengan a la inteligencia artificial como su tecnología principal.

Enfoque de Google respecto de la IA ética

Google considera que la inteligencia artificial (IA) lo está revolucionando todo y ayudando a superar muchos de los desafíos más acuciantes a gran escala. No obstante, esta increíble oportunidad conlleva la responsabilidad de desarrollar una IA que beneficie a todos.

Propósitos globales de Google respecto de la IA ética (1)

Google aspira a crear tecnologías que resuelvan problemas importantes y ayuden a las personas en su vida diaria. Son optimistas sobre el increíble potencial de la IA y otras tecnologías avanzadas para empoderar a las personas, beneficiar ampliamente a las generaciones actuales y futuras y trabajar por el bien común.

Google cree que estas tecnologías promoverán la innovación y promoverán su misión de organizar la información del mundo y hacerla universalmente accesible y útil.

(1) https://ai.google/principles/#:~:text=Avoid%20creating%20or%20reinforcing%20unfair%20bias.&text=We%20will%20seek%20to%20avoid,and%20political%20or%20religious%20belief.

Reconocen que estas mismas tecnologías también plantean desafíos importantes que debemos abordar de manera clara, reflexiva y afirmativa. Estos principios establecen su compromiso de desarrollar tecnología de manera responsable y establecen áreas de aplicación específicas que no abordarán. Evaluaremos las aplicaciones de IA en vista de los siguientes objetivos. **Google cree que la IA debería**:

Aportar un beneficio a la sociedad

El alcance ampliado de las nuevas tecnologías cada vez tiene más impacto en la sociedad. Los avances en IA tendrán un impacto transformador en una amplia variedad de ámbitos, incluidos la salud, la seguridad, la energía, el transporte, la construcción y el entretenimiento. Cuando se considera un potencial desarrollo y uso de tecnologías de IA, se tienen en cuenta numerosos factores sociales y económicos, y solo se procederá al desarrollo en los casos en los que comités multidisciplinares internos de la compañía evalúen si los beneficios generales superan sustancialmente a los riesgos y desventajas previsibles. Además, será fundamental el seguimiento estricto, proactivo y anticipado de las normativas más restrictivas.

Evitar crear o reforzar sesgos

Los datos y los algoritmos de IA pueden reflejar, reforzar o reducir los sesgos injustos. Google reconoce que distinguir los sesgos justos de los injustos no siempre es simple, y difiere entre culturas y sociedades. Se busca evitar impacto injusto en las personas, particularmente aquel relacionado con aspectos como raza, etnia, género, nacionalidad, ingresos, orientación sexual, o capacidades y creencias políticas o religiosas.

Que el modelo sea construido y testeado siguiendo los estándares de seguridad

Google desarrolla y aplica sólidas prácticas de seguridad y protección para evitar resultados no deseados que generen riesgos dañinos. Diseñan sus sistemas de IA para que sean cautelosos y buscan desarrollarlos de acuerdo con las mejores prácticas actuales en temas de seguridad de la IA. Cuando procede, se prueban las tecnologías de IA en entornos restringidos y se monitoriza su funcionamiento después de la implementación.

Mostrar responsabilidad ante la sociedad

Diseñar sistemas de IA que promuevan el *feedback*, proporcionando explicaciones relevantes y resulten atractivos para las personas.

Incorporar principios de un diseño centrado en la privacidad

Google incorpora sus principios de privacidad en el desarrollo y uso de sus soluciones de IA. Fomentan arquitecturas con que garanticen privacidad y proporciona transparencia y control adecuados sobre el uso de los datos.

Mantener un alto estándar de excelencia científica

La innovación tecnológica tiene sus raíces en el método científico y el compromiso con la investigación abierta, el rigor intelectual, la integridad y la colaboración. Las herramientas de IA tienen el potencial de desbloquear nuevos campos de investigación y conocimiento científicos en áreas críticas como la biología, la química, la medicina y las ciencias ambientales. Google aspira a los más altos estándares de excelencia científica mientras trabaja para avanzar en el desarrollo de la IA.

Estar disponible para usos que estén de acuerdo con estos principios

Muchas tecnologías tienen múltiples usos. Google trabaja para limitar las aplicaciones potencialmente dañinas o abusivas. A medida que se desarrollan e implementan tecnologías de IA, se evalúan los usos más probables de acuerdo con los siguientes factores:

- **Propósito y uso principal**: el propósito principal y el uso probable de una tecnología y aplicación, incluido el grado de relación o adaptabilidad de la solución a un uso nocivo.
- **Naturaleza y singularidad**: si se está poniendo a disposición pública una tecnología que es única o una más accesible.
- **Escala**: si el uso de esta tecnología tendrá un impacto significativo.
- **Naturaleza de la participación de Google**: si proporcionan herramientas de propósito general, integran herramientas para clientes o desarrollan soluciones personalizadas.

El enfoque de Google

De igual manera, Google define las **aplicaciones de la IA en las que no está interesada** en participar, teniendo en cuenta que a medida que profundizan en su experiencia en la inteligencia artificial, **esta lista puede evolucionar**:

- Tecnologías que causan o es probable que causen un daño general. Cuando exista un riesgo material de daño, Google procederá solo cuando crea que los beneficios superan sustancialmente los riesgos e incorporarán las restricciones de seguridad adecuadas.
- Armas u otras tecnologías cuyo propósito principal o implementación es causar o facilitar directamente lesiones a las personas.
- Tecnologías que recopilan o utilizan información para la vigilancia violando las normas internacionalmente aceptadas.
- Tecnologías cuyo propósito contraviene los principios ampliamente aceptados del derecho internacional y los derechos humanos.

Estrategia de Google para lograr una IA responsable (1)

Principios de la IA

Desde el 2018, los principios de la IA de Google han sido sus pilares y les han motivado a aunar esfuerzos por un mismo fin. El equipo de Innovación responsable marca las directrices que deben seguir para poner en práctica estos principios en toda su organización y guía la estrategia de Google en cuanto al desarrollo de tecnologías avanzadas, la realización de investigaciones y la redacción de sus políticas.

Llevar a la práctica sus principios

Hacer evaluaciones estrictas es un paso clave para desarrollar una IA con éxito. Para asegurar que en Google se ajustan a sus principios de la IA, dos organismos de revisión distintos realizan rigurosos análisis éticos y evaluaciones de riesgos y oportunidades tanto de todos los productos tecnológicos que desarrollan como de los acuerdos alcanzados en las etapas iniciales que implican encargos personalizados.

Herramientas y formación

Las herramientas de IA responsable son una forma cada vez más efectiva de analizar modelos de IA y entender cómo funcionan. Google desarrolla recursos como la IA explicable, *Model Cards* y el conjunto de herramientas de código abierto de *TensorFlow* para proporcionar transparencia en sus modelos de una forma estructurada y accesible.

(1) https://cloud.google.com/responsible-ai#section

Los beneficios de una IA fundamentada en valores (1)

Google expone una serie de beneficios que, a según su criterio, son el resultado de adoptar una IA fundamentada en valores y principios éticos:

Permite ofrecer productos más seguros y responsables

Las tecnologías avanzadas tienen más éxito cuando sus ventajas están al alcance de todo el mundo. Evaluar los sistemas de IA, tanto cuando funcionan según lo previsto como cuando no, es un factor crucial para desarrollar productos responsables.

Puede ayudar a ganar y mantener la confianza de los clientes

Debido en parte a la desconfianza en la IA, la implementación de estos sistemas aún no ha arraigado en el ámbito empresarial, y cada vez son más las organizaciones que eligen un producto de IA en función de sus prácticas responsables. Con una estrategia de IA responsable se puede ganar y mantener la confianza de tus clientes.

Facilita una cultura de innovación responsable

Si se proporcionan herramientas adecuadas a los desarrolladores y responsables de toma de decisiones de IA que les permitan tener en cuenta los aspectos éticos, podrán encontrar nuevas formas innovadoras de llevar a cabo los proyectos.

(1) https://cloud.google.com/responsible-ai#section

Cuidar de los negocios con IA responsable (1)

En este apartado reproducimos un resumen de un artículo de *Tracy* (*Tracy Pizzo*) Frey. Una de sus responsabilidades está dentro de la IA responsable. En el artículo, *Tracy* integra sus experiencias con los planteamientos de Google a la hora de poner en práctica los principios de IA, con algunas reflexiones que son comunes en la industria de la IA responsable.

En menos de 10 años, algunos predicen que la inteligencia artificial (IA) puede ser el principal impulsor del crecimiento del PIB mundial. Esta es una predicción asombrosa. Durante la próxima década, veremos una adopción e innovación increíbles

A pesar de este entusiasmo, la confianza en la IA es una barrera cada vez mayor para la adopción por parte de las empresas. En una encuesta de ejecutivos de negocios globales, más del 90% informó haber encontrado problemas éticos en relación con la adopción de un sistema de IA. De estos, el 40% abandona completamente el proyecto. Sin una evaluación sólida de las preocupaciones éticas y la construcción y el despliegue responsables de la IA, corremos el riesgo de que los beneficios de esta tecnología no se materialicen.

En Google, creen que las evaluaciones rigurosas de cómo crear IA de manera responsable no solo son lo correcto, sino que son un componente fundamental para crear una IA exitosa.

Comenzamos a desarrollar nuestros Principios de IA a mediados de 2017 y los publicamos un año después, en junio de 2018. Son una constitución viva que usamos para guiar nuestro enfoque para desarrollar tecnologías avanzadas, realizar investigaciones y redactar nuestras políticas.

Nuestros Principios de IA nos mantienen motivados por un propósito común, nos guían para usar tecnologías avanzadas en el mejor interés de las sociedades de todo el mundo y nos ayudan a tomar decisiones que están alineadas con la misión y los valores fundamentales de Google. También son inseparables del éxito a largo plazo de la IA implementada.

Después de este tiempo, lo que sigue siendo cierto es que nuestros Principios de IA rara vez nos dan respuestas directas a nuestras preguntas sobre cómo construir nuestros productos. No nos permiten, y no deberían, permitirnos eludir conversaciones difíciles. Son una base que establece lo que representamos, lo que construimos y por qué lo construimos, y son fundamentales para el éxito de nuestras ofertas de IA empresarial.

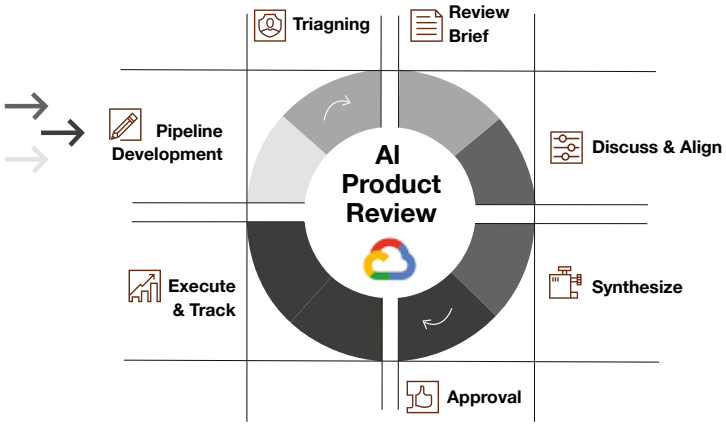
(1) https://cloud.google.com/blog/products/ai-machine-learning/taking-care-of-business-with-responsible-ai

Cómo Google pone en práctica sus Principios de IA

Nuestros procesos de gobierno están diseñados para implementar nuestros Principios de IA de manera sistemática y repetible. Estos procesos incluyen: revisiones de productos y acuerdos, mejores prácticas para el desarrollo de aprendizaje automático, educación interna y externa, herramientas y productos como la IA explicable de *Cloud*, así como orientación sobre cómo consultamos y trabajamos con nuestros clientes y socios.

En *Google Cloud* existen dos procesos de revisión separados. Uno se enfoca en los productos que construimos con tecnologías avanzadas, y el otro se enfoca en acuerdos en etapas iniciales que involucran trabajo personalizado más allá de nuestros productos generalmente disponibles.

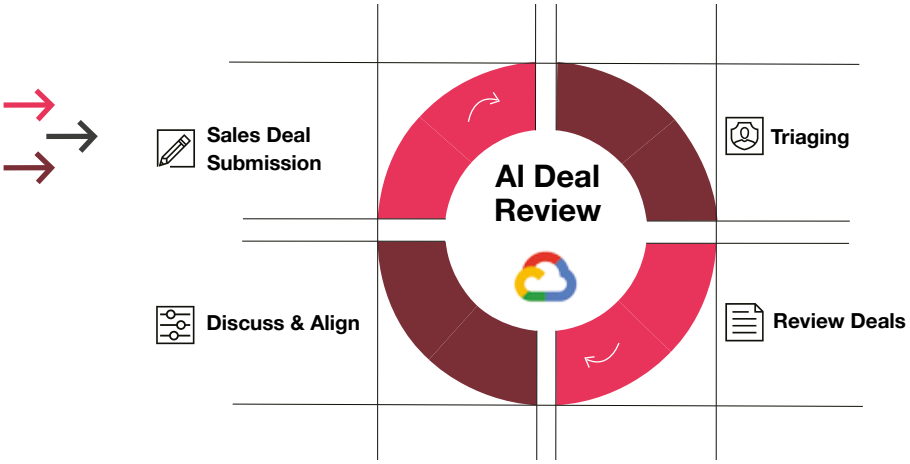
Alinear nuestro desarrollo de productos



Cuando nos propusimos desarrollar nuevos productos y servicios que involucran tecnologías avanzadas, así como productos y servicios similares que ya estaban disponibles en general cuando comenzamos este proceso, llevamos a cabo análisis éticos rigurosos y profundos, evaluando riesgos y oportunidades en cada principio y realizando sólidas revisiones en vivo que pueden involucrar conversaciones complejas, pero a menudo inspiradoras. Discutimos libremente temas críticos pero desafiantes, como la equidad del aprendizaje automático, el sesgo implícito, cómo nuestras propias experiencias pueden diferir en gran medida de las que podrían verse afectadas por la IA y una variedad de otras consideraciones que pueden afectar si avanzamos con un producto determinado. Dedicamos un tiempo considerable a crear y mantener la seguridad psicológica entre nuestros equipos de revisión para garantizar que todas las voces y los temas sean escuchados y valorados.

Alinear las ofertas de nuestros clientes usando IA personalizada

También hacemos una revisión de nuestros acuerdos comerciales en etapa inicial, mucho antes de que se firmen, lo que implicará que construyamos una IA personalizada para un cliente. Trabajamos para determinar desde el principio si el proyecto implicará el uso de una tecnología avanzada de manera que pueda contravenir nuestros Principios de IA.



Fuente: Google

A medida que adquirimos experiencia en nuestras evaluaciones, desarrollamos "jurisprudencia" a lo largo del tiempo y pensamos profundamente sobre dónde podríamos trazar las líneas de nuestra responsabilidad, hemos reiterado lo que consideramos que está dentro o fuera del alcance de nuestro compromiso. Debido a que hemos realizado revisiones exhaustivas de nuestros productos disponibles de forma general y hemos creado planes de alineación para cada uno de ellos, hemos podido centrar nuestras revisiones de acuerdos en casos de uso únicos y personalizados sobre cómo se podrían aplicar nuestros productos de disponibilidad general.

Nuestras revisiones de productos informan nuestra hoja de ruta y nuestras mejores prácticas

En los últimos años, hemos desarrollado una serie de mejores prácticas que a menudo comparto cuando trabajo con organizaciones que buscan implementar sus propios principios. Estos son:

- **No hay lista de verificación de ética**. Puede ser tentador crear una lista de árboles de decisión que fácilmente catalogue un conjunto de cosas como buenas y un conjunto de otras cosas como malas. Estos enfoques no siempre son operativos ya que las intersecciones de la tecnología, los datos, el caso de uso y dónde y cómo se aplican estas listas de verificación es compleja y pueden ser similares, pero nunca iguales.
- **Responsabilidad por diseño**. Nuestro objetivo es realizar revisiones en profundidad al principio del ciclo de vida del desarrollo, lo que ha demostrado ser esencial para garantizar la alineación.
- **Diversidad**. Los miembros permanentes de nuestros comités de revisión son multifuncionales, técnicos y no técnicos. Múltiples miembros tienen formación en ciencias sociales, filosofía y ética. Somos deliberadamente multinivel, desde miembros desde el principio de su carrera hasta altos ejecutivos. Los aportes internos y externos son una prioridad para nosotros, especialmente cuando nuestras propias experiencias vividas no pueden informar de manera sólida nuestras decisiones. El aporte directo de expertos en dominios externos y grupos afectados también es clave. Por ejemplo, las Evaluaciones de Impacto en los Derechos Humanos proporcionadas por organizaciones como BSR como un componente de nuestra Diligencia en Derechos Humanos nos han ayudado enormemente. Garantizar que se escuchen todas las voces es esencial, y no siempre es fácil.
- **Una cultura de apoyo**. Un mandato de alto nivel es necesario pero no suficiente. La transformación cultural en toda una organización es clave. Parte de esto proviene de la capacitación en ética tecnológica para conectar activamente la ética con la tecnología que, de otro modo, podría considerarse neutral en cuanto a valores. No siempre es cómodo hablar sobre cómo estas tecnologías transformadoras pueden ser dañinas, pero es importante construir ese compromiso modelándolo, incluso cuando sea difícil.
- **Transparencia**. La confianza es el núcleo de por qué tenemos estos principios y prácticas vigentes. Si bien no somos transparentes con cada decisión, ya que eso iría en contra de nuestros compromisos con la privacidad, creo que ese nivel de transparencia no es necesario. No existe un mundo en el que cualquier decisión que tomemos sea recibida con aprobación universal. Esto está bien. Lo que buscamos es comprender cómo tomamos nuestras decisiones, ya que esto puede generar confianza incluso en medio del desacuerdo.
- **Un enfoque humilde**. La IA está cambiando rápidamente y nuestro mundo no es estático. Tratamos de recordar conscientemente que siempre estamos aprendiendo, y aunque certificar un producto como perfecto es una meta inalcanzable, siempre podemos mejorar.
- **El trabajo no es (fácilmente) medible**. La alineación de los principios es a menudo no cuantificable, subjetiva, culturalmente relativa e imposible de precisar. Sin embargo, es precisamente porque existen estos desafíos que tomamos medidas.

El enfoque de Google

Enfoques respecto de la gobernanza

La IA puede aportar grandes beneficios a las economías, a la Sociedad y apoyar una toma de decisiones más justa, segura, inclusiva e informada. Pero esta promesa no se hará realidad, sin tener en cuenta la importancia de cómo debe gobernarse su desarrollo y uso, y qué grado de supervisión legal y ética es necesario.

Hasta la fecha, los enfoques autorreguladores y co-reguladores basados en la legislación vigente y en las perspectivas de las empresas, el mundo académico y los organismos técnicos asociados han tenido mucho éxito a la hora de frenar el uso de la IA no garantista. Google entiende necesario que el regulador actúe para aportar la contribución necesaria a este debate abierto sobre la gobernanza de la IA. En concreto, Google destaca cinco áreas (1) en las que los reguladores, en colaboración con la Sociedad civil y los profesionales de la IA, deben desempeñar un papel clave a la hora de aclarar las expectativas respecto del uso de la IA:

Normas de explicabilidad	• Reunir una colección de explicaciones de mejores prácticas junto con comentarios sobre sus características para proporcionar referencias y que sirvan de inspiración práctica. • Proporcionar directrices para casos de uso hipotéticos, de modo que la industria pueda calibrar cómo equilibrar las ventajas de utilizar sistemas complejos de IA con las limitaciones prácticas que imponen los distintos estándares de explicabilidad. • Describir los estándares mínimos aceptables en diferentes sectores industriales y contextos de aplicación.
Valoración de la equidad	• Articular marcos que equilibren objetivos y diferentes definiciones de equidad. • Aclarar la prioridad relativa de los factores que compiten entre sí en algunas situaciones hipotéticas comunes, considerando la posibilidad de que esto difiera entre culturas.
Consideraciones de seguridad	• Esbozar flujos de trabajo básicos y normas documentadas para contextos de aplicación específicos que sean suficientes para demostrar el cumplimiento en la realización de controles de seguridad. • Establecer marcas de certificación de seguridad para significar que un servicio ha sido evaluado para superar las pruebas especificadas para aplicaciones críticas.
Colaboración entre humanos e inteligencia artificial	• Determinar los contextos en los que la toma de decisiones no debería ser totalmente automatizada por un sistema de IA, sino que requeriría un “humano en el bucle (2)“ (en inglés, *human-in-the-loop).* • Evaluar diferentes enfoques para permitir la revisión y supervisión humana de los sistemas de IA.
Marcos de responsabilidad	• Evaluar las posibles deficiencias de las normas de responsabilidad existentes y explorar normas complementarias para aplicaciones específicas de alto riesgo. • Considerar la posibilidad de establecer marcos de seguridad específicos para cada sector y límites de responsabilidad en ámbitos en los que se prevea que las leyes de responsabilidad puedan desalentar la innovación socialmente beneficiosa. • Explorar alternativas de coberturas de seguridad cuando se desarrolle IA en ámbitos territoriales o espacios en los que las normas tradicionales de responsabilidad son inadecuadas o inviables.

Como afirma Google, aunque las diferentes sensibilidades y prioridades culturales pueden dar lugar a variaciones entre las regiones, debería ser posible acordar una lista de control de alto nivel de los factores que deben tenerse en cuenta en el desarrollo y uso de la IA, y en un largo plazo, trabajar con organismos de normalización (como ISO e IEEE) para establecer algunas normas mundiales como procesos de mejores prácticas para demostrar el cumplimiento ("diligencia debida") en relación con el desarrollo y la aplicación de la IA.

(1) https://ai.google/static/documents/perspectives-on-issues-in-ai-governance.pdf

(2) El humano en el bucle (*human-in-the-loop*) es la persona que forma, prueba o ajusta un sistema de IA para ayudarlo a producir resultados más fiables. Se trata de dar supervisión humana a un sistema de inteligencia artificial y aprovechar así de forma simultánea dos tipos totalmente diferentes de inteligencia. Por ejemplo, piénsese en un sistema al que se enseña a identificar la vida marina, puede aprender fácilmente a distinguir un pulpo de otras criaturas debido a su forma única, pero le cuesta distinguir a peces que tienen formas y colores similares. Un humano en el bucle puede intervenir para enseñarle al sistema qué características distintivas debe tener en cuenta y orientarlo hacia respuestas más precisas.

Marcos de responsabilidad

Existen un amplio debate sobre otorgar personalidad jurídica propia a los sistemas de IA y robots, y por tanto que sean considerados sujetos responsables. En este debate, Google es claro al afirmar que las organizaciones deben seguir siendo responsables de las decisiones que tomen y de la forma de actuar, que debe hacerse en consecuencia (ya sea con IA, con humanos o con ambos). Por muy complejo que sea el sistema de IA, deben ser las personas u organizaciones las responsables en última instancia de las acciones de los sistemas de IA diseñados o controlados por ellos garantizando mecanismos adecuados para su gestión y supervisión.

Google no considera acertado otorgar personalidad jurídica a la IA o a los robots, sino que debe ser una responsabilidad de las personas u organizaciones, por las siguientes razones (1) :

- No es necesario: Siempre habrá una persona física o una organización responsable dentro de la legislación aplicable.
- Es inviable: Incluso si fuera posible llegar a una definición de los robots o la IA que garantice la personalidad jurídica (lo que no es ni mucho menos un hecho), sería imposible hacer que esas entidades respondieran por el incumplimiento de sus obligaciones. Por decirlo de otro modo, ¿cómo se puede castigar a una máquina que carece de conciencia o sentimientos?
- Es inmoral: la responsabilidad es una propiedad intrínsecamente humana. Es moralmente inapropiado trasladar la responsabilidad a “personas sintéticas” en forma de máquinas o códigos.
- Se presta a abusos: Facilitaría que los malos actores se protegieran de la responsabilidad por las actividades ilegales realizadas por las máquinas que han creado.

Sin embargo, las cosas están menos claras en lo que respecta a las expectativas de comportamiento que se aplican a los proveedores de IA. Pocas organizaciones fuera del ámbito de la tecnología desarrollarán sus sistemas de IA únicamente utilizando la experiencia interna. Lo más habitual es que colaboraren con proveedores de IA externos, que tienen la experiencia y las herramientas para ayudar a diseñar y poner en funcionamiento un sistema de IA que satisfaga necesidades de la organización, mucho más rápido y con mayor calidad. La responsabilidad de los proveedores de IA es ayudar a sus clientes a comprender los riesgos inherentes al uso de los sistemas de IA, para que puedan tomar decisiones sobre cómo mitigar y vigilar los riesgos que comportan esta tecnología (por ejemplo, advertir sobre las limitaciones de rendimiento de los modelos estándar).

Google recomienda al regulador un enfoque cauteloso con respecto a la responsabilidad en los sistemas de IA, ya que los marcos equivocados podrían culpar injustamente, perjudicar a la innovación o incluso reducir la seguridad. Cualquier cambio en el marco general de la responsabilidad debería venir sólo después de una investigación exhaustiva que establezca el fracaso de las leyes existentes. En caso de que se identifique la necesidad de actuar en áreas que impliquen mayores riesgos para los usuarios finales (por ejemplo, la atención sanitaria y la investigación en materia de salud, los servicios financieros, o el transporte), esto debería ocurrir de forma específica para cada sector, añadiendo una nueva normativa sólo cuando haya una laguna clara y de forma que se minimice el impacto.

En definitiva, Google cree que el régimen de gobernanza óptimo es aquel que es flexible y capaz de seguir el ritmo de los avances, respetando al mismo tiempo las diferencias culturales. Así, los enfoques de autorregulación y co-regulación seguirán siendo la forma práctica más eficaz de abordar y prevenir los problemas relacionados con la IA en la gran mayoría de los casos, dentro de los límites ya establecidos por la normativa sectorial.

(1) En abril de 2018, un grupo de destacados expertos en IA y robots expusieron en una carta abierta a la Comisión Europea su oposición a la noción de personalidad jurídica para la IA y los robots. Google comparte esta opinion: https:// bit.ly/2xfMToe. Para más información sobre este tema, véase Bryson, J.J. et al. 2017 “*Of, for, and by the people: the legal lacuna of synthetic persons”* at https://bit.ly/2Auvjf

Buenas prácticas generales

Google está comprometido a avanzar en el desarrollo responsable de la IA y a compartir sus conocimientos, investigaciones, herramientas, conjuntos de datos y otros recursos con toda la comunidad.

Los sistemas de IA centrados en el usuario confiables y efectivos deben diseñarse siguiendo las mejores prácticas generales para los sistemas de software, junto con prácticas que aborden consideraciones exclusivas del aprendizaje automático que desarrollamos a continuación (1):

Utilizar un enfoque de diseño centrado en el ser humano

- La forma en que los usuarios finales experimentan con los sistemas inteligentes es esencial para evaluar el correcto impacto de las predicciones, recomendaciones y decisiones llevadas a cabo por estos.
- La claridad y el control son cruciales para una buena experiencia de usuario, por ello Google recomienda integrar las descripciones adecuadas sobre las características de diseño.
- Otro punto que tiene suma importancia es la asistencia al usuario dependiendo de la situación. Proponer una única respuesta puede ser apropiado cuando existe una alta probabilidad de que dicha objeción satisfaga a la gran mayoría de usuarios y casos de uso. Pero en otros casos, puede ser óptimo que se sugieran varias opciones. Además, técnicamente, es mucho más complicado lograr la precisión adecuada en una única respuesta que en varias.
- También recomienda considerar las posibles ineficiencias de los modelos cuanto antes por parte de los usuarios al principio del proceso de diseño, seguida de continuas pruebas e iteraciones para un pequeño grupo antes de la implementación completa.
- Por último, Google aconseja interactuar con un diverso conjunto de usuarios y escenarios de casos de uso, incorporando los comentarios pertinentes antes y durante el desarrollo del proyecto. Con estas acciones se pretende conseguir crear una gran variedad de perspectivas de usuario en el proyecto y aumentar el número de personas que se beneficiarán de la tecnología.

Identificar múltiples métricas para evaluar la capacitación y el seguimiento

- El uso de varias métricas en lugar de una sola nos proporcionará un mejor entendimiento de las compensaciones entre los diferentes tipos de errores y experiencias. Siempre y cuando nos aseguremos de que las métricas utilizadas son apropiadas para el contexto y los objetivos de nuestro sistema inteligente.
- Google recomienda considerar métricas que incluyan comentarios de encuestas de usuarios que rastreen el rendimiento general del sistema y el estado del producto tanto a corto como a largo plazo y que consideren también las tasas de falsos positivos y falsos negativos.
- Asegurar que las métricas utilizadas para evaluar un modelo sean apropiadas para éste. Por ejemplo, una alarma debe tener siempre un alto *recall*, en castellano exhaustividad, que se conoce como la ratio del número de sucesos relevantes acertados respecto al número total de sucesos relevantes, aunque ello implique alguna falsa alarma. En otras palabras, es importante que la alarma salte siempre que exista un riesgo, aunque ello pueda ocasionar alguna falsa alarma. El riesgo de no detectar un problema real es mucho más relevante que se ocasionen falsas alarmas.
- Google también está investigando proactivamente en nuevas métricas que garanticen no solo el rendimiento de los sistemas, sino también su adecuación al objetivo del servicio y su capacidad de adaptación al contexto. En los sistemas inteligentes, las métricas absolutas son importantes, pero es posible que no alcancen a medir los aspectos más innovadores de las nuevas generaciones de sistemas inteligentes.

(1) https://ai.google/responsibilities/responsible-ai-practices/

Cuando sea posible, examinar directamente los datos sin procesar

Los modelos de aprendizaje automático reflejan los datos en los que están entrenados, por lo que se deben analizar los datos con cuidado antes de su procesamiento para asegurarnos de que se comprenden.

En esta recomendación, Google propone considerar los siguientes puntos:

- Asegurarse de que los datos de partida no contienen ningún error; por ejemplo, etiquetas incorrectas, valores que faltan, etc.
- Ratificar que la manera en la que se muestrean los datos representa a todos los usuarios (por ejemplo, si el modelo va a ser usado para todas las edades no se pueden utilizar datos solo de gente adulta) o la configuración del mundo real (si el sistema se usará durante todo el año, no podemos utilizar solo datos de verano).
- Para conseguir hacer frente al sesgo entrenamiento-servicio (diferencia entre el rendimiento del sistema durante el entrenamiento y el rendimiento durante el servicio), durante el entrenamiento se deben intentar identificar posibles sesgos y trabajar para abordarlos, incluso ajustando los datos de entrenamiento. De igual manera, durante la evaluación, se debe continuar intentado obtener datos de evaluación del entorno implementado que sean lo más representativos posibles.
- Utilizar el modelo más simple que cumpla con nuestros objetivos de rendimiento, es decir, eliminar aquellas características de nuestro modelo que sean redundantes o innecesarias.

Comprender las limitaciones de nuestro conjunto de datos y modelo

- Un modelo que ha sido entrenado para detectar correlaciones no debe utilizarse para hacer inferencias causales, es decir, no implica que pueda hacerlo. Por ejemplo, nuestro modelo puede aprender que las personas que compran zapatillas de baloncesto en promedio son más altas, pero esto no implica que un usuario que compra zapatillas de baloncesto vaya a ser más alto como resultado de esa compra. En otras palabras, correlación no implica causalidad.
- Los modelos de aprendizaje automático actuales son en gran medida un reflejo de los patrones de sus datos de entrenamiento. Por tanto, es importante comunicar el alcance del entrenamiento, aclarando así la capacidad y las limitaciones de estos. Por ejemplo, un detector de zapatos entrenado con fotos de archivo puede funcionar mejor con fotos de archivo, pero tiene una capacidad limitada cuando se prueba con fotos de teléfonos móviles generadas por el usuario.
- Por ello, siempre se recomienda comunicar dichas limitaciones a los usuarios cuando sea posible y de esta manera también se podrá conseguir una mejor retroalimentación de los usuarios sobre el funcionamiento del sistema.

El enfoque de Google

Probar, probar y probar

Google considera de gran importancia el realizar todo tipo de pruebas en todos y cada uno de los puntos del ciclo de vida para conseguir acercarnos lo máximo posible al control absoluto del sistema. Para ello recomienda:

- Aprender de las mejores prácticas de prueba de ingeniería de software y calidad para asegurarnos de que el sistema de inteligencia artificial funciona según lo previsto y sea de confianza.
- Realizar pruebas unitarias rigurosas para probar cada componente del sistema de forma aislada.
- Realizar pruebas de integración para comprender cómo los componentes individuales de aprendizaje automático interactúan con otras partes del sistema en general.
- Detectar de forma proactiva las posibles desviaciones de entrada probando las estadísticas de las entradas al sistema de IA para asegurarnos de que no cambian de forma inesperada.
- Utilizar un conjunto de datos estándar de alta calidad para probar el sistema y asegurarnos de que continúe comportándose como se espera.
- Actualizar el conjunto de pruebas anterior con regularidad de acuerdo con los cambios de usuarios y casos de uso.
- Realizar pruebas de usuario iterativas para incorporar un conjunto diverso de necesidades de los usuarios en los ciclos de desarrollo.
- Aplicar el principio de ingeniería de calidad de *poka-yoke*: crear controles de calidad en un sistema para que, o bien no se den fallos no intencionales, o bien se desencadene una respuesta inmediata (por ejemplo, si una característica importante – características que explican la etiqueta – falta por algún motivo no contemplado, el sistema de IA no generará ninguna acción).

Continuar monitoreando y actualizando el sistema después de la implementación

- Con el monitoreo continuo se garantiza que nuestro modelo tenga un rendimiento adecuado en el mundo real, así como poder valorar los comentarios de los usuarios.
- Cualquier modelo es imperfecto casi por definición, por ello surgirán problemas. Para intentar minimizarlos se recomienda incorporar un tiempo de margen a la planificación de los productos para hacerles frente.
- En cuanto a las soluciones planteadas para dichos problemas que puedan surgir, se aconseja considerar acciones tanto a corto plazo como a largo. Una solución simple puede ayudarnos a resolver un problema de manera rápida, pero puede que no sea la solución óptima a largo plazo, por ello es importante equilibrar las soluciones simples a corto plazo con las soluciones aprendidas a más largo plazo.
- Y, finalmente, antes de actualizar cualquier modelo implementado, se debe analizar en qué se va a diferenciar el candidato de los modelos ya implementados, y cómo la actualización influirá en la calidad general del sistema y la propia experiencia del usuario.

Principios éticos sobre los que Google se focaliza

Google pone el foco en cuatro principios a la hora de trabajar la inteligencia artificial desde un punto de vista ético y responsable. Estos cuatro principios son:

- **Equidad**
- **Interpretabilidad**
- **Privacidad**
- **Seguridad**

Las definiciones de estos principios que se realizan no son dogmáticas. Son la representación que Google otorga a estos principios y, sobre todo, cómo los trata a nivel de modelos de inteligencia artificial, *Machine y Deep Learning*. Es por ello por lo que las definiciones de los principios aquí descritas pueden ser interpretaciones distintas a las que otras entidades puedan realizar, y por eso las explicaremos al principio del capítulo dedicado a cada principio.

El enfoque de Google

Principio de Equidad

En esta sección vamos a desarrollar:

- Qué es la Equidad para Google
- Prácticas recomendadas por Google
  - Diseñar el modelo utilizando objetivos concretos de equidad e inclusión
  - Utilizar conjuntos de datos representativos para entrenar y probar el modelo
  - Verificar el sistema en busca de sesgos injustos
  - Analizar el desempeño
- Investigación sobre la equidad con casos de uso
  - Medición de sesgos
  - Creación de conjuntos de datos y modelos de ML justos
  - Técnicas de mitigación
  - Comprensión de los usuarios y la sociedad
- Recursos y tecnologías para gestionar la equidad

Qué es la Equidad para Google

Los sistemas de inteligencia artificial (IA) están brindando nuevas experiencias y habilidades. Más allá de recomendar libros y programas de televisión, estos pueden utilizarse para realizar tareas más críticas, como predecir la presencia de una afección médica y su gravedad, poner en contacto personas con trabajos o incluso identificar si una persona está cruzando la calle. Estos sistemas de toma de decisiones tienen el potencial de ser más justos e inclusivos que los procesos de toma de decisiones basados en juicios humanos o en reglas determinadas, pero el riesgo de cometer cualquier injusticia con ellos conlleva un impacto a gran escala. Por lo tanto, a medida que aumenta el impacto de la IA en todos los sectores y sociedades, es fundamental trabajar hacia sistemas que sean justos e inclusivos para todos. A continuación, se comentan algunos de los desafíos con mayor notoriedad existentes hoy en día en la consecución de una IA justa:

Los modelos de *Machine Learning* (ML) y *Deep Learning* (DL) **aprenden de los datos existentes recopilados del mundo real, por lo que un modelo preciso puede aprender, o incluso ampliar, los sesgos problemáticos preexistentes** en los datos basados en la raza, el género, la religión u otras características. Por ejemplo, un sistema de búsqueda de empleo puede "aprender" a favorecer a los usuarios masculinos a la hora de seleccionar candidatos para las rondas de entrevistas si dispone de un histórico de contrataciones mayoritariamente masculino.

Incluso con la preparación y las pruebas más rigurosas, **es un gran desafío garantizar que un sistema sea justo en todas las situaciones y entornos**. Por ejemplo, un sistema de reconocimiento de voz apto para adultos estadounidenses puede ser justo e inclusivo en este contexto, pero si lo usan los adolescentes, el sistema puede no ser capaz de reconocer ciertas palabras o expresiones utilizadas por este grupo. Incluso, si el sistema se implementa en España, puede que tenga más dificultad de entendimiento con algunos acentos que con otros. Conforme se utilizan estos sistemas, se pueden descubrir algunos puntos ciegos injustos e involuntarios que en las fases iniciales de desarrollo eran difíciles de identificar.

No existe una definición estándar de equidad, independientemente de que las decisiones las tomen los humanos o las máquinas. La identificación de criterios de equidad adecuados para un sistema requiere tener en consideración distintos aspectos, como la experiencia del usuario, las consideraciones culturales, sociales, históricas, políticas, legales y éticas, varias de las cuales pueden discrepar entre ellas mismas. Incluso en situaciones que pueden parecernos simples, la gente puede no estar de acuerdo sobre lo que es y no es justo y puede que no esté del todo claro qué punto de vista debería venir dictado por ley, especialmente en un contexto globalizado como el actual.

En la actualidad, conseguir una IA justa e inclusiva está siendo una tarea de investigación activa. Un ejemplo de ello son las prácticas llevadas a cabo, como la evaluación de los conjuntos de datos de entrenamiento para evitar posibles fuentes de sesgo, la creación de modelos de entrenamiento para eliminar o corregir dichos sesgos problemáticos o las pruebas continuas de los sistemas finales en busca de resultados injustos. De hecho, los propios modelos de ML también se pueden usar para identificar algunos de los prejuicios y barreras humanas en términos de inclusión que han sido desarrollados, consciente e inconscientemente, y se han perpetuado a lo largo de la historia. Es por ello por lo que la equidad en la IA presenta tanto una oportunidad como un desafío. Por esta razón Google se compromete a avanzar en todas estas áreas y a crear herramientas, conjuntos de datos y otros recursos para la comunidad. A continuación, se expone el pensamiento de Google, en cuanto a prácticas recomendadas, técnicas, recursos y herramientas con las que abordar este tema.

Prácticas recomendadas

En el caso de que el *Machine Learning* pueda ayudar a proporcionar una solución adecuada a alguno de nuestros problemas específicos, al igual que no existe un modelo "correcto" único para todas las tareas de ML, tampoco existe una técnica única que garantice la equidad en cada situación. En la práctica, los investigadores y desarrolladores deberían considerar el uso de una variedad de enfoques para ensayar y mejorar. Por ello desde Google se definen las siguientes recomendaciones para diversas fases de diseño de un modelo IA.

Diseñar el modelo utilizando objetivos concretos de equidad e inclusión

- Resulta conveniente interactuar con profesionales del ámbito de la sociología y otros expertos en las relaciones humanas para que el producto comprenda y tenga en consideración distintas perspectivas.
- Considerar cómo la tecnología y su desarrollo a lo largo del tiempo afectarán a los diferentes casos de uso realizándose, por ejemplo, las siguientes preguntas: ¿Qué puntos de vista están representados? ¿Qué tipo de datos están representados? ¿Qué queda fuera del alcance del sistema? ¿Qué resultados permite esta tecnología y cómo se comparan para diferentes usuarios y grupos? ¿Qué sesgos, experiencias negativas o resultados discriminatorios podrían ocurrir?
- Establecer objetivos para que el sistema funcione de manera justa en los casos de uso previstos: por ejemplo, en 'X' idiomas diferentes o en 'Y' grupos de edad distintos. Supervisar estos objetivos a lo largo del tiempo y ampliarlos según corresponda.
- Si es posible, diseñar los algoritmos para reflejar los objetivos de equidad. En el caso de que se utilicen herramientas de mercado de modelos IA de terceros, analizar que dichos algoritmos ofrecidos por las herramientas cumplan dichos criterios de equidad.
- Actualizar los datos de entrenamiento y prueba con frecuencia en función de quién usa la tecnología y cómo la usa.

Utilizar conjuntos de datos representativos para entrenar y probar el modelo

- Evaluar la equidad en los conjuntos de datos, lo que incluye la identificación de las limitaciones correspondientes, así como la detección de correlaciones perjudiciales o discriminatorias entre características, etiquetas y grupos. La visualización, la agrupación en *clústeres* y la clasificación de datos pueden ayudar con esta evaluación.
- Los conjuntos de datos de entrenamiento públicos a menudo deberán ampliarse para reflejar, de la manera más fidedigna posible, los cambios producidos en los comportamientos de las personas, eventos y atributos del mundo real, sobre los cuales el sistema hará las predicciones.
- Entender los distintos puntos de vista, experiencias y objetivos de las personas que se encargan de definir lo que se quiere medir/predecir con los datos (*data labeling* (1)), conocido en castellano como etiquetado de datos. ¿Cómo entienden el éxito los diferentes trabajadores y cuáles son las compensaciones entre el tiempo dedicado a la tarea y el beneficio de dicha tarea?
- Si se está trabajando con equipos de *labeling*/etiquetado, vincularse estrechamente con ellos para diseñar tareas claras, incentivos y mecanismos de comunicación. Tener en cuenta la variabilidad humana, incluida la accesibilidad, la memoria muscular y los sesgos innatos en las personas, por ejemplo, mediante el uso de un conjunto estándar de preguntas con respuestas conocidas.

(1) *Data labeling*: proceso de identificar datos sin procesar y agregar una o más etiquetas significativas e informativas para proporcionar contexto y permitir el entrenamiento del modelo.

Verificar el sistema en busca de sesgos injustos

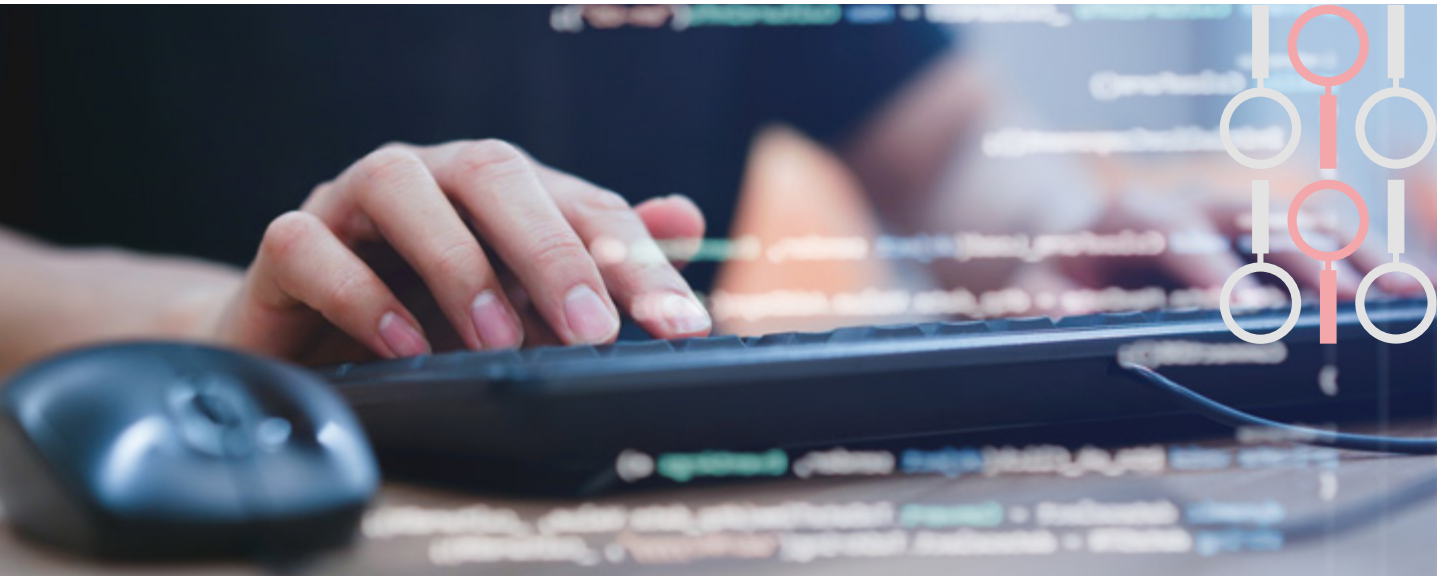
- Por ejemplo, organizar un grupo de personas diverso y con experiencia contrastada que puedan probar el sistema de manera opuesta al funcionamiento lógico del sistema y así incorporar distintas entradas incoherentes en las pruebas unitarias. Esto puede ayudar a la identificación de efectos adversos inesperados y de problemas ocultos en las métricas previamente agregadas al sistema. Incluso una tasa de error baja puede dar lugar a errores ocasionales muy graves.
- Al diseñar métricas para entrenar y evaluar el sistema, también se recomienda incluir métricas para examinar el rendimiento en diferentes subgrupos. Por ejemplo, la tasa de falsos positivos (1) y la de falsos negativos (2) por subgrupo pueden ayudar a comprender qué grupos tienen un funcionamiento desproporcionadamente peor o mejor.
- Además de las métricas estadísticas de división de grupos, es aconsejable crear un conjunto de ensayos que pongan a prueba el sistema en casos complejos. Esto permitirá evaluar rápidamente si el sistema está funcionando bien en casos que pueden ser particularmente perjudiciales o problemáticos cada vez que este se actualice. Este conjunto de estudios, al igual que con todos los conjuntos de prueba, también debe actualizarse continuamente a medida que el sistema evoluciona, se agregan o eliminan funciones y tiene más comentarios de los usuarios.
- Considerar los efectos de los sesgos creados por decisiones tomadas por el sistema anteriormente y los bucles de retroalimentación que esto puede crear.

Analizar el desempeño

- Tener en cuenta las diferentes métricas que han sido definidas. Por ejemplo, la tasa de falsos positivos de un sistema puede variar entre diferentes subgrupos de los datos, y las mejoras en una métrica pueden afectar negativamente a otra.
- Evaluar la experiencia del usuario en escenarios del mundo real en una amplia gama de usuarios, casos de uso y contextos. Primero probar y repetir en las pruebas internas, y luego continuar con pruebas rutinarias después del lanzamiento del producto.
- Incluso si todo en el diseño general del sistema se elabora cuidadosamente para abordar los problemas de equidad, los modelos basados en ML rara vez funcionan con un 100% de perfección cuando se aplican a datos reales y en vivo. Cuando ocurre un problema en un producto en continuo cambio, considerar si se alinea con alguna desventaja social existente y cómo se verá afectado por las soluciones a corto y largo plazo.

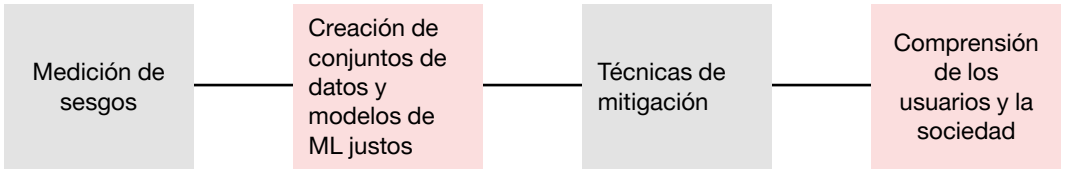
(1) Tasa de falsos positivos: probabilidad de que se dé un resultado positivo cuando el valor real es negativo.

(2) Tasa de falsos negativos: probabilidad de que se dé un resultado negativo cuando el valor real es positivo.



Investigación sobre la equidad con casos de uso

En los últimos tiempos Google ha realizado numerosas investigaciones, junto con el desarrollo de distintas técnicas, para profundizar en los análisis de equidad. En este apartado se nombran algunas de estas publicaciones más destacables en las cuestiones de:



Medición de sesgos

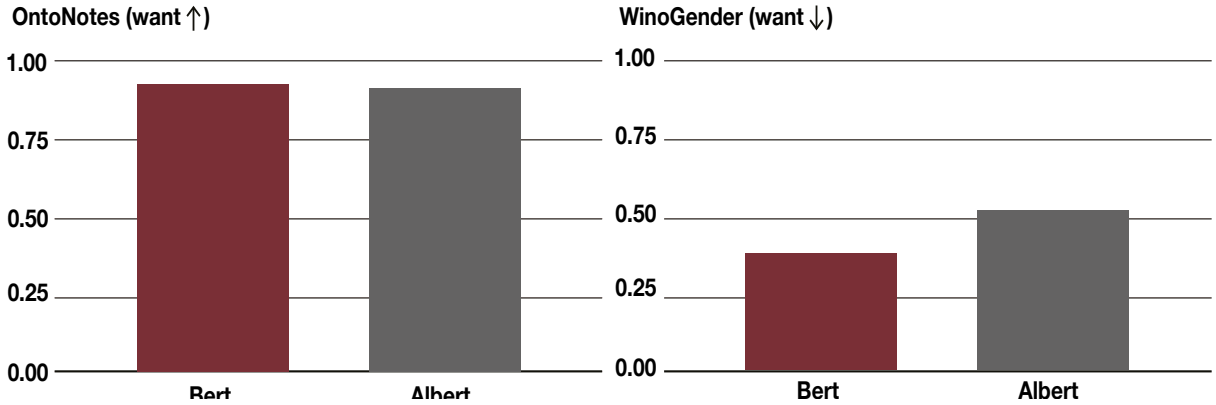
Medición de correlaciones de género en modelos de Procesado de Lenguaje Natural (PLN) previamente entrenados

Este caso de estudio relacionado con BERT (técnica de aprendizaje automático basada en transformadores para la formación previa del procesamiento del lenguaje natural desarrollada por Google) y ALBERT (equivalente de BERT de menor capacidad), analiza las correlaciones vinculadas con el género en los datos de entrenamiento y plantea una serie de prácticas recomendadas para usar modelos de PLN (Procesamiento de Lenguaje Natural).

> La enfermera notificó al paciente que su turno terminaría en una hora

Una de las pruebas realizadas en el estudio fue el análisis de la correferencia, que es la capacidad que permite a los modelos comprender la palabra antecesora correcta de un pronombre dado en una oración. Por ejemplo, en la siguiente oración, el modelo debe reconocer que el pronombre “su” se refiere a la enfermera y no al paciente.

En este ensayo, se utilizaron dos pruebas para medir la precisión de la correferencia de un modelo, la prueba *OntoNotes* y la *WinoGender*. La primera representa una distribución de datos y de ahí se concluye la precisión; y la segunda proporciona datos adicionales diseñados para identificar cuándo las asociaciones de los modelos entre género y profesión influyen de manera negativa en la correferencia. Los valores altos de la métrica *WinoGender* (cercanos a 1) indican que el modelo está basando sus decisiones en asociaciones normativas entre género y profesión (por ejemplo, asociar a la enfermera con el género femenino y no masculino). Cuando la puntuación es cero, el modelo no tiene una asociación preconcebida entre género y profesión, lo que sugiere que las decisiones se basan en alguna otra información, como la estructura de la oración o la propia semántica.



Fuente: Google

El enfoque de Google

Como se puede observar en los anteriores gráficos, ni el modelo *BERT* ni *ALBERT* logran una puntuación cero en la métrica *WinoGender*, a pesar de lograr una precisión muy alta en *OntoNotes*, cerca del 100%. Esto se debe a que los modelos utilizan preferentemente correlaciones de género en sus razonamientos, lo que no resulta del todo sorprendente ya que hay una gran variedad de pistas disponibles para comprender el texto y es posible que un modelo general capte cualquiera de ellas, o todas. Sin embargo, no es deseable que un modelo haga sus predicciones basadas principalmente en correlaciones de género aprendidas en lugar de llevarlas a cabo con las evidencias disponibles en los datos de entrada (1).

Igualdad de oportunidades

Google propone una metodología para medir y prevenir la discriminación basada en un conjunto de atributos que pueden generar sesgos. El marco de actuación no solo ayuda a analizar los parámetros de predicción para descubrir posibles cuestiones injustas, sino que también muestra cómo ajustar un parámetro de predicción dado para lograr una mejor compensación entre la precisión de la clasificación y la no discriminación, si es necesario (2).

Los modelos de inserción de texto contienen sesgos: por eso son importantes

En este estudio, Google analiza el efecto del sesgo introducido en modelos de inserción de texto, como por ejemplo el análisis del sentimiento en el uso del procesamiento del lenguaje natural, el análisis de texto y la lingüística computacional para identificar, extraer, cuantificar y estudiar sistemáticamente los estados afectivos y la información subjetiva (3).

Equidad en la clasificación de recomendaciones a través de comparaciones de experimentos revisados y aprobados por la comunidad

En este artículo, Google ofrece un conjunto de novedosas métricas para evaluar las preocupaciones sobre la equidad algorítmica en los sistemas de recomendación. En particular, se muestra cómo la medición de la equidad basada en comparaciones de experimentos aleatorios, revisados y aprobados por la comunidad IA, proporciona un medio apto para razonar sobre la equidad en las clasificaciones de los sistemas de recomendación (4).

(1) https://ai.googleblog.com/2020/10/measuring-gendered-correlations-in-pre.htm

(2) https://ai.googleblog.com/2016/10/equality-of-opportunity-in-machine.html

(3) https://developers.googleblog.com/2018/04/text-embedding-models-contain-bias.html

(4) https://research.google/pubs/pub48107/

Creación de conjuntos de datos y modelos de ML justos

"Todos quieren hacer el trabajo del modelo, no el trabajo de los datos": Cascadas de datos en IA de alto riesgo

Las "cascadas de datos" se originan cuando los modelos han sido entrenados en conjuntos de datos libres de desviaciones (condiciones "ideales") y se implementan en el mundo real, con sus imperfecciones. A pesar de su importancia, los datos son un aspecto subestimado del desarrollo de la IA. En este documento, Google discute cómo el potenciar la excelencia en los datos de entrenamiento se traduce en una IA más segura y sólida (1).

Equidad composicional práctica: comprensión de la equidad en los sistemas de clasificación de componentes múltiples

En este documento, Google se enfoca en el estudio de la composición de la equidad sobre múltiples elementos en sistemas reales, y descubren que esta característica de un sistema de un extremo a otro se puede lograr en gran medida mejorándola en los componentes individuales (2).

(1) https://research.google/pubs/pub49953/

(2) https://arxiv.org/abs/1911.01916

Equidad sin datos demográficos a través del aprendizaje ponderado contradictoriamente

En este estudio se analiza cómo se puede entrenar un modelo de ML para mejorar la equidad cuando ni siquiera se conocen, por cuestiones de privacidad y/o regulación, las características de los propios miembros en los que se van a evaluar los atributos protegidos (característica que divide a una población en grupos cuyos resultados deben tener paridad; incluyen raza, género, nivel social y religión) (1).

No clasificación sin representación

Se evalúan los problemas de geodiversidad en distintos conjuntos de datos disponibles en el mundo desarrollado y se demuestra el sesgo existente de representación en América y Europa centrales en dos grandes conjuntos de datos de imágenes disponibles públicamente (2).

Poner en práctica los principios de equidad: desafíos, métricas y mejoras

Un estudio sobre la aplicación de la equidad en la investigación del ML a un sistema de clasificación de reglas de producción (programa informático utilizado para proporcionar alguna forma de IA, que consiste principalmente en un conjunto de reglas sobre el comportamiento, pero también incluye el mecanismo necesario para seguir esas reglas a medida que el sistema responde a los estados del mundo) y nuevos conocimientos sobre cómo medir y abordar los problemas de equidad de los algoritmos de los sistemas de IA (3).

(1) https://arxiv.org/abs/2006.13114

(2) https://research.google/pubs/pub46553/

(3) https://research.google/pubs/pub47763/

Técnicas de mitigación

Mejora de la detección de sonrisas con diversidad racial y de género

Esta investigación demuestra cómo utilizar el aprendizaje por transferencia (proceso de entrenar un modelo en un conjunto de datos a gran escala y luego usar ese modelo previamente entrenado para llevar a cabo el aprendizaje para otra tarea posterior) en un conjunto limitado de ejemplos de raza y género para mejorar el rendimiento del sistema en general (1).

Satisfacer los objetivos del mundo real con restricciones de conjuntos de dato.

En este documento Google demuestra métodos para manejar múltiples objetivos en múltiples conjuntos de datos mediante el entrenamiento con restricciones de conjuntos de datos (2).

Asignación justa de recursos en un mercado volátil

En este estudio Google presenta algoritmos de aproximación (algoritmo que se encarga de encontrar soluciones aproximadas a problemas de optimización con elevadas garantías demostrables sobre que la solución devuelta se encuentra muy próxima de la solución óptima) para encontrar una solución de compensación de mercado; solución en la cual la oferta de todo lo que se comercialice se ajuste a la demanda y no quede ni oferta ni demanda sobrante, proporcionalmente justa para la asignación de anuncios en línea (3).

Decisiones de datos e implicaciones teóricas cuando se aprenden negativamente representaciones justas

Conecta un enfoque de aprendizaje de múltiples tareas adversas que penaliza el modelo por predecir características sensibles con diferentes métricas de equidad (4).

(1) https://research.google/pubs/pub46513/

(2) https://research.google/pubs/pub46324/

(3) https://research.google/pubs/pub46038/

(4) https://research.google/pubs/pub46295/

El enfoque de Google

Comprensión de los usuarios y la sociedad

Hacia una metodología de carrera crítica en equidad algorítmica

En este trabajo, se examina la forma en que la raza y las categorías raciales se adoptan en los marcos de equidad algorítmica. Se enfocan en la historia de las categorías raciales; históricamente la raza era utilizada para establecer una jerarquía social, y recurren a la teoría crítica de la raza y su trabajo sociológico y de la etnia para fundamentar un entendimiento global conjunto de lo que se entiende por el término "raza" para la investigación de la equidad, aprovechando las lecciones de la salud pública, la investigación biomédica y la investigación de encuestas sociales (1).

Una exploración cualitativa de las percepciones de la equidad algorítmica

Investiga cómo se sienten los miembros de las comunidades potencialmente afectadas por los problemas de equidad en los algoritmos de los modelos (2).

(1) https://arxiv.org/abs/1912.03593

(2) https://research.google/pubs/pub47451/

ML justo en la práctica

En la práctica, hoy en día Google tiene en marcha numerosos proyectos e iniciativas para intentar conseguir sistemas de Machine Learning lo más justos posibles. Algunos ejemplos de ellos se muestran a continuación:

Traducción específica de género en *Google Translate*

Para ayudar a reducir los sesgos de géneros en las traducciones, Google proporciona en su web de traducción la traducción femenina y masculina en algunas palabras de género neutrales (palabras que no tienen género en sí mismas, sino que lo obtienen en función del artículo que llevan delante: el líder, la líder; la taxista, el taxista, etc.) (1).

Garantizar la equidad en el ML para promover la equidad en salud

Este artículo describe cómo el diseño del modelo, los sesgos en los datos y las interacciones de las predicciones del modelo con las de los médicos y los pacientes pueden incrementar las desigualdades en la atención médica. En él se exponen algunas acciones para favorecer la equidad, especialmente aquellas que garantizan la igualdad en los resultados del paciente y la asignación de recursos, y guía a los médicos sobre cuándo deben priorizar cada principio (2).

Aplicación Crowdsource

Proporciona una herramienta para que cualquier persona ayude a entrenar la IA de Google. El objetivo es generar datos de entrenamiento y grandes conjuntos de datos abiertos que representen la diversidad de culturas de todo el mundo (3).

Proyecto Respect

Este proyecto tiene como objetivo crear un conjunto de datos abiertos que recopile declaraciones realizadas por miembros del colectivo LGTBIQ+ (sigla que hace referencia a: lesbiana, gay, trans, bisexual, intersexual y queer) y otras comunidades marginadas, para detectar y abordar los sesgos injustos. De esta manera, los sistemas pueden entrenarse con los propios términos que utilizan estos grupos para expresarse y poder lograr un mejor entendimiento de los datos (4).

(1) https://ai.googleblog.com/2018/12/providing-gender-specific-translations.html

(2) https://www.acpjournals.org/doi/10.7326/M18-1990

(3) https://play.google.com/store/apps/details?id=com.google.android.apps.village.boond

(4) https://www.blog.google/technology/ai/fairness-matters-promoting-pride-and-respect-ai/

Imágenes Inclusivas

Una competición para la comunidad de desarrolladores e investigación global para capacitar modelos y crear sistemas de reconocimiento de imágenes que puedan funcionar igualmente bien en un conjunto de imágenes de prueba extraídas de diferentes distribuciones geográficas con las que no fueron entrenados (1).

Imágenes abiertas extendidas

Un conjunto de datos abierto para ayudar a expandir la diversidad de culturas y personas representadas en los datos de entrenamiento de imágenes (2).

(1) https://www.kaggle.com/c/inclusive-images-challenge

(2) https://research.google/tools/datasets/open-images-extended-crowdsourced/

Recursos y tecnologías para gestionar la equidad

Google proporciona un conjunto de recursos, herramientas y tecnologías para gestionar una IA responsable. Estos elementos son explicados con detalle en el presente documento, en el apartado “Recursos de Google para gestionar los principios”. En el mencionado apartado se describen con detalle dichos elementos para un ciclo de vida que Google estructura en 5 fases y que también se detalla en dicho apartado:

1. Definir el problema
2. Construir y preparar los datos
3. Crear y entrenar un modelo
4. Evaluar el modelo
5. Implementar y supervisar

Para el principio de Equidad, este es un resumen de los **principales** recursos y tecnologías Google que pueden ser utilizados bajo dicho ciclo de vida específicamente para la gestión de este principio. **Tengamos en cuenta que** varios principios **pueden tener relación entre sí** en función del caso de uso. Por ejemplo, la interpretabilidad (que veremos más adelante) y sus herramientas también pueden servir para gestionar la equidad dado que dicho principio ofrece información sobre el modelo que puede servir para identificar puntos clave acerca de la equidad del mismo.

Fase	Recurso
Definir el problema	• *People + AI Research (PAIR) Guidebook* • *PAIR Explorables*
Construir y preparar los datos	• *Know Tour Data* • *Data Cards* • *TF Data Validation*
Crear y entrenar un modelo	• *TF Constrained Optimization* • *TF Model Remediation* • *TF Lattice*
Evaluar el modelo	• *Fairness Indicators*
Implementar y supervisar	• *Model Card Toolkit - MCT*

Principio de Interpretabilidad

En esta sección vamos a desarrollar:

- Qué es la Interpretabilidad para Google.
- Prácticas recomendadas por Google.
- Investigación sobre la interpretabilidad con casos de uso.
- Recursos y tecnologías para gestionar la interpretabilidad.

Qué es la Interpretabilidad para Google

Las predicciones y toma de decisiones automáticas forman parte de la vida cotidiana, desde las recomendaciones de música hasta la monitorización de las constantes vitales de un paciente. Se considera la interpretabilidad como la capacidad de explicar los datos y el modelo a distintos niveles, poder cuestionar, entender y sobre todo confiar en los sistemas de IA.

Al contrario que el software tradicional, entender y *testear* sistemas de IA supone analizar millones de parámetros y operaciones matemáticas. Es por ello por lo que un buen diseño de la solución es clave para la lograr el mayor grado de interpretabilidad posible.

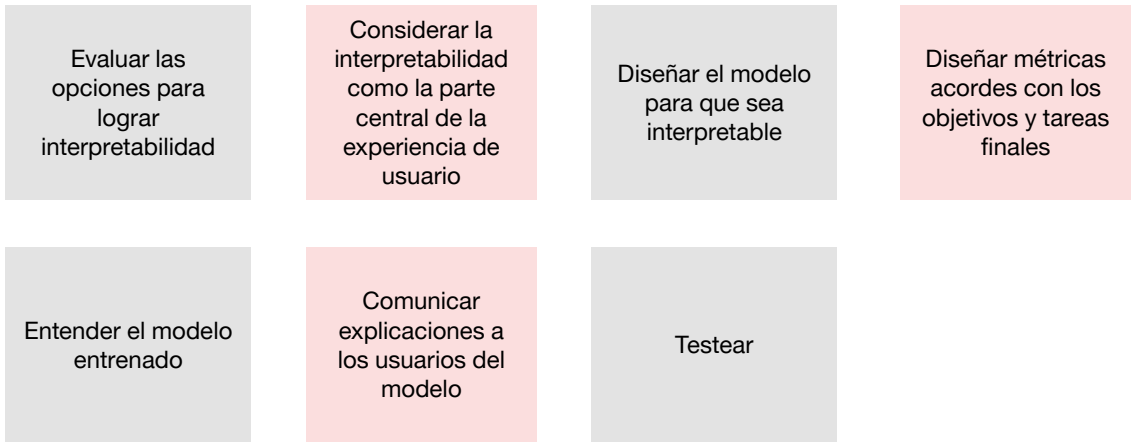
En general, un sistema de IA se comprende mejor a través de los datos y proceso de entrenamiento y del modelo de IA resultante. Es decir, cuánto mayor entendimiento y control se tenga sobre el set de datos de entrada, el proceso de entrenamiento del modelo, y el diseño del modelo en sí, mejor será su interpretabilidad. Si bien es cierto que este escenario plantea nuevos desafíos, se está haciendo un esfuerzo continuo para elaborar guías, establecer mejores prácticas y desarrollar herramientas, mejorando constantemente nuestra capacidad para entender, controlar y *debuggear* (1) sistemas de IA.

Concretamente, desde Google, se está trabajando en proporcionar una serie de recomendaciones, métodos, *frameworks* y herramientas que ayuden en el propósito de potenciar la interpretabilidad de los modelos de IA, a distintos perfiles y fases del ciclo de vida de la solución inteligente.

(1) Acción que permite la ejecución controlada de un programa o código, siguiendo cada instrucción ejecutada para así localizar y corregir errores.

Prácticas recomendadas

Aunque una solución completa a la interpretabilidad no es todavía una realidad y es un área en la que tanto Google como la comunidad de IA siguen investigando, a continuación, se presentan algunas de las prácticas recomendadas por Google hasta la fecha. Dichas recomendaciones se agrupan bajo los siguientes fundamentos:



Evaluar las opciones para lograr interpretabilidad

- Es importante trabajar estrechamente con expertos en el campo en el que vaya a trabajar el modelo (sanidad, *retail*, etc.) para identificar las características claves que se han de entender y el por qué. En algunas pocas ocasiones, hay escenarios o sistemas en los que una interpretabilidad muy granular no es necesaria, al existir suficiente evidencia empírica. Este análisis previo permitirá determinar el nivel de interpretabilidad necesario, ya que demasiada transparencia podría generar entradas a vectores de ataque, como los descritos en el apartado del principio de seguridad descrito en el presente documento, aumentando la vulnerabilidad del sistema.
- Estudiar la viabilidad de analizar los datos de entrada, identificando si hay ciertos campos que son confidenciales, ya que en ocasiones se trabajará con ese tipo de datos a los cuales no se tendrá acceso, o que resultarán difíciles de explicar sin vulnerar políticas de confidencialidad.
- Evaluar si cabe la posibilidad de diseñar un nuevo modelo, o existen restricciones a usar un modelo previamente entrenado. El participar en cada una de las fases de desarrollo de la solución inteligente, desde el diseño, permitirá tener mayor control sobre las prácticas, métodos o herramientas que se emplean para potenciar la interpretabilidad. En caso de modelos ya entrenados, esto se traduce en analizar la viabilidad de acceso al interior del modelo y evaluar si se cumplen los objetivos de interpretabilidad para cada tipo de modelo (modelos de caja negra vs modelos de caja blanca, siendo un modelo de caja negra aquel en el cual solo se conocen las entradas y las salidas).

Considerar la interpretablidad como la parte central de la UX

- Es importante establecer un contacto iterativo con los distintos tipos de usuario (los usuarios que etiquetan los datos, descritos en el apartado de equidad, aquellos que forman parte del testeo y aquellos que utilizarán el modelo), desde la fase de desarrollo hasta la de testeo, redefiniendo las suposiciones sobre las necesidades y objetivos del usuario para que el resultado final sea fruto de un proceso colaborativo y la toma de decisiones contenta el *feedback* proporcionado por los usuarios.
- Diseñar la UX de manera que los usuarios tengan clara la estructura del sistema. Si no se proporciona información clara y concisa, los usuarios tienden a elaborar sus propias teorías sobre cómo funciona el sistema de IA, lo que puede repercutir negativamente en cómo hacen uso del modelo.
- En la medida de lo posible, facilitar el que los usuarios puedan llevar a cabo sus propios *sensitivity analyses*, es decir, empoderarles para que testeen cómo diferentes entradas afectan a la salida del modelo. De esta manera serán capaces de extraer sus propias conclusiones sobre el rendimiento y funcionamiento del modelo, aumentando así la auto-explicabilidad, entendiendo ésta como la capacidad de este de ser entendido de forma sencilla.

El enfoque de Google

Diseñar el modelo para que sea interpretable

- Usar el conjunto de *inputs* más pequeño posible a la hora de establecer los objetivos de rendimiento. De esta forma, resulta más fácil identificar qué factores afectan al modelo y explicar la toma de decisiones. En línea con este concepto, es recomendable usar el modelo más simple que cumpla con dichos objetivos, simplificando así la tarea de interpretabilidad de la solución.
- En la medida de lo posible, usar relaciones causales y no correlaciones (por ejemplo, usar la estatura y no la edad para predecir si es seguro para un niño subirse a una montaña rusa). Es decir, la correlación no implica causalidad, ya que dos variables pueden tener una alta correlación, pero no sean causa la una de la otra.
- Durante la fase de entrenamiento, definir los objetivos de manera que se ajusten a la realidad. Es decir, no solo es importante garantizar que el modelo es preciso, sino que también se debe asegurar que se está entrenando para que los resultados sean satisfactorios en el caso de uso concreto. En ocasiones se tiende a la focalización en la precisión a costa de perder interpretabilidad.
- Restringir el modelo a que produzca relaciones *input-output* que reflejen conocimiento experto del entorno en el que trabaja (por ejemplo, será más probable que se recomiende una tienda de café que esté cercana al usuario, estando el resto de las variables en igualdad de condiciones – precio, tamaño, etc.).

Elegir métricas que reflejen el objetivo y la tarea finales

- Las métricas que se consideren deben abordar los riesgos y beneficios específicos al contexto del caso de uso. Por ejemplo, un sistema de detección de incendios debe tener una sensibilidad baja, incluso si esto implica ocasionales falsas alarmas.

Entender el modelo entrenado

- Se están desarrollando muchas técnicas para obtener mayor grado de conocimiento sobre el modelo. Cuanto más se conozca del modelo, más sencilla resultará la tarea de potenciar la interpretabilidad.
- Analizar la sensibilidad del modelo a diferentes entradas, para diferentes subconjuntos de datos. De esta forma, aumentará el entendimiento sobre qué variables influyen en la toma de decisión del modelo, y cómo lo hacen.

Comunicar explicaciones a los usuarios del modelo

- Proporcionar explicaciones entendibles y adecuadas al público objetivo (los detalles técnicos serán pertinentes al público experto y académicos, mientras para los usuarios de carácter general será de mayor ayuda tener elementos con los que interactuar, visualizaciones, resúmenes, etc). Las explicaciones deben basarse en un conjunto de consideraciones filosóficas, psicológicas, informáticas, legales y éticas sobre qué es considerada una “buena explicación” en diferentes contextos.
- Identificar los escenarios en los que las explicaciones puedan no ser apropiadas o pertinentes (si las explicaciones derivan en mayor confusión en los usuarios, si las explicaciones pueden ser aprovechadas por agentes externos para vulnerar el sistema, o las explicaciones revelan información confidencial).
- Considerar alternativas si las explicaciones requeridas no se pueden compartir con el usuario que las pide, o si no es posible proporcionar una explicación clara y concisa. En estos casos sería mejor “rendir cuentas” a través de otros mecanismos, como la auditoría, o el permitir que los usuarios impugnen o influyan en decisiones futuras mediante el *feedback* que compartan.
- Priorizar explicaciones que describan acciones concretas que el usuario pueda realizar para corregir predicciones imprecisas o erróneas. Además, se debe de procurar no dar a entender que las explicaciones suponen causalidad a menos que así sea, para evitar derivar en conclusiones erróneas.
- Entender todas las partes del sistema de ML/DL (especialmente las entradas) y cómo interactúan y se integran todas las partes ayuda a los usuarios a crear más claramente mapas mentales de los modelos/sistemas. De hecho, estos modelos mentales se ajustan más al rendimiento real del sistema, proporcionando una experiencia más confiable y expectativas más precisas para el aprendizaje futuro.
- Ser conscientes de las limitaciones de las explicaciones. Es decir, explicaciones de partes particulares pueden no ser generalizables, y podrían proporcionar explicaciones contradictorias para dos muestras/ejemplos aparentemente similares.

Probar, probar y probar

Es importante aprender de las *best practices* de la ingeniería de software y la ingeniería de calidad para asegurar que el modelo de IA está rindiendo como se espera y es confiable. Para ello, se deben llevar a cabo diferentes pruebas que lo garanticen. A continuación, se enumeran algunas de las más relevantes:

- Llevar a cabo tests unitarios rigurosos para testear cada componente del sistema aislado.
- Detectar de manera proactiva desviaciones en la entrada, testeando las estadísticas de dichas entradas al sistema de IA, para asegurarse de que no están cambiando de forma inesperada.
- Usar lo que se denomina un *gold standard datset (1)* para testear el sistema y asegurar que continúa comportándose como se espera. Actualizar este conjunto de prueba regularmente, conforme existan cambios en los usuarios o casos de uso.
- Llevar a cabo *testing* con usuarios de forma iterativa para incorporar un conjunto diverso de necesidades de usuario en los ciclos de desarrollo, incluyendo sus comentarios y propuestas
- Aplicar el principio de ingeniería de calidad *poka-yoke*, descrito en el apartado 'Probar, probar y probar' de las buenas prácticas generales de Google.
- Realizar pruebas/tests de integración: entender cómo interactúa el modelo de IA con otros sistemas, y qué bucles de retroalimentación se pueden crear (si los hay). Por ejemplo, recomendar una noticia porque es popular puede hacer que esa noticia sea más popular y por tanto hacer que se recomiende más.

(1) Set de datos aceptado como el más preciso y confiable de tu tipo, que puede usarse como medida de dichas cualidades en otros conjuntos de datos.

El enfoque de Google

Investigación sobre la interpretabilidad con casos de uso

Google está trabajando intensamente en la interpretabilidad en los sistemas de IA. A continuación, se presentan algunas de las principales líneas de investigación existentes acerca del diseño y el entendimiento del modelo para poder gestionar la interpretabilidad del mismo.

- *Constrained training* (1): en este paper se examina el uso de limitaciones del clasificador durante el entrenamiento para asegurar que se depuran las responsabilidades, incluyendo el cumplir con métricas de equidad en los grupos y garantizar la toma de decisiones correcta del clasificador en ejemplos representativos.
- *Towards a rigorous science of interpretable machine learning* (2). Incluye sugerencias sobre cuándo se necesita interpretabilidad y cuándo no, y una taxonomía para evaluar de forma rigurosa de un modelo de ML interpretable.
- *Human-in-the-Loop Interpretability Prior* (3). Ayuda a aprender sobre modelos interpretables preguntando a las personas qué modelos son más fácilmente interpretables durante el entrenamiento del modelo.
- *Concept-based model explanations for Electronic Health Records* (4). Estudio que evalúa cómo se puede utilizar el TCAV para generar explicaciones basadas en conceptos sobre el análisis de registros clínicos electrónicos mediante RNNs (redes neuronales recurrentes).
- *On Completeness-Aware concept-based explanations in Deep Neural Networks* (5). Se explora cómo las DNNs pueden ofrecer explicaciones sobre decisiones de alto nivel similares a las que daría un humano, en cuanto a los conceptos clave en los que está basada la predicción.
- *Integrated gradients* (6) *and SmoothGrad* (7). Estudios que definen cómo identificar características de la entrada (ejemplo, pixels) que son relevantes para las predicciones.
- *Deepdream y Building blocks of interpretability*. Exploran qué entradas activan diferentes partes de las redes neuronales, y cómo se pueden combinar la visualización de características, asignación y la factorización matricial.
  - Enlace al artículo sobre *Deepdream* (8).
  - Enlace al artículo *Building blocks of interpretability* (9).
- XRAI: Método que ayuda a desarrolladores a entender qué regiones de una imagen han sido más relevantes para el modelo a la hora de realizar la predicción.
  - Enlace a la publicación (10).
- *Closing the Accountability Gap: Defining a Framework for Internal Algorithmic Auditing* (11). *Framework* para auditoría de algoritmos que respalda el desarrollo *end-to-end* de sistemas de IA, aplicable a lo largo del ciclo de vida del desarrollo de la organización interna.

(1) Goh, G., Cotter, A., Gupta, M., & Friedlander, M. (June 2016) *Satisfying Real-world Goals with Dataset Constraints.*

(2) Doshi-Velez, F., & Kim B. (February 2017). *Towards A Rigorous Science of Interpretable Machine Learning.*

(3) Lage, I., Ross, A.S., Kim B., Gershman, S.J. & Doshi-Velez, F. (May 2018). H*uman-in-the-Loop Interpretability Prior.*

(4) Mincu, D., Loreaux, E., Hou, S., Baur, S., Protsyuk, M.G.S., Mottram, A., Tomasev, N., Karthikesanlingam, A., & Schrouff, J. (April 2021). *Concept-based model explanations for Electronic Health Records. Association for Computing Machinery*.

(5) Yeh, C., Kim, B., Arik, S.O., Li, C., Pfister, T., & Ravikumar, P. (October 2019). *On Completeness-aware Concept-Based Explanations in Deep Neural Networks*.

(6) Sundararajan M., Taly, A., & Yan, Q. (March, 2017). *Axiomatic Attribution for Deep Networks*.

(7) Smilkov, D., Thorat, N., Kim, B., Viégas, F., & Wattenberg M. (June 2017). *SmoothGrad: removing noise by adding noise*.

(8) https://ai.googleblog.com/2015/06/inceptionism-going-deeper-into-neural.html

(9) https://distill.pub/2018/building-blocks/

(10) https://cloud.google.com/blog/products/ai-machine-learning/explaining-model-predictions-on-image-data

(11) Raji, I.D., Smart, A., White, R.N., Mitchel, M., Gebru, T., Hutchinson, B., Smith-Loud, J., Theron, D., & Barnes P. (January, 2020). *Closing the AI Accountability Gap: Defining an End-to-End Framework for Internal Algorithmic Auditing. FAT Conference 2020*.

- *Estimating training* data *influence by tracking Gradient Descent* (1). Método llamado *TracIn* que mide la influencia de un ejemplo de entrenamiento en una predicción realizada por el modelo.
- *To trust or not to trust a classifier*. Método que determina si se puede confiar en la predicción de un clasificador para un uso más responsable de los modelos de ML (2).
- *Sanity Checks for Saliency Maps. Saliency maps* es un método de interpretabilidad post-entrenamiento para explicar la "evidencia" de las predicciones. Sin embargo, este trabajo expone que esto tiene poco que ver con las predicciones del modelo. Algunos *Saliency maps* son visualmente indistinguibles antes y después de aleatorizar los pesos de la red (es decir, produciendo predicciones poco fiables) (3).
- *Critiquing protein family classification models using sufficient input subsets. Framework* explicativo local que interpreta funciones de caja negra para encontrar un subconjunto de entrada suficiente que sea un set mínimo de características de entrada cuyos valores observados por sí solos sean suficientes para alcanzar la misma decisión (4).

(1) Pruthi., G., Liu, F., Sundararajan, M., & Kale, S. (February, 2020). *Estimating Training Data Influence by Tracking Gradient Descent. 34th Conference on Neural Information Processing Systems*.

(2) Jiang, H., Kim, B., Guan, M., & Gupta, M. (October, 2018). *To Trust Or Not To Trust a Classifier. 32nd Conference on Neural Information Processing Systems*.

(3) Adebayo, J., Gilmer, J., Muelly, M.C., Goodfellow, I., Hardt, M., & Kim, B. (October, 2018). *Sanity Checks for Saliency Maps. 32nd Conference on Neural Information Processing Systems*.

(4) https://github.com/google-research/google-research/tree/master/sufficient_input_subsets

Recursos y tecnologías para gestionar la interpretabilidad

Google proporciona un conjunto de recursos, tecnologías y herramientas para gestionar una IA responsable. Estos elementos son explicados con detalle en el presente documento, en el apartado "Recursos de Google para gestionar los principios". En el mencionado apartado se describen con detalle dichos elementos para un ciclo de vida que Google estructura en 5 fases y que también se detalla en dicho apartado:

1. Definir el problema
2. Construir y preparar los datos
3. Crear y entrenar un modelo
4. Evaluar el modelo
5. Implementar y supervisar

Para el principio de Interpretabilidad, este es un resumen de los principales recursos y tecnologías Google que pueden ser utilizados bajo dicho ciclo de vida específicamente para la gestión de este principio de interpretabilidad.

Tengamos en cuenta que, para una gestión responsable de estos principios éticos, una buena práctica es hacerlo desde un punto de vista holístico, teniendo en cuenta la relación entre todos ellos aplicada a las particularidades de cada caso de uso.

Fase	Recurso
Definir el problema	• *People + AI Research (PAIR) Guidebook* • *PAIR Explorables*
Construir y preparar los datos	• *Know Tour Data* • *TF Data Validation*
Crear y entrenar un modelo	• *TF Lattice*
Evaluar el modelo	• *Explanaible AI* • *Language Interpretability Tool* • *TensorBoard* • *TF Model Analysis*
Implementar y supervisar	• *Model Card Toolkit - MCT* • *ML Metadata* • *Model Cards*

El enfoque de Google

Principio de Privacidad

En esta sección vamos a desarrollar:

- Qué es la privacidad para Google.
- Cuáles son las prácticas recomendadas por Google.
  - Recopilar y manejar datos de una manera responsable.
  - Aprovechar el procesamiento en dispositivo local cuando sea posible.
  - Proteger adecuadamente la privacidad de los modelos de ML.
- Investigación sobre la privacidad con casos de uso.
- Recursos y tecnologías para gestionar la privacidad.

Qué es la Privacidad para Google

Los modelos de *Machine Learning (ML)* aprenden de los datos de entrenamiento y hacen predicciones a partir de los datos de entrada. A veces, los datos de entrenamiento, los datos de entrada, o ambos pueden ser de carácter confidencial o privado. Un ejemplo de este tipo de datos podrían ser los datos que se obtienen a partir de informes médicos como enfermedades, radiografías, etc.

A pesar de que pueden existir grandes beneficios al construir modelos que operen con datos confidenciales (por ejemplo, un detector de cáncer entrenado en un *dataset* de imágenes de biopsia y utilizado en escaneos de pacientes), es esencial tener en cuenta las posibles implicaciones en la privacidad del uso de datos confidenciales.

Esto incluye no solo respetar los requisitos legales y normativos, sino también considerar las normas sociales y las expectativas individuales típicas. Las expectativas subjetivas de cada persona sobre el nivel de privacidad de sus datos. Por ejemplo, ¿qué tipo de protección debe implementarse para garantizar la privacidad de las personas considerando que los modelos de ML pueden recordar o revelar aspectos de los datos a los que has estado expuestos? ¿Qué medidas son necesarias para garantizar a los usuarios una transparencia y un control adecuados de sus datos?

Afortunadamente, la posibilidad de que los modelos de ML revelen datos se puede minimizar aplicando adecuadamente varias técnicas de una manera precisa. Google está desarrollando constantemente este tipo de técnicas para proteger la privacidad en los sistemas de inteligencia artificial (IA). Esta es un área de investigación en curso en la comunidad de ML con un gran margen de crecimiento (1), (2).

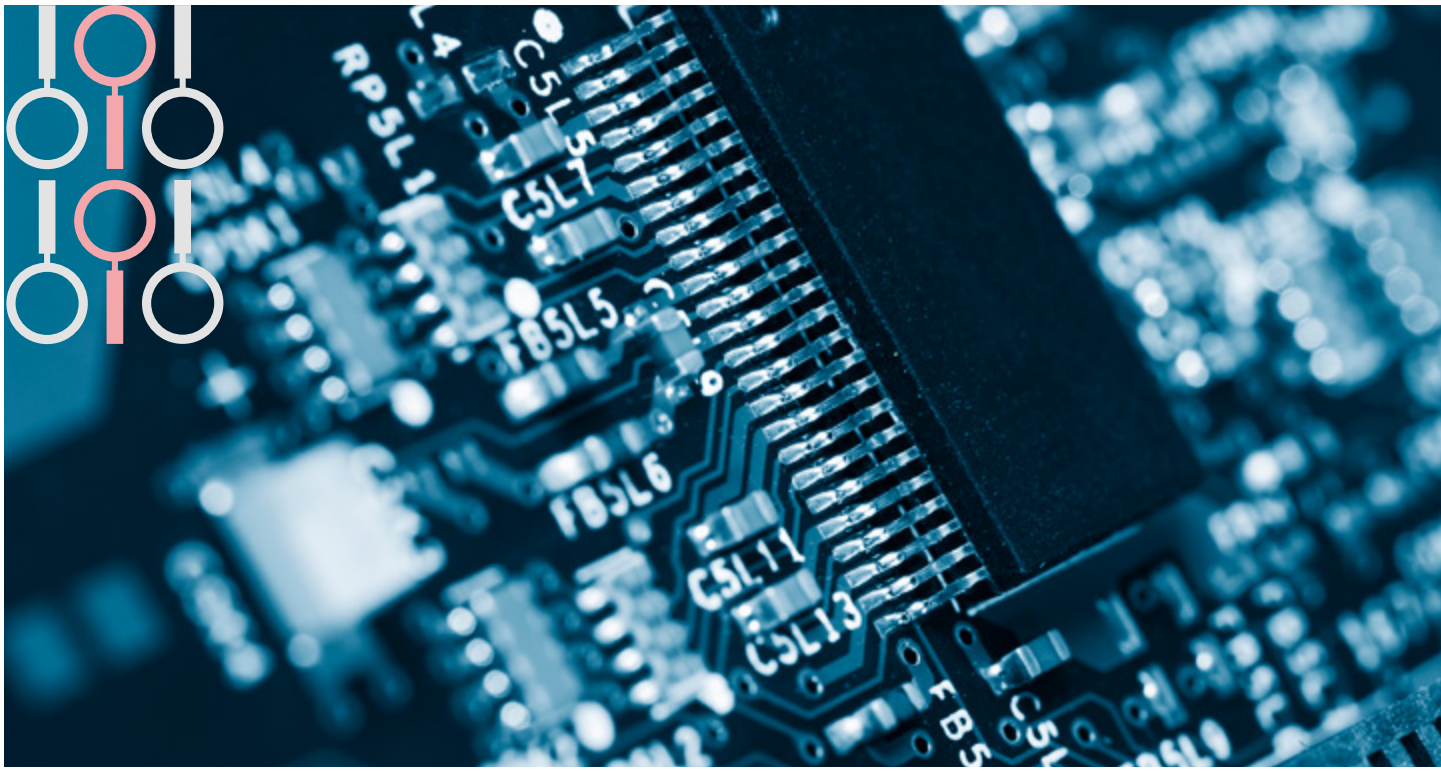
(1) Gibert, D., Mateu, C., & Planes, J. (2020). *The rise of machine learning for detection and classification of malware: Research developments, trends and challenges. Journal of Network and Computer Applications*, 153, 102526.
(2) Ahmad, I., Shahabuddin, S., Malik, H., Harjula, E., Leppänen, T., Loven, L., ... & Riekki, J. (2020). *Machine learning meets communication networks: Current trends and future challenges. IEEE Access*, 8, 223418-223460.

Prácticas recomendadas

Al igual que no existe un modelo "correcto" único para todas las tareas de ML, tampoco hay un único enfoque correcto para la protección de la privacidad en todos los escenarios. En la práctica, los investigadores y desarrolladores deben iterar para encontrar un enfoque que equilibre adecuadamente la privacidad y la utilidad para la tarea en cuestión y, para que este proceso tenga éxito, se necesita una definición clara de privacidad. Para conseguir esto último, Google propone, por ejemplo, usar “PATE” (*Private Aggregation of Teacher Ensembles*), que consiste en transferir a un modelo ‘estudiante’ el conocimiento de un conjunto de modelos ‘maestros’ con una privacidad inmediata de los datos confidenciales, proporcionada por el entrenamiento de los maestros con datos disjuntos, es decir, que provienen de distintas entradas; y una fuerte privacidad garantizada por la agregación de ruido en los datos que los modelos maestros proporcionan a los modelos ‘alumnos’.

Recopilar y manejar datos de una manera responsable

- Identificar si el modelo de ML puede entrenarse sin el uso de datos confidenciales, por ejemplo, utilizando un *dataset* de datos no confidenciales, o una fuente de datos pública existente. No obstante, como se describe en el principio de Equidad, una fuente de datos pública deberá ser analizada en mayor detalle para asegurar que es suficientemente representativa, habiendo que aumentarla si requiere para ajustarla mejor a la realidad.
- Si es imprescindible procesar datos de entrenamiento confidenciales, se deberá realizar un esfuerzo por minimizar el uso de dichos datos. Es necesario manejar los datos confidenciales con precaución; por ejemplo, cumpliendo con las leyes y normas requeridas, proporcionando a los usuarios un aviso claro y brindándoles los controles necesarios sobre el uso de datos, siguiendo las mejores prácticas como el cifrado en tránsito, que consiste en el cifrado de datos que van a ser transferidos, y cifrado en reposo, que corresponde al cifrado de los datos que van a ser almacenados.
- Se deberá anonimizar y agregar los datos de entrada utilizando las mejores prácticas de depuración de datos considerando, por ejemplo, la eliminación de la información de identificación personal, PII; algunos ejemplos de este tipo de información sería el nombre completo, el número de seguro social, el número de licencia de conducir, el número de cuenta bancaria, etc. y la eliminación de los valores atípicos o de metadatos que podrían permitir la desanonimización, que se define como el proceso mediante el cual los usuarios maliciosos son capaces de acceder a los datos confidenciales que habían sido anonimizados previamente. Este análisis incluye los metadatos implícitos, como el orden de llegada, que se pueden eliminar mediante mezcla aleatoria, como en el caso de *Prochlo*, explicado en la sección *Prochlo*; o la API de *Cloud Data Loss Prevention* para descubrir y redactar automáticamente datos confidenciales e identificativos). Al evitar la desanonimización, se consigue eliminar la posibilidad de acceder a la información de identificación personal, previamente definida.



Aprovechar el procesamiento en dispositivo local cuando sea posible

- Si el objetivo es conocer las estadísticas de las interacciones individuales, por ejemplo, la frecuencia con la que se utilizan ciertos elementos de la interfaz de usuario, se deberá recopilar solo las estadísticas que se hayan calculado localmente, en el dispositivo, en lugar de los datos de interacción en bruto, que pueden incluir información sensible.
- Se deberá considerar la inclusión de técnicas como el aprendizaje federado, detallado posteriormente en este documento; en el cual una flota de dispositivos se coordina para entrenar un modelo global compartido a partir de datos de entrenamiento almacenados localmente y, de esta manera, al descentralizar los datos los haces más inaccesibles, mejorando la privacidad en el sistema.
- Cuando sea posible, se deben aplicar diversas operaciones como: agregación, que es la suma de datos de distintas fuentes u orígenes; aleatorización, que consiste en mezclar el conjunto de datos con el que se va a trabajar; y depuración, que se refiere a la eliminación de datos irrelevantes para nuestro modelo, en el dispositivo. Algunos ejemplos serían la agregación segura (RAPPOR) y el paso de codificación de *Prochlo*. Hay que tener en cuenta que estas operaciones solo pueden proporcionar un nivel de privacidad limitado, a menos que las operaciones empleadas estén acompañadas de pruebas que demuestren su fiabilidad.

Proteger adecuadamente la privacidad de los modelos de ML

Debido a que los modelos de ML pueden exponer detalles sobre sus datos de entrenamiento tanto a través de sus parámetros internos como de su comportamiento visible desde el exterior, es crucial considerar el impacto sobre la privacidad de cómo se construyeron los modelos y cómo se puede acceder a ellos. Es necesario:

- Estimar si su modelo está memorizando o exponiendo involuntariamente datos confidenciales utilizando pruebas basadas en mediciones de “exposición”, que evalúan cuantitativamente el riesgo de que las secuencias de datos sean memorizadas, o en la evaluación de la inferencia de los miembros (*membership inference*), que se define como un tipo de ataque que permite detectar los datos utilizados para entrenar un modelo de ML.
- Experimentar con los parámetros para la minimización de datos (por ejemplo, agregación, umbrales de valores atípicos y factores de aleatorización) para comprender las compensaciones e identificar la configuración óptima para su modelo. Con la minimización de los datos se consigue reducir el volumen de datos confidenciales que el modelo utiliza para su correcto funcionamiento.
- Entrenar modelos de ML utilizando técnicas que establezcan garantías matemáticas para la privacidad, ofreciendo de esta manera garantías solidas de que el modelo que se está evaluando no aprende ni recuerda los detalles sobre ningún usuario especifico, un ejemplo de estas técnicas podría ser la librería *TensorFlow Privacy*, descrito en la fase *Define problem* del ciclo de vida del presente documento. Se deberá tener en cuenta que estas garantías analíticas no son garantías sobre el sistema operativo completo.
- Seguir los procesos de mejores prácticas establecidos para el software criptográfico y crítico para la seguridad, por ejemplo, el uso de enfoques basados en principios y siendo estos demostrables, la publicación de nuevas ideas revisadas y aprobadas por la comunidad, el código abierto de componentes de software críticos y la contratación de expertos para su revisión en todas las etapas de diseño y desarrollo. Además, los modelos deben protegerse de los ordenadores cuánticos, los cuales aprovechan algunos de los fenómenos de la mecánica cuántica para ofrecer grandes avances en cuanto a potencia de procesamiento y se prevé que tendrán un gran crecimiento en la próxima década; para ello se debería progresar en el ámbito de la criptografía cuántica.

Investigación sobre la privacidad con casos de uso

Google lleva mucho tiempo apoyando los esfuerzos de investigación y desarrollo de técnicas de mejora de la privacidad y anonimato para los sistemas de IA, incluyendo becas de investigación para profesores y becas de doctorado para la investigación académica en privacidad y seguridad. Además, Google ha trabajado para publicar nuevas técnicas y códigos de fuente abierta como mejores prácticas de protección de la privacidad. El propio trabajo de Google en esta área ha tenido un impacto significativo, como se detalla a continuación.

- **Aprendizaje federado:** El aprendizaje federado permite a los dispositivos móviles aprender de manera colaborativa un modelo de predicción compartido, manteniendo todos los datos de entrenamiento en el dispositivo y desvinculando la capacidad de hacer ML de la necesidad de almacenar los datos de la nube. Funciona de la siguiente manera: un dispositivo móvil descarga el modelo actual, lo mejora aprendiendo de los datos del teléfono y sube el modelo actualizado (mejorado) a la nube para que otros usuarios puedan seguir mejorándolo.
  - Enlace al ejemplo de trabajo (1)

(1) https://ai.googleblog.com/2017/04/federated-learning-collaborative.html

- **Reconstrucción federada (aprendizaje federado parcialmente local)**: Se diferencia del aprendizaje federado en que, en este caso, el entrenamiento se hace de una manera parcialmente local, en vez de que este sea global. Mediante esta modificación se puede mejorar el rendimiento del modelo tal y como se indica en el siguiente *paper* (1).
- ***Prochlo***: Consiste en un proyecto de monitorización de datos cuya arquitectura se basa en ESA (*Encode, Shuffle, Analyze*), la cual proporciona una mayor utilidad de los datos y su anonimato. El funcionamiento se divide en tres pasos principales: el primero es la codificación de los datos; a continuación, se procede a mezclar estos datos codificador; por último, los datos codificados y barajados se analizan mediante un motor de análisis específico para mejorar aún más la seguridad (2).
- ***RAPPOR (Randomized Aggregatable Privacy-Preserving Ordinal Response)***: Es un proyecto desarrollado por Google para proteger la seguridad y la privacidad de sus usuarios. Se basa en el concepto de respuesta aleatoria y permite aprender estadísticas sobre el comportamiento del software de los usuarios garantizando la privacidad del cliente (3).

(1) https://research.google/pubs/pub50086/
(2) https://research.google/pubs/pub46411/
(3) https://ai.googleblog.com/2014/10/learning-statistics-with-privacy-aided.html

Recursos y tecnologías para gestionar la Privacidad

Google proporciona un conjunto de recursos y herramientas para gestionar una IA responsable. Estos elementos son explicados con detalle en el presente documento, en el apartado “Recursos de Google para gestionar los principios”. En el mencionado apartado se describen con detalle dichos elementos para un ciclo de vida que Google estructura en 5 fases y que también se detalla en dicho apartado:

1. Definir el problema
2. Construir y preparar los datos
3. Crear y entrenar un modelo
4. Evaluar el modelo
5. Implementar y supervisar

Para el principio de Privacidad, este es un resumen de los **principales** recursos y tecnologías Google que pueden ser utilizados bajo dicho ciclo de vida **específicamente para la gestión de este principio de Privacidad**.

Tengamos en cuenta que, para una gestión responsable de estos principios éticos, una buena práctica es hacerlo desde **un punto de vista holístico**, teniendo en cuenta la relación entre todos ellos aplicada a las particularidades de cada caso de uso. Por ejemplo, en el caso del principio de Privacidad, su relación con el principio de Seguridad es muy importante.

Fase	Recurso
Definir el problema	• *PAIR Explorables* • *People + AI Guidebook*
Construir y preparar los datos	
Crear y entrenar un modelo	• *TF Federated*
Evaluar el modelo	• *TF Privacy* • *TF Privacy Test*
Implementar y supervisar	

El enfoque de Google

Principio de Seguridad

En esta sección vamos a desarrollar:

- Qué es la seguridad para Google.
- Prácticas recomendadas por Google.
  - Identificar las posibles amenazas al sistema.
  - Desarrollar un sistema para combatir las amenazas.
  - Seguir aprendiendo a mantenerse a la vanguardia.
- Investigación sobre la Seguridad con casos de uso.
- Recursos y tecnologías para gestionar la Seguridad.

Qué es la Seguridad para Google

La seguridad implica garantizar que los sistemas de inteligencia artificial (IA) se comporten según lo previsto, independientemente de cómo los atacantes intenten interferir. Es esencial considerar y abordar la seguridad de un sistema de IA antes de que se confíe plenamente en él, sobre todo en aplicaciones donde un nivel de seguridad alto es crítico.

La seguridad de los sistemas de IA puede generar muchos problemas. Por ejemplo, es difícil predecir todos los escenarios con antelación, especialmente cuando se aplica el *Machine Learning (ML)* o el *Deep Learning (DL)* a problemas que son difíciles de resolver para los humanos. También es difícil construir sistemas que proporcionen tanto las restricciones necesarias para la seguridad como la flexibilidad necesaria para generar soluciones creativas o adaptarse a datos de entrada inusuales. A medida que se desarrolle la tecnología de IA, los atacantes seguramente encontrarán nuevos medios de ataque; y será necesario desarrollar nuevas soluciones en paralelo. A continuación, se muestran las recomendaciones actuales de Google.

Prácticas recomendadas

La investigación sobre la seguridad en el campo de ML abarca una amplia gama de amenazas, como, por ejemplo:

- El envenenamiento de datos de entrenamiento, que consiste en la modificación de este tipo de datos de forma maliciosa para engañar y/o provocar un malfuncionamiento del modelo.
- La sustracción de datos confidenciales con los que se ha entrenado al modelo.
- El robo de modelos y ejemplos adversos, que son aquellos que se generan a partir del uso de modelos de ML basados en el aprendizaje adversario o *adversarial learning*. El *adversarial learning* consiste en el uso de una red neuronal para generar ejemplos adversos que pueden engañar a un sistema, junto con una segunda red para tratar de detectar el fraude.

Google invierte en investigación relacionada con todas estas áreas, y parte de este trabajo está relacionado con prácticas en inteligencia artificial y privacidad. Uno de los focos de la investigación sobre seguridad en Google ha sido el aprendizaje adversario previamente definido.

Actualmente, las mejores defensas contra ejemplos adversos aún no son lo suficientemente fiables para su uso en un entorno de producción. Se trata de un área de investigación en curso y extremadamente activa. Dado que aún no existe una defensa eficaz, los desarrolladores deben pensar si su sistema es susceptible de ser atacado, considerando las posibles consecuencias de un ataque exitoso y, en la mayoría de los casos, simplemente no construir sistemas donde dichos ataques puedan tener un impacto negativo significativo.

Identificar las posibles amenazas al sistema

- Se deberá considerar si alguien pudiera tener algún incentivo para hacer que el sistema se comporte mal. Por ejemplo, si un desarrollador crea una aplicación que ayuda a un usuario a organizar sus propias fotos, sería fácil para los usuarios maliciosos modificar las fotos para organizarlas incorrectamente, pero los usuarios maliciosos tendrían un incentivo limitado para hacerlo, ya que tan solo podrían modificar el orden de las fotos, pero no podrían realizar modificaciones de mayor gravedad como por ejemplo la modificación o eliminación de fotos.
- Identificar las consecuencias que resultarían si el sistema cometiera un error; y evaluar la probabilidad y la gravedad de estas consecuencias.
- Desarrollar un modelo de amenaza riguroso para comprender todos los posibles vectores de ataque. Por ejemplo, un sistema que permitiera a un atacante cambiar los datos de entrada al modelo ML podría ser mucho más vulnerable que un sistema que procese metadatos recopilados por el servidor, como marcas de tiempo de las acciones que realizó el usuario, ya que es mucho más difícil que un usuario malicioso modifique los datos de entrada si no se requiere su participación directa.

Desarrollar un sistema para combatir las amenazas

Algunas aplicaciones como por ejemplo el filtrado de *spam*, pueden tener éxito con las técnicas de defensa actuales a pesar de la dificultad del ML/DL adverso, que se refiere a las técnicas de ML y DL que utilizan el aprendizaje adversario.

- Se deberá probar el rendimiento de sus sistemas en el entorno adverso. En algunos casos, esto se puede hacer utilizando herramientas como *CleverHans*.
- Crear un equipo rojo interno, es decir, un grupo de trabajo que desempeñe el papel de enemigo, para llevar a cabo diferentes pruebas, u organizar un concurso o programa de recompensas que anime a terceros a probar la defensa de su sistema utilizando algoritmos o modelos basados en el aprendizaje adversario.

Seguir aprendiendo a mantenerse a la vanguardia

- Todo aquel involucrado se deberá mantener actualizado sobre los últimos avances en investigación. La investigación sobre el ML/DL adverso está ofreciendo un mejor rendimiento para las defensas, como se detalla posteriormente mediante 'CAT-Gen', y algunas técnicas de defensa están comenzando a ofrecer garantías demostrables (1).
- Más allá de los ataques con los datos de entrada, es posible que haya otras vulnerabilidades en la cadena de suministro de ML. Aunque tal ataque aún no ha ocurrido, es importante considerar la posibilidad y estar preparado.

(1) Wong, E., & Kolter, Z. (2018, July). *Provable defenses against adversarial examples via the convex outer adversarial polytope*. *In International Conference on Machine Learning* (pp. 5286-5295). PMLR.

Ejemplos de trabajo en el área de Seguridad de Google

Google lleva a cabo investigaciones en materia de seguridad y protección de la IA y comparte este trabajo a través de publicaciones, talleres y códigos de fuente abierta. Google también apoya a investigadores externos en esta área a través de becas de investigación para profesores y becas de doctorado. Algunos ejemplos de este trabajo se encuentran a continuación.

- ***CleverHans***: Es una librería de *Python* para evaluar la vulnerabilidad de los sistemas de ML/DL ante ejemplos adversos. La biblioteca *CleverHans* está en continuo desarrollo, y abierta para que todos los usuarios aporten su ayuda a la hora de resolver las diferentes tareas que los propios desarrolladores proponen (1).
- **CAT-Gen: *Improving Robustness in NLP Models via Controlled Adversarial Text Generation***: NLP, Natural ***Language Processing***, es un área dentro de ML/DL que se centra en el estudio de las interacciones mediante el uso del lenguaje natural entre los seres humanos y las máquinas. Al desarrollar modelos de NLP podremos realizar diferentes tareas de manera automática como traducir textos, resumirlos, etc. El problema de este tipo de modelos es que son muy poco robustos frente a pequeñas modificaciones en los datos de entrada. Sin embargo, en este *paper* se demuestra como al introducir el aprendizaje adversario, *adversarial learning*, en este tipo de modelos, la robustez frente a cambios en los datos de entrada mejora notablemente (2).
- ***Machine learning security contest***: Es un concurso creado en 2017 y organizado en *Kaggle*, una plataforma de competiciones de *data science*; que tiene como objetivo analizar cómo se pueden utilizar diferentes técnicas de ML y DL para mejorar los sistemas de defensa frente a posibles futuros ataques (3).
- ***On Evaluating Adversarial Robustness***: La tarea de evaluar las defensas frente a ejemplos adversos es extremadamente complicada. No obstante, en este *paper* se discute sobre nuevos métodos que pueden resultar eficaces a la hora de proteger nuestro sistema frente a diferentes ataques que utilizan modelos basados en el aprendizaje adversario.

(1) https://github.com/cleverhans-lab/cleverhans
(2) https://arxiv.org/abs/2010.02338
(3) https://www.technologyreview.com/2017/07/20/68124/ai-fight-club-could-help-save-us-from-a-future-of-super-smart-cyberattacks/

Recursos y tecnologías para gestionar la Seguridad

Google proporciona un conjunto de recursos y herramientas para gestionar una IA responsable. Estos elementos son explicados con detalle en el presente documento, en el apartado “Recursos de Google para gestionar los principios”. En el mencionado apartado se describen con detalle dichos elementos para un ciclo de vida que Google estructura en 5 fases y que también se detalla en dicho apartado:

1. Definir el problema
2. Construir y preparar los datos
3. Crear y entrenar un modelo
4. Evaluar el modelo
5. Implementar y supervisar

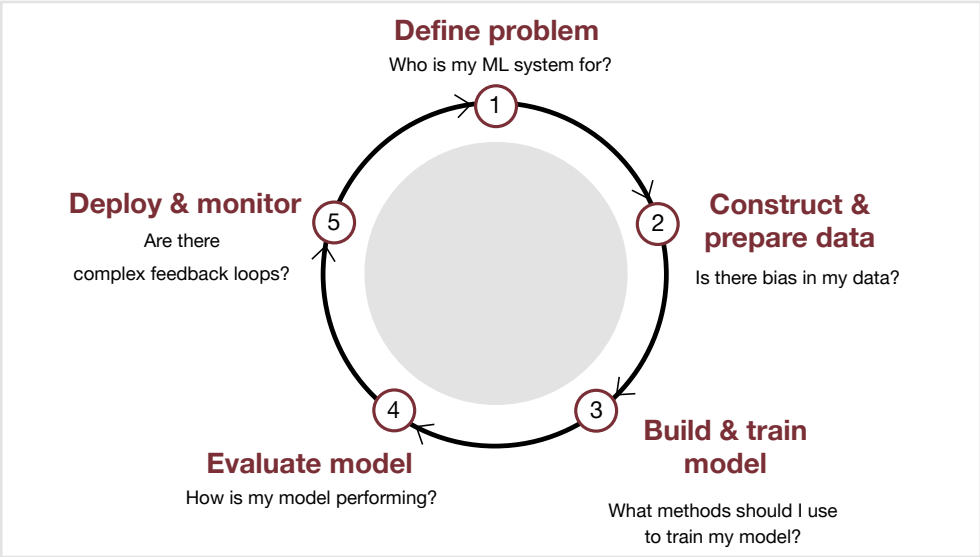
Para el principio de Seguridad, este es un resumen de los **principales** recursos y tecnologías Google que pueden ser utilizados bajo dicho ciclo de vida **específicamente** para la gestión de este principio de Seguridad.

Tengamos en cuenta que, para una gestión responsable de estos principios éticos, una buena práctica es hacerlo desde **un punto de vista holístico**, teniendo en cuenta la relación entre todos ellos aplicada a las particularidades de cada caso de uso. En el caso del principio de Seguridad, su relación con el principio de Privacidad es muy importante.

Fase	Recurso
Definir el problema	• *PAIR Explorables* • *People + AI Guidebook*
Construir y preparar los datos	• *TF Data Validation*
Crear y entrenar un modelo	• *TF Federated*
Evaluar el modelo	• *TF Privacy Test* • *What-if-Tool*
Implementar y supervisar	

Los recursos de Google para gestionar los principios

Google enfoca la gestión operativa de estos principios éticos basándose en un ciclo de vida de las soluciones inteligentes formado por 5 pasos. En cada uno de los pasos del flujo de trabajo hay que incorporar las prácticas de *Responsible AI*. A continuación, se describe dicho flujo, incluyendo algunas preguntas clave que deben tenerse en cuenta en cada etapa.



Fuente: Google

① Definir el problema

¿A quién está dirigido mi sistema de aprendizaje automático?

La forma en que los usuarios reales experimentan el sistema es fundamental para evaluar el verdadero impacto de sus predicciones, recomendaciones y decisiones. Asegúrate de que un conjunto de usuarios diverso proporcione sus opiniones al comienzo del proceso de desarrollo.

¿Cuáles son los recursos disponibles?

- *People + AI Research (PAIR) Guidebook*
- *PAIR Explorables*

② Construir y preparar los datos

¿Estoy usando un conjunto de datos representativo?

¿Tu muestreo de datos representa a los usuarios? (p. ej. se utilizará para todas las edades, pero solo tienes datos de entrenamiento correspondientes a adultos mayores). ¿El muestreo representa escenarios de la vida real? (p. ej. se utilizará para todo el año, pero solo hay datos de entrenamiento correspondientes al verano).

¿Hay algún grado de sesgo humano o del mundo real en mis datos?

Es posible que los sesgos subyacentes en los datos contribuyan a ciclos de reacción complejos que acentúen los estereotipos existentes.

¿Cuáles son los recursos disponibles?

- *Know your data (Beta)*
- *TF Data validation*
- *Data Cards*

③ Crear y entrenar un modelo

¿Qué métodos debería utilizar para entrenar mi modelo?

Usa métodos de entrenamiento que aporten equidad, interpretabilidad, privacidad y seguridad al modelo.

¿Cuáles son los recursos disponibles?

- *TF Model Remediation*
- *TF Privacy*
- *TF Federated*
- *TF Constrained Optimization (TFCO)*
- *TF Lattice (TFL)*

④ Evaluar el modelo

¿Cuál es el rendimiento de mi modelo?

Evalúa la experiencia del usuario en escenarios de la vida real teniendo en cuenta una amplia variedad de usuarios, casos de uso y contextos de uso. Primero ensaya y realiza iteraciones mediante pruebas internas; luego haz pruebas de forma constante después del lanzamiento.

¿Cuáles son los recursos disponibles?

- *Fairness Indicators*
- *TF Model Analysis*
- *What-If Tool*
- *Language Interpretability Tool*
- *Explainable AI*
- *TF Privacy Test*
- *TensorBoard*

⑤ Implementa y supervisa

¿Hay ciclos de reacción complejos?

Incluso si se pensó minuciosamente cada detalle del diseño del sistema en general, los modelos pocas veces funcionan con una efectividad absoluta cuando se aplican a datos reales y en tiempo real. Si surge algún problema en producción, evalúa si responde a alguna desventaja social existente y el impacto que tendrán en el producto las soluciones a corto y largo plazo.

¿Cuáles son los recursos disponibles?

- *Model Card Toolkit*
- *ML Metadata*
- *Model cards*

Recursos para definir el problema

People + AI Research (PAIR) Guidebook (1)

Es un conjunto de guías, mejores prácticas y ejemplos para diseñar una IA responsable.

Está compuesta por una serie de artículos y casos de estudio que ayudan a construir y definir el problema que se pretende resolver con IA y a ajustar las expectativas y metas en los diferentes pasos de un proyecto de data y modelado.

Para ello, se proponen una serie de patrones que responden a preguntas que ayudan a dar transparencia a los procesos: ver en qué situaciones usar IA, que esperar de ella, etc. También tiene 6 capítulos en los que ayuda a definir y responder a dudas y definir expectativas a la hora de trabajar con modelos e IA.

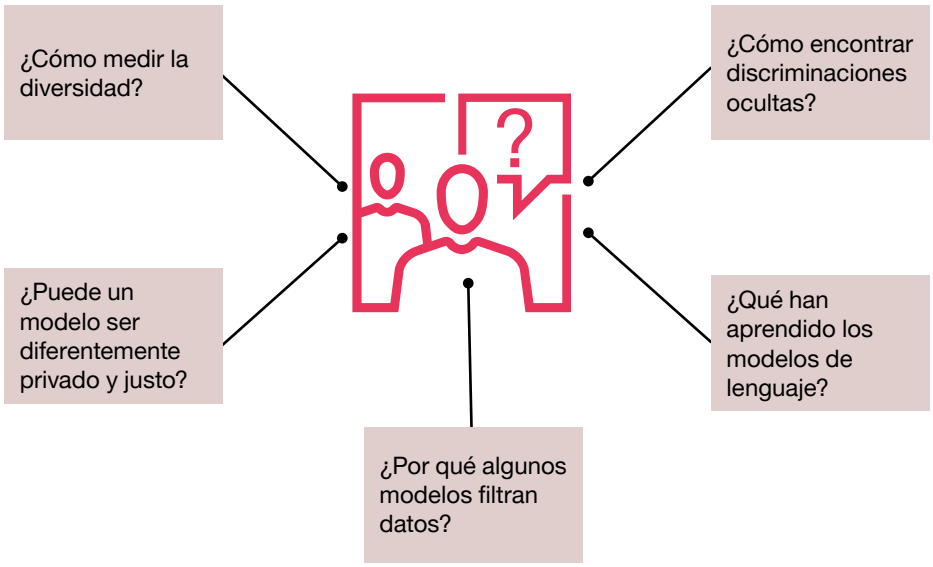
De manera adicional, tiene cinco casos de uso en los que poner estos recursos en práctica. Los cuáles resultan útiles para entender cómo estos patrones y capítulos ayudan a definir las metas y expectativas previas al comienzo de un proyecto.

Por último, presenta materiales de tipo *Workshop* con los que poner en práctica los conceptos aprendidos en una serie de ejercicios dirigidos cortos. Todo esto puede ayudar al usuario a entender que debe esperar de un modelo, como gestionar las expectativas con el usuario final, y como construirlo de manera responsable.

PAIR Explorables (2)

Son una serie de ensayos en los que se pretende dar respuesta a algunas de las preguntas más comunes que surgen debido al rápido crecimiento del uso del *Machine Learning*.

Está formada por una serie de sencillas demos interactivas y explicadas que pretenden responder a algunas preguntas como:



(1) https://pair.withgoogle.com/guidebook/
(2) https://pair.withgoogle.com/explorables/

Recursos para construir y preparar los datos

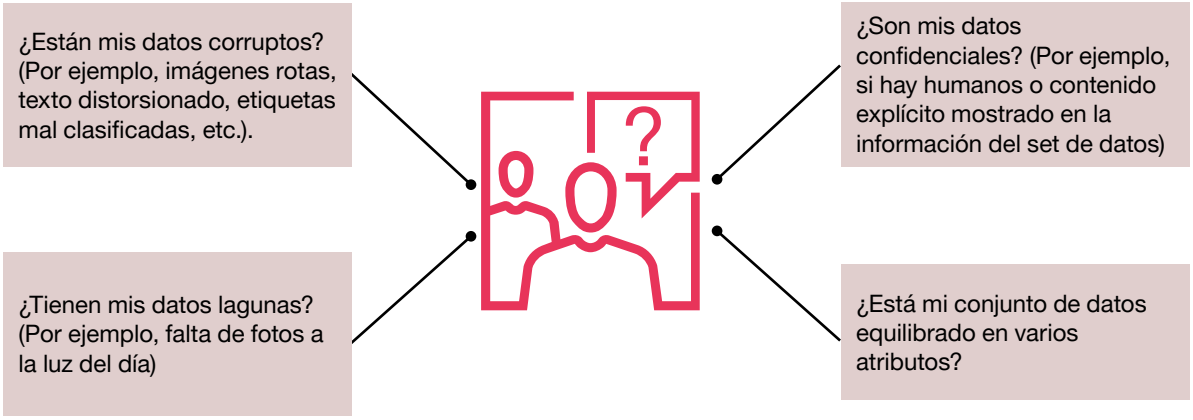
Know your data (1)

Es una herramienta *opensource* cuyo código está disponible en *GitHub* (2) que se encuentra en versión Beta y que está en desarrollo ahora mismo, que consiste en un *dashboard* interactivo por el cual es posible explorar un set de datos en formato imagen.

La versión inicial de *Know Your Data* admite conjuntos de datos de imágenes compatibles con la API de conjuntos de datos de *TensorFlow*, y en un futuro admitirá conjuntos de datos tabulares y de texto.

Esta herramienta permite filtrar y explorar en tiempo real a la vez que se clasifican las imágenes, de tal forma que es posible realizar un análisis exhaustivo de los datos disponibles. Además, permite evaluar la equidad dentro de los datos con métricas que también se calculan en tiempo real. Esta herramienta es muy interesante para conocer y explorar en profundidad datos en formato imagen.

Know Your Data tiene como objetivo responder a las siguientes preguntas:



(1) https://knowyourdata.withgoogle.com/
(2) https://github.com/pair-code/knowyourdata

TF Data Validation (1)

Se trata de una librería *open-source* diseñada para modelos *TensorFlow* y que permite realizar el análisis de distribución de las diferentes variables de un conjunto de datos. Esta librería ofrece la media y la desviación de cada variable, así como el porcentaje de nulos o una representación de las distribuciones. Además, permite identificar distribuciones no uniformes, con alto porcentaje de valores nulos o amplias diferencias en escala dentro de una variable, todo mediante un sistema de filtrado y ordenación bastante intuitivo. Esta librería resulta especialmente útil a la hora de realizar un análisis de distribución de variables, algo que suele ser recomendable en la mayoría de los casos de uso antes de comenzar con el modelado y que permite ver el aspecto que tienen tus datos, para detectar grupos minoritarios que pueden ser discriminados.

Hay muchas razones para analizar y transformar los datos:

- Para encontrar problemas en los datos. Los problemas comunes incluyen:
  - Faltan datos, como entidades con valores vacíos.
  - Etiquetas tratadas como características, para que el modelo pueda ver la respuesta correcta durante el entrenamiento.
  - Funciones con valores fuera del rango esperado.
  - Anomalías de datos.
  - El modelo de transferencia aprendido tiene un procesamiento previo que no coincide con los datos de entrenamiento.
- Diseñar conjuntos de funciones más eficaces. Por ejemplo, puede identificar:
  - Funciones especialmente informativas.
  - Funciones redundantes.
  - Funciones que varían tanto en escala que pueden ralentizar el aprendizaje.
  - Funciones con poca o ninguna información predictiva única.

La validación de datos de *TensorFlow* identifica anomalías en el entrenamiento y la entrega de datos, y puede crear un esquema automáticamente al examinar los datos. El componente se puede configurar para detectar diferentes clases de anomalías en los datos. Puede:

- Realizar verificaciones de validez comparando estadísticas de datos con un esquema que codifique las expectativas del usuario.

  La validación de datos de *TensorFlow* identifica cualquier anomalía en los datos de entrada al comparar las estadísticas de datos con un esquema. El esquema codifica las propiedades que se espera que satisfagan los datos de entrada, como tipos de datos o valores categóricos, y el usuario puede modificarlos o reemplazarlos.

  En lugar de construir un esquema manualmente desde cero, un desarrollador puede confiar en la construcción automática de esquemas de Validación de datos de *TensorFlow*. Específicamente, *TensorFlow Data Validation* construye automáticamente un esquema inicial basado en estadísticas calculadas sobre los datos de entrenamiento disponibles en la canalización. Los usuarios pueden simplemente revisar este esquema generado automáticamente, modificarlo según sea necesario, registrarlo en un sistema de control de versiones y enviarlo explícitamente a la canalización para su posterior validación. El esquema generado automáticamente es el mejor esfuerzo y solo intenta inferir propiedades básicas de los datos. Se espera que los usuarios lo revisen y modifiquen según sea necesario.

- Detectar el sesgo de servicio de entrenamiento comparando ejemplos en datos de entrenamiento y servicio.

  La validación de datos de *TensorFlow* puede detectar un sesgo de distribución entre el entrenamiento y la entrega de datos. El sesgo de distribución se produce cuando la distribución de valores de características para los datos de entrenamiento es significativamente diferente de los datos de servicio.

(1) https://www.tensorflow.org/tfx/guide/tfdv?hl=es-419

- Detectar la deriva de datos observando una serie temporal de datos.

  La detección de desviación se admite entre tramos consecutivos de datos (es decir, entre el tramo N y el tramo N + 1), como entre diferentes días de datos de entrenamiento. Puede establecer la distancia de umbral para recibir advertencias cuando la desviación sea mayor de lo aceptable. Establecer la distancia correcta es típicamente un proceso iterativo que requiere experimentación y conocimiento del dominio.

La validación de datos de *TensorFlow* proporciona herramientas para visualizar la distribución de los valores de las funciones. Mediante el examen de estas distribuciones en un cuaderno *Jupyter* usando facetas que puede detectar los problemas comunes con los datos. Puede identificar errores comunes en sus datos utilizando una pantalla de Resumen de facetas para buscar distribuciones sospechosas de valores de características lo que nos permite encontrar:

- **Datos desequilibrados**

  Puede identificar errores comunes en sus datos utilizando una pantalla de Resumen de facetas para buscar distribuciones sospechosas de valores de características. Las características más desequilibradas se enumerarán en la parte superior de cada lista de tipos de características.

- **Datos distribuidos uniformemente**

  Una característica distribuida uniformemente es aquella en la que todos los valores posibles aparecen con casi la misma frecuencia. Al igual que con los datos desequilibrados, esta distribución puede ocurrir de forma natural, pero también puede producirse por errores de datos.

- **Datos vacíos**

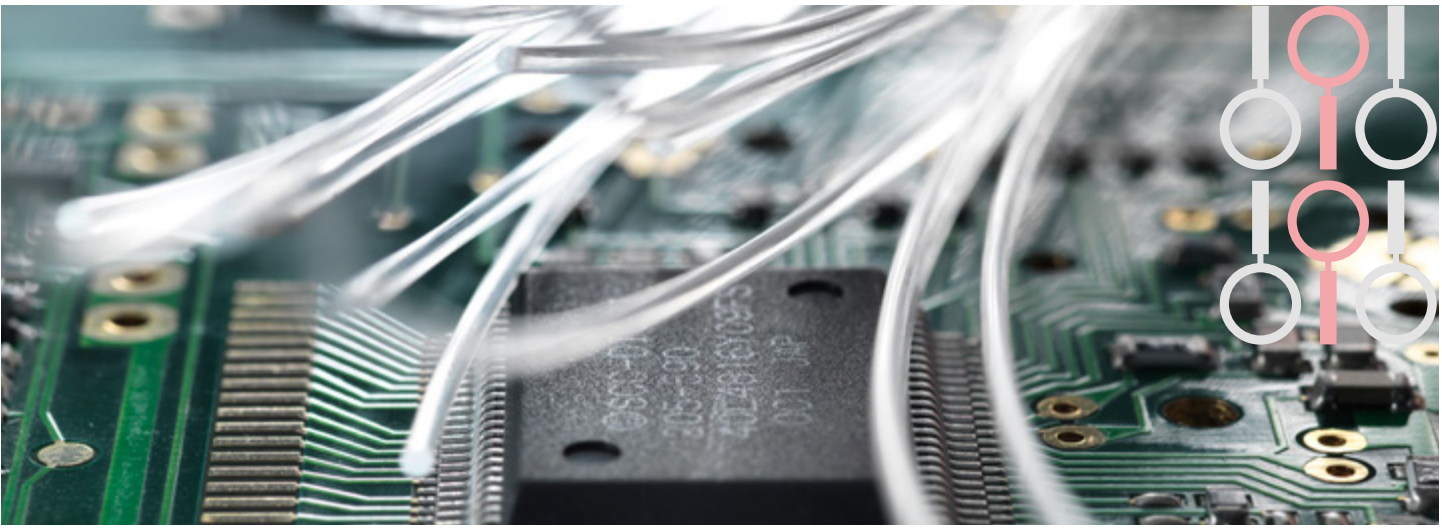
  Se puede comprobar si a una variable tiene valores nulos o vacíos. Mirando la columna *missing* para ver el porcentaje de instancias con valores nulos para una característica.

Si sus características varían mucho en escala, es posible que el modelo tenga dificultades para aprender. Por ejemplo, si algunas características varían de 0 a 1 y otras varían de 0 a 1,000,000,000, tiene una gran diferencia de escala. Considere la posibilidad de normalizar los valores de las funciones para reducir estas amplias variaciones.

Los estimadores de *TensorFlow* tienen restricciones sobre el tipo de datos que aceptan como etiquetas. Por ejemplo, los clasificadores binarios normalmente sólo funcionan con etiquetas {0, 1}. Mediante esta herramienta, se pueden validar los valores de la etiqueta en “Facetas” y asegurarse de que se ajustan a los requisitos de los estimadores.

Data Cards

Esta herramienta permite crear informes de los datos que se van a usar en el modelado, incluidos los diferentes subconjuntos que puedan aparecer como el de entrenamiento, validación, test, etc. Estos informes pretenden dar transparencia a la parte inicial de los datos, creando una documentación de estos que sea fácil de leer y que permita entender los datos. Esta herramienta también resulta útil en casi cualquier caso de uso en el que queremos documentar nuestros datos y darles transparencia.



Recursos para crear y entrenar un modelo

TF Model Remediation (1)

Es una librería *open source* publicada en *GitHub* (2) diseñada para modelos en *Keras* que se implementa para corregir las discriminaciones en la predicción de los modelos que son discriminatorias para ciertas poblaciones minoritarias. Esta librería modifica la forma en la que se entrena el modelo, añadiendo penalizaciones en esas discriminaciones a las distribuciones mayoritarias para que las predicciones de los diferentes grupos se igualen. Esta librería se puede aplicar en casos de uso en los que se sabe que existen minorías en nuestros datos y se detecta que el modelo está discriminando a esas minorías por su baja frecuencia en los datos. En general, existen tres tipos principales de intervenciones técnicas para solucionar problemas de sesgo:

- **Cambiar los datos de entrada:** recopilar más datos, generar datos sintéticos, ajustar las ponderaciones y las tasas de muestreo de las distintas porciones, etc.
- **Modificar el modelo:** cambiar el modelo estableciendo o modificando sus objetivos, agregando restricciones, etc.
- **Realizar un procesamiento posterior de los resultados:** modificar los resultados del modelo o la interpretación de ellos para mejorar el rendimiento en todas las métricas.

Dentro de la herramienta, existe la funcionalidad *MinDiff*, que es una técnica de solución de modelos cuya función es equiparar dos distribuciones. Se puede usar para equilibrar las tasas de error en distintas porciones de datos mediante la penalización de las diferencias en la distribución. Se puede aplicar *MinDiff* en los casos en que su modelo se desempeña bien en general, pero produce errores dañinos con mayor frecuencia en ejemplos que pertenecen a un grupo sensible y desea cerrar la brecha de rendimiento. Los grupos sensibles de interés pueden variar según su caso de uso, pero a menudo incluyen clases protegidas, como raza, religión, género, orientación sexual y más.

MinDiff penaliza al modelo durante el entrenamiento por la diferencia de distribución de las puntuaciones entre ambos conjuntos. Cuantas menos diferencias existan entre los conjuntos en función a la puntuación de predicción, menor será la penalización que se aplicará. También ha demostrado ser consistentemente eficaz cuando se aplica a clasificadores binarios. Es posible adaptar el método para otras aplicaciones, pero cualquier uso de *MinDiff* sobre estos u otros tipos de modelos debe ser considerado experimental.

(1) https://www.tensorflow.org/responsible_ai/model_remediation?hl=es-419
(2) https://github.com/tensorflow/model-remediation

TF Privacy (1)

Es una librería *open-source* publicada en *GitHub* (2) para modelos de *TensorFlow* que está diseñada para ayudar a mantener la privacidad de los datos en los modelos cambiando unas pocas líneas de código. Esta librería se basa en la idea de que existen técnicas a lo largo del proceso, como los métodos de optimización, que permiten ocultar la información sensible por la cual es posible identificar a individuos para construir modelos que preserven la privacidad. Esta librería es útil en todos los casos de uso en los que pretendemos modelar a partir de datos sensibles, es fácil de implementar y de forma rápida ayuda a que tu modelo tenga menos puntos de fallo en lo que a la privacidad respecta.

(1) https://www.tensorflow.org/responsible_ai/privacy/guide?hl=es-419
(2) https://github.com/tensorflow/privacy

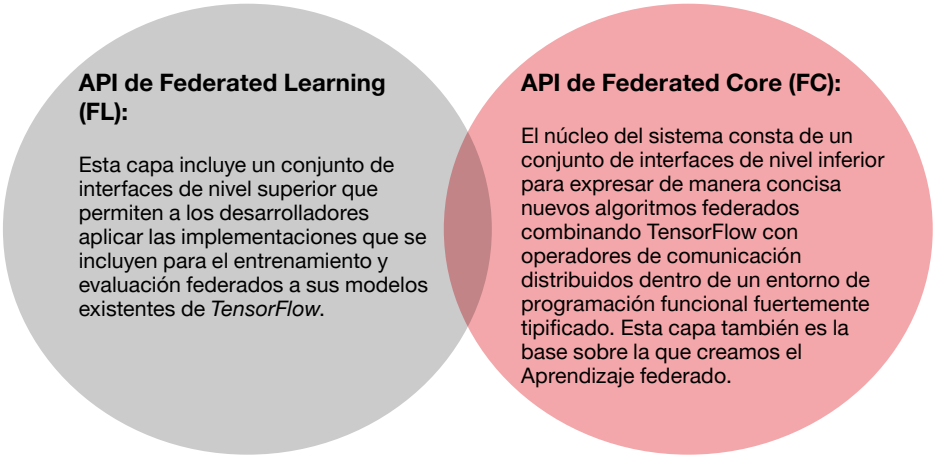
El enfoque de Google

Se consigue usando el descenso de gradiente estocástico con privacidad diferencial (DP-SGD), que es una modificación del algoritmo estándar de descenso de gradientes estocástico (SGD) del aprendizaje automático. Los modelos entrenados con DP-SGD presentan mejoras cuantificables de privacidad diferencial (DP), lo que ayuda a mitigar el riesgo de exponer datos de entrenamiento sensibles. Dado que el propósito de la DP es evitar que se identifiquen datos individuales, un modelo entrenado con DP no debería verse afectado por ningún ejemplo de entrenamiento en su conjunto de datos de entrenamiento. Las técnicas de DP-SGD también se pueden utilizar en el aprendizaje federado para ofrecer una privacidad diferencial a nivel del usuario.

TF Federated (1)

Es un marco de referencia *open-source* publicado en *GitHub* (2) de ML y otros cálculos sobre datos descentralizados, en la que *Tensor Flow* pretende fomentar el desarrollo e investigación de *Federated Learning*, un nuevo enfoque en MI que permite entrenar modelos de forma descentralizada de forma global y compartida entre los usuarios que mantienen sus datos de entrenamiento de forma local. Este nuevo enfoque se ha usado para entrenar, por ejemplo, modelos de teclado de teléfonos móviles sin subir información sensible de los teclados de los usuarios a un servidor centralizado. Esta tecnología está en desarrollo y es bastante novedosa, pero podría ser útil en casos de uso en los que los datos de entrenamiento de cada usuario sean sensibles, ya que no necesitan dar esos datos para entrenar el modelo.

TFF permite a los desarrolladores simular los algoritmos de aprendizaje federado que están incluidos en sus modelos y datos, y experimentar con algoritmos nuevos. Los componentes básicos que proporciona TFF también se pueden usar para implementar cálculos que no sean de aprendizaje, como estadísticas agregadas sobre datos descentralizados. Las interfaces de TFF se organizan en dos capas:



TFF permite a los desarrolladores expresar cálculos federados de manera declarativa, de modo que se puedan implementar en diferentes entornos de ejecución. TFF también incluye un entorno de ejecución que simula una sola máquina para realizar experimentos.

(1) https://www.tensorflow.org/federated?hl=es-419
(2) https://github.com/tensorflow/federated

TF Contrained Optimization (1)

Es una librería *open-source* diseñada para modelos *TensorFlow* que permite crear métricas personalizadas en los casos en los que hay diferentes distribuciones. En la mayoría de los casos se aplica a la función objetivo del algoritmo, por lo que es posible crearla a voluntad para evitar que los modelos sean discriminatorios. La construcción de estas funciones puede ser complicada, por lo que en la librería se hace por ratios, lo que lo simplifica y permite su construcción de forma sencilla. La creación de estas métricas personalizadas puede resultar muy útil en ciertos casos y puede ayudar con distribuciones discriminatorias, lo cual es su principal objetivo.

(1) https://github.com/google-research/tensorflow_constrained_optimization/blob/master/README.md

TF Lattice (1)

Es una herramienta *open source* publicada en *GitHub* (2) formada por un conjunto de estimadores prediseñados que permiten entrenar modelos de forma que identifique si un *input* puede o debería tener un efecto únicamente monótono en el *output*. Por ejemplo, el *input* "tiempo desde que se ha realizado una tarea por última vez" solo debería de tener un impacto positivo en predecir la probabilidad de "es hora de realizar la tarea" (3).

Permite al usuario crear modelos flexibles, controlados e interpretables de manera sencilla. Esta librería permite inyectar el conocimiento específico de cada negocio a los modelos mediante modelos creados con formas especificadas por diferentes normas o reglas que proporciona el usuario. Esta librería resulta muy útil para crear modelos o métricas basadas en distribuciones creadas manualmente por el usuario que permiten introducir conocimiento de negocio al modelo.

La librería implementa modelos basados en celosías (*lattices*). Una red (*lattice*) es una tabla de consulta interpolada que puede aproximar relaciones entrada-salida arbitrarias en sus datos. Para un punto de prueba x, f(x) se interpola linealmente a partir de los valores de celosía que rodean x. La función f(x) puede capturar las interacciones no lineales entre características. Con D características y 2 vértices a lo largo de cada dimensión, una red regular tendrá 2^D parámetros.

Dado que los parámetros de cada capa son el resultado de esa capa, es fácil analizar, comprender, depurar e interpretar cada parte del modelo. El uso de celosías de grano fino, puede obtener funciones arbitrariamente complejas con una sola capa de celosía. El uso de múltiples capas de calibradores y celosías a menudo funciona bien en la práctica y puede igualar o superar a los modelos DNN (*Deep Neural Network*) de tamaños similares.

Es posible que los datos de entrenamiento del mundo real no representen suficientemente los datos en tiempo de ejecución. Las soluciones de aprendizaje automático flexibles, como DNN, a menudo actúan de forma inesperada e incluso incontrolada en partes del espacio de entrada que no están cubiertas por los datos de entrenamiento. Este comportamiento es especialmente problemático cuando se pueden violar las restricciones de política o equidad. Aunque las formas comunes de regularización pueden resultar en una extrapolación más sensata, los regularizadores estándar no pueden garantizar un comportamiento razonable del modelo en todo el espacio de entrada, especialmente con entradas de alta dimensión. Por otra parte, cambiar a modelos más simples con un comportamiento más controlado y predecible puede tener un costo severo para la precisión del modelo.

(1) https://github.com/google-research/tensorflow_constrained_optimization/blob/master/README.md
(2) https://www.tensorflow.org/lattice/overview?hl=es-419
(3) https://ai.googleblog.com/2017/10/tensorflow-lattice-flexibility.html

El enfoque de Google

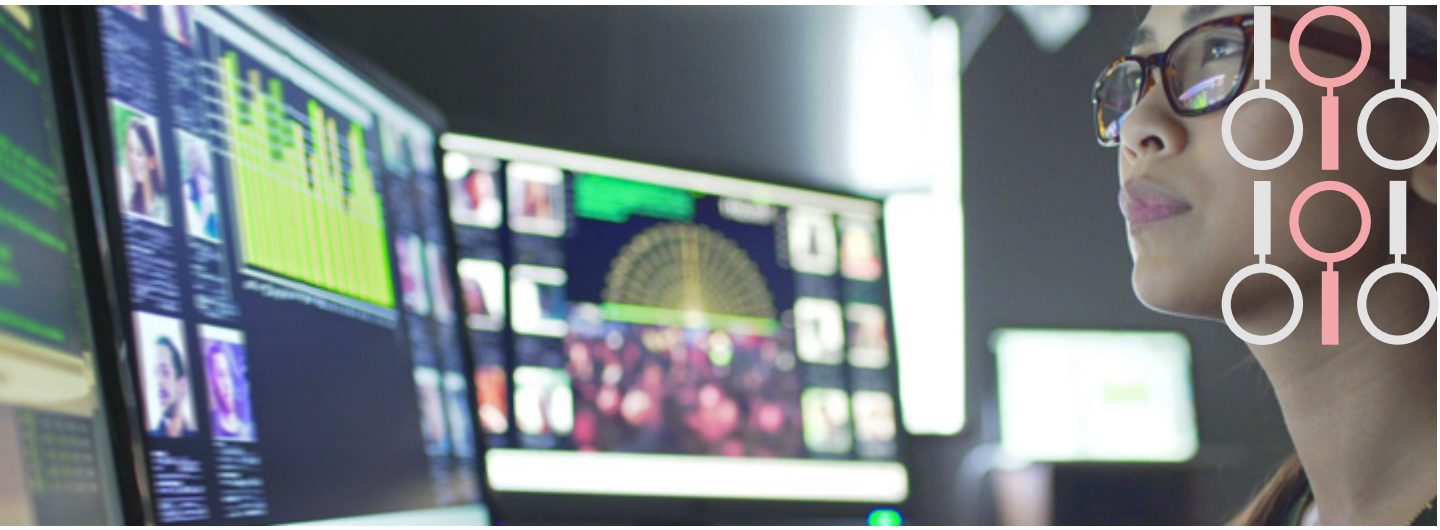
Hay varias formas de restricciones que se pueden imponer en las capas de *TensorFlow Lattice* para inyectar conocimiento del dominio del problema por parte del usuario en el proceso de entrenamiento:

- **Monotonicidad**: Se puede especificar que la salida sólo debería aumentar / disminuir con respecto a una entrada.
- **Convexidad / concavidad**: puede especificar que la forma de la función puede ser convexa o cóncava. Mezclado con monotonicidad, esto puede obligar a la función a representar rendimientos decrecientes con respecto a una característica dada.
- **Unimodalidad**: Se puede especificar que la función debe tener un pico único o singular valle.
- **Confianza por pares**: Esta restricción trabaja en un par de características y sugiere que una característica de entrada refleja semánticamente confianza en otra función. Por ejemplo, una mayor cantidad de reseñas le da más confianza en la calificación promedio de estrellas de un restaurante. El modelo será más sensible con respecto a la calificación de estrellas (es decir, tendrá una pendiente mayor con respecto a la calificación) cuando el número de reseñas sea mayor.
- **Dominancia por pares**: esta restricción sugiere que el modelo debe tratar una característica como más importante que otra característica. Esto se hace asegurándose de que la pendiente de la función sea mayor con respecto a la característica dominante.

Además de las restricciones anteriores, *TensorFlow Lattice* proporciona una serie de regularizadores para controlar la flexibilidad y suavidad de la función para cada capa.

- **Laplaciano regularizador**: Las salidas de la red / calibración vértices / puntos significativos se regularizan hacia los valores de sus respectivos vecinos. Esto resulta en una función más plana.
- **Hessian regularizador**: Este penaliza la primera derivada de la capa de calibración PWL para hacer la función más lineal.
- **Arrugas regularizador**: Este penaliza la segunda derivada de la capa de calibración PWL para evitar cambios bruscos de la curvatura. Hace que la función sea más suave.
- **Regularizador de torsión**: El modelo se regularizará hacia la independencia entre las contribuciones de las características.

Las capas *TF Lattice* se pueden combinar con otras capas de *Keras* para construir modelos parcialmente restringidos o regularizados.



Recursos para evaluar un modelo

Fairness Indicators (1)

En el momento de evaluar el rendimiento del modelo, *TensorFlow* propone *Fairness Indicators*, una librería *open-source* publicada en *GitHub* (2) que permite evaluar la justicia de cada modelo, para ello emplea una serie de métricas de justicia, así como análisis de los resultados en los diferentes subgrupos de datos, para poder así identificar discriminaciones dentro de los resultados y evaluar como de justo es un modelo. Esta librería es agnóstica al origen del modelo y se puede emplear en todos los casos de uso en los que entrenemos un clasificador, nos permitirá saber si este está siendo justo con todos nuestros datos y si el entrenamiento realizado es suficiente. Con el conjunto de herramientas de indicadores de equidad, se puede hacer lo siguiente:

- Procesar las métricas de equidad que se identifican de manera habitual para los modelos de clasificación.
- Comparar el rendimiento de los modelos entre subgrupos con un modelo de referencia o con otros modelos.
- Usar intervalos de confianza para mostrar disparidades de importancia estadística.
- Realizar la evaluación en varios umbrales.

(1) https://www.tensorflow.org/responsible_ai/fairness_indicators/guide?hl=es-419
(2) https://github.com/tensorflow/fairness-indicators

TF Model Analysis (1)

Es una librería *open-source* diseñada para analizar modelos *TensorFlow* ya entrenados. Permite generar un *dashboard* en *Jupyter* sobre el total de los datos o sobre subgrupos de estos para poder analizar la predicción del modelo, la influencia de las variables o cómo afectan las modificaciones a las métricas de precisión del modelo. Esta herramienta es muy útil para visualizar de forma interactiva los resultados de un modelo ya entrenado, aunque solo funciona con modelos de TF.

A medida que se modifica un modelo durante su desarrollo, se debe verificar si los cambios están mejorando el modelo. Verificar la precisión puede no ser suficiente. *TensorFlow Model Analysis* te permite realizar evaluaciones de modelos y ver métricas y gráficos para ayudarnos a entender el modelo y las modificaciones del mismo. Específicamente, puede proporcionar:

- Métricas calculadas en todo el conjunto de datos de entrenamiento y retención, así como en las evaluaciones del día siguiente.
- Seguimiento de métricas a lo largo del tiempo.
- Desempeño de la calidad del modelo en diferentes segmentos de características.
- Validación del modelo para garantizar que el modelo mantenga un rendimiento constante.

(1) https://www.tensorflow.org/tfx/model_analysis/install?hl=es-419

El enfoque de Google

What-if tool (1)

Herramienta *open source* publicada en *GitHub* (2), es una herramienta interactiva creada dentro de *TensorBoard*, utilizable en *Jupyter, Colaboratory, JupyterLab* y *Cloud AI Platform*, que proporciona una interfaz fácil de usar para ampliar la comprensión de un modelo ML de clasificación o regresión de caja negra. Con el complemento, se puede realizar la inferencia en un gran conjunto de ejemplos y visualizar inmediatamente los resultados de diversas maneras. Además, los ejemplos pueden editarse manualmente o mediante programación y volver a ejecutarse a través del modelo para ver los resultados de los cambios. Contiene herramientas para investigar el rendimiento y la equidad del modelo en subconjuntos de un conjunto de datos.

El propósito de la herramienta es ofrecer una forma sencilla, intuitiva y potente de jugar con un modelo de ML entrenado en un conjunto de datos a través de una interfaz visual sin necesidad de código (3).

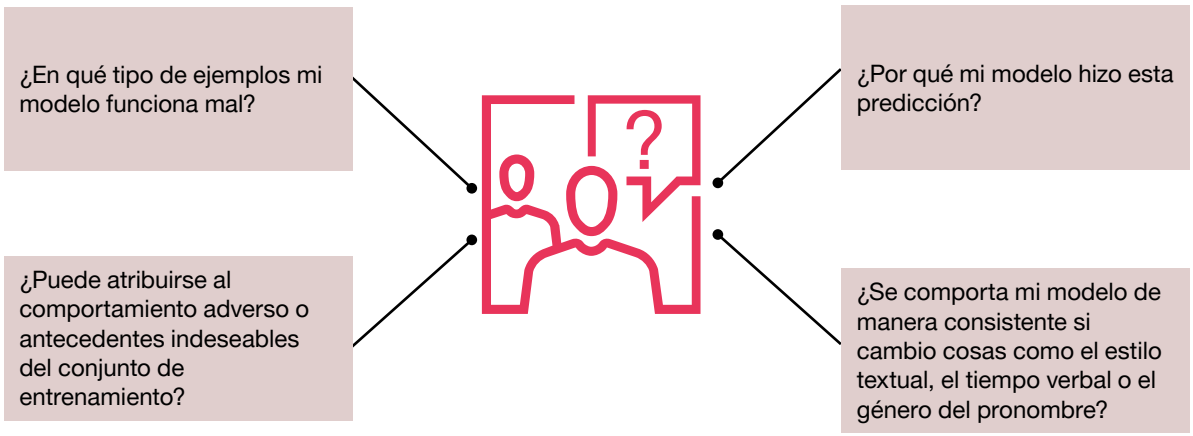
(1) https://pair-code.github.io/what-if-tool/
(2) https://github.com/pair-code/what-if-tool
(3) https://pair-code.github.io/what-if-tool/

Language Interpretability Tool (1)

Es una herramienta que pretende ayudar a los usuarios que están construyendo un modelo de procesamiento natural del lenguaje o NLP. Esta herramienta permite analizar el modelo de forma totalmente visual. Contiene una serie de métricas para analizar el modelo y permite crear métricas personalizadas, visualizaciones, técnicas de interpretabilidad, etc.

Además, está en desarrollo para otro tipo de modelos de no texto, pero para estos se recomienda otras herramientas como la de *What-If* descrita anteriormente. Esta herramienta es ideal en el caso de uso en el que se construya un modelo de texto, ya que la interpretabilidad de estos suele ser complicada, aunque puede llegar a ser interesante en muchos otros casos a medida que se adapte a otros modelos.

LIT se usa para hacer y responder preguntas como:



LIT se puede ejecutar como un servidor independiente o dentro de entornos de portátiles *Python* como *Colab, Jupyter* y *Google Cloud Vertex AI Notebooks*.

(1) https://pair-code.github.io/lit/

Explainable AI (1)

Es un conjunto de herramientas y librerías que pretende ayudar a interpretar y entender los resultados proporcionados por un modelo. Con ello, se puede mejorar y arreglar un modelos y su desempeño. De esta forma, es posible monitorizar las predicciones del modelo en tiempo real, construir herramientas propias para la monitorización de modelos y lanzar modelos a producción con confianza y seguridad. Construir herramientas aquí puede ayudar a tener un sistema de monitorización y control de fallos en los casos de uso que se desee poner el modelo en producción.

Con este conjunto de instrumentos se pude depurar y mejorar el rendimiento de los modelos, así como ayudar a otras personas a entender su comportamiento. Permiten compilar desde cero sistemas de IA que sean interpretables e inclusivos gracias a herramientas diseñadas para ayudar a detectar y resolver sesgos, desvíos y otros vacíos en los datos y modelos. *AI Explanations* en *AutoML* Tables, predicciones de *Vertex AI* y *Notebooks* nos ayudan a fortalecer la confianza del usuario final y mejorar la transparencia con explicaciones de los modelos de aprendizaje automático que pueden ser interpretadas por humanos. Entre las acciones que puede realizar se encuentran:

AI Explanations

Proporciona una puntuación que indica cómo contribuye cada factor al resultado final de las predicciones de los modelos en *AutoML Tables*, dentro de tu *notebook*, o a través de *Vertex AI Prediction API*.

Herramienta de hipótesis

Utiliza la herramienta de hipótesis integrada en *Vertex AI* para valorar el rendimiento de los modelos en una amplia gama de casos prácticos del conjunto de datos, para estudiar estrategias de optimización e incluso para manipular valores de puntos de datos concretos.

Evaluación continua

Realiza muestreos de las predicciones de los modelos de aprendizaje automático que has preparado y desplegado en *Vertex AI*, y proporciona etiquetas validadas para las entradas de predicción mediante la función de evaluación continua. El servicio de etiquetado de datos compara las predicciones de los modelos con las etiquetas validadas para ayudarte a mejorar el rendimiento de estos.

(1) https://cloud.google.com/explainable-ai?hl=es-419

TF Privacy Test (1)

Es un módulo desarrollado para comprobar la privacidad de los modelos, la librería *TF Privacy* (explicada anteriormente en la fase de Crear y entrenar un modelo), introduce ruido en los datos para que sea más difícil identificarlos y tener una fuga de privacidad. Sin embargo, se demostró que es posible crear un modelo para averiguar qué miembros del set de datos de entrenamiento contienen ruido y cuáles no, dejando a estos últimos vulnerables y su privacidad expuesta. La librería *TF Privacy* Test realiza comprobaciones sobre cómo de fácil es encontrar estos individuos expuestos en nuestros datos e indica lo fuerte que es la privacidad de un modelo.

(1) https://blog.tensorflow.org/2020/06/introducing-new-privacy-testing-library.html?hl=es-419

Tensor Board (1)

Es una herramienta *opensource* publicada en *GitHub* (2) que pretende proporcionar la métricas y visualizaciones necesarias durante todo el proceso de modelado de un algoritmo de ML. Permite comprobar métricas experimentales de precisión, visualizar gráficos del modelo o utilizar modelos de reducción de la dimensionalidad, entre otros. Esta herramienta parece especialmente útil en muchas situaciones de proceso de ML y es posible combinarlo con otras herramientas como ***What-If***, descrita previamente.

TensorBoard proporciona la visualización y las herramientas necesarias para experimentar con el aprendizaje automático:

- **Escalares de *TensorBoard*: registro de métricas de entrenamiento en *Keras***

  El aprendizaje automático implica invariablemente la comprensión de métricas clave como la pérdida y cómo cambian a medida que avanza el entrenamiento. Estas métricas pueden ayudar a entender si existe sobreajuste, o si se está entrenando innecesariamente durante demasiado tiempo. Es posible comparar estas métricas en diferentes ejecuciones de entrenamiento para ayudar a depurar y mejorar el modelo.

- **Visualización de datos de imágenes en *TensorBoard***

  Utilizando la *API Resumen TensorFlow* imagen, pueden conectarse fácilmente tensores e imágenes arbitrarias y verlos en *TensorBoard*. Esto puede ser extremadamente útil para muestrear y analizar los datos de entrada, o para visualizar los pesos de capa y tensores generados.

- **Gráfico de *TensorFlow***

  El tablero de gráficos de *TensorBoard* es una herramienta poderosa para examinar el modelo de *TensorFlow*. Se puede ver rápidamente un gráfico conceptual de la estructura del modelo y asegurarse de que coincida con el diseño previsto. También se puede ver un gráfico de nivel de operación para comprender cómo *TensorFlow* entiende el programa.

- **Visualización de datos de texto en *TensorBoard***

  Con la API de resumen de texto de *TensorFlow*, se puede registrar fácilmente texto arbitrario y verlo en *TensorBoard*. Esto puede ser extremadamente útil para muestrear y examinar datos de entrada, o para registrar metadatos de ejecución o texto generado. También se puede registrar datos de diagnóstico como texto que puede ser útil en el curso del desarrollo de su modelo.

- **Ajuste de hiperparámetros con *HParams Dashboard***

  Al crear modelos de aprendizaje automático, se debe elegir varios hiperparámetros, como la tasa de abandono en una capa o la tasa de aprendizaje. Estas decisiones afectan las métricas del modelo, como la precisión. Por lo tanto, un paso importante en el flujo de trabajo de aprendizaje automático es identificar los mejores hiperparámetros para el problema. El tablero de HParams en TensorBoard proporciona varias herramientas para ayudar con el proceso de identificar los hiperparámetros más prometedores.

- **Visualización de datos con el proyector de incrustaciones de *TensorBoard***

  Con el proyector de incrustaciones de *TensorBoard*, puede representar gráficamente incrustaciones de alta dimensión. Esto puede ser útil para visualizar, examinar y comprender sus capas incrustadas.

- **Evaluación de modelos con el panel de indicadores de equidad [Beta]**

  Los indicadores de equidad *TensorBoard* permiten un fácil cálculo de métricas de equidad comúnmente identificadas para clasificadores binarios y multiclase. En particular, los indicadores de equidad para *TensorBoard* permiten evaluar y visualizar el rendimiento del modelo, dividido en grupos definidos de usuarios.

(1) https://www.tensorflow.org/tensorboard/get_started?hl=es-419

(2) https://github.com/tensorflow/tensorboard/blob/master/docs/get_started.ipynb

- **TensorFlow Profiler: rendimiento del modelo de perfil**

  Los algoritmos de aprendizaje automático suelen ser computacionalmente costosos. Por lo tanto, es vital cuantificar el rendimiento de la aplicación de aprendizaje automático para asegurarse de que está ejecutando la versión más optimizada del modelo. *TensorFlow Profiler* se usa para perfilar la ejecución del código de *TensorFlow*.

- **Usar TensorBoard con Notebooks**

  TensorBoard se puede utilizar directamente en experiencias portátiles tales como *Colab y Jupyter*. Esto puede ser útil para compartir resultados, integrar *TensorBoard* en flujos de trabajo existentes y usar *TensorBoard* sin instalar nada localmente.

TensorBoard.dev te permite alojar los resultados de tu experimento, hacerles un seguimiento y compartirlos fácilmente.

Recursos para implementar y supervisar un modelo

Model Card Toolkit -MCT-(1)

Es una librería que pretende dar transparencia e interpretabilidad a los modelos, para ello, se crean unas tarjetas en formato HTML que contienen la información de dicho modelo, con detalles de los sets de datos utilizados, del modelo y de los resultados de este, para favorecer y facilitar la explicabilidad e interpretabilidad de dichos modelos. Esta librería es muy útil cuando se desarrolla un modelo y se desea que este sea lo más transparente posible, y, sobre todo, cuando se tenga que presentar dicho modelo a terceros. Integrando el kit de herramientas *Model Card* en tus canalizaciones de AA, podrás compartir los metadatos y métricas de tu modelo con investigadores, desarrolladores, periodistas o cualquier otra tercera persona.

El MCT almacena campos de tarjetas de modelos mediante un esquema de JSON. puede propagar esos campos automáticamente para los usuarios de TFX mediante los Metadatos de AA (MLMD). Los campos de tarjetas de modelos también se pueden propagar de forma manual mediante una API de *Python*. Estos son algunos casos de uso de tarjetas de modelos:

- Facilitar el intercambio de información entre los compiladores de modelos y los desarrolladores de productos.
- Brindar información a los usuarios de modelos de AA para que tomen decisiones mejor fundamentadas sobre cómo utilizarlos (o cómo no utilizarlos).
- Proveer la información del modelo necesaria para asegurar la supervisión pública y responsabilidad eficientes.

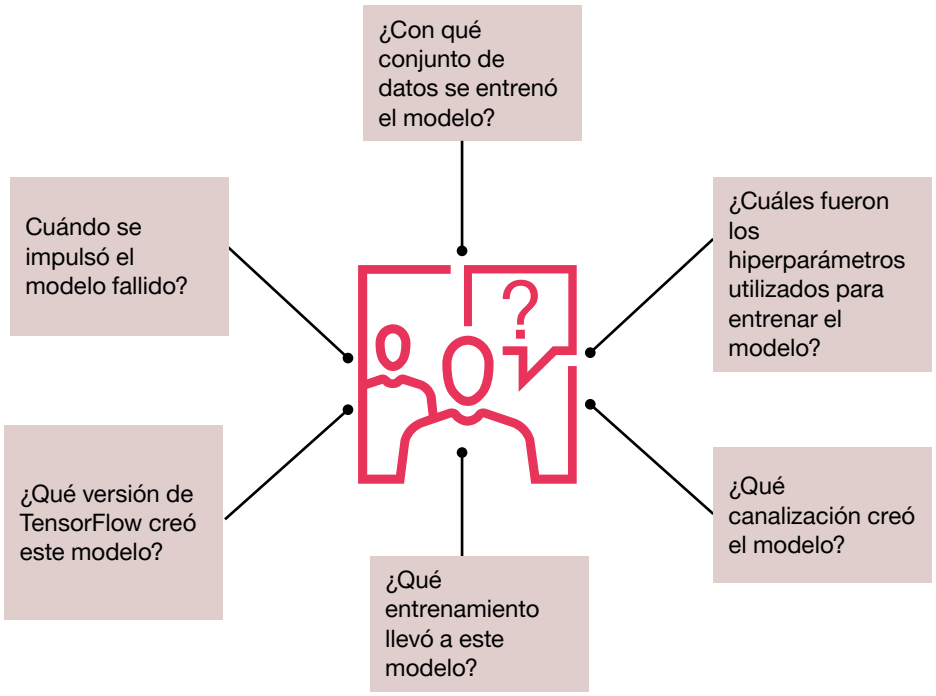
(1) https://www.tensorflow.org/responsible_ai/model_card_toolkit/guide?hl=es-419

MLMetadata (1)

Es una librería que forma parte de *TensorFlow Extended*, pero que ha sido desarrollada para usarse en solitario. Esta librería permite crear un sistema de *logging* dentro de un proceso de ML, es decir, almacena la información de cada paso dentro del proceso, como tiempo de ejecución, los componentes y los resultados, para que en el caso de que algo no funcione dentro de este proceso esa información esté disponible y sea posible encontrar errores y solucionarlos. Esta librería puede resultar útil siempre que se construya un modelo de ML, ya que, este tipo de información almacenada permite encontrar de forma fácil y rápida los errores, algo muy útil a medida que los procesos de modelado se hacen grandes y complicados.

Cada ejecución de una canalización de ML de producción genera metadatos que contienen información sobre los diversos componentes de la canalización, sus ejecuciones y los artefactos resultantes.

En caso de errores o comportamientos inesperados de la canalización, estos metadatos se pueden aprovechar para analizar el linaje de los componentes de la canalización y depurar problemas. MLMD lo ayuda a comprender y analizar todas las partes interconectadas de su canal de ML en lugar de analizarlas de forma aislada y puede ayudarlo a responder preguntas como:



(1) https://www.tensorflow.org/tfx/guide/mlmd?hl=es-419

Model Cards

Una mayor transparencia para los modelos de aprendizaje automático puede beneficiar a todos, por lo que las tarjetas modelo están dirigidas tanto a expertos como a no expertos. Los desarrolladores pueden usarlos para diseñar aplicaciones que enfaticen las fortalezas de un modelo mientras evitan o informan a los usuarios finales sobre sus debilidades. Para periodistas y analistas de la industria, pueden proporcionar información que facilite la explicación de tecnología compleja a una audiencia general.

La transparencia en la IA debería ser un bien común. Es por eso que las tarjetas modelo no pretenden ser un producto de Google, sino un marco compartido y en evolución formado por una variedad de voces. Esto incluye usuarios, desarrolladores, partes interesadas de la sociedad civil y empresas de toda la industria, así como colaboraciones con organizaciones como *Partnership on AI* y su *ABOUT ML Project*.

(1) https://modelcards.withgoogle.com/about

3. Framework GuIA

Cómo aterrizar cada principio ético

El enfoque de Microsoft

El enfoque de Microsoft para gestionar los principios éticos

En esta sección dedicada a los planteamientos de Microsoft vamos a desarrollar el marco de trabajo de dicho fabricante para conseguir una IA ética. Para ello profundizaremos sobre los siguientes contenidos:

1. **Propósitos globales** de Microsoft respecto de la IA ética.
2. **Principios éticos** sobre los que Microsoft se focaliza.
3. **Organización y recursos internos** de Microsoft para asegurar que sus productos utilizados por terceros cumplen con sus propósitos globales y principios éticos sobre los que Microsoft se focaliza.
4. **Recomendaciones metodológicas** a lo largo del ciclo de vida de las soluciones inteligentes y que son comunes a todos los principios éticos para conseguir una IA ética y responsable.
5. Para cada principio ético, el **toolkit *de Microsoft con las tecnologías, herramientas y Guidelines*** que facilitan su gestión.

De esta manera iremos profundizando **desde una visión global a una aterrizada**. Los tres primeros puntos permitirán a cualquier entidad entender planteamientos globales que poder adoptar en su estrategia, su organización y procesos de gestión. El punto 4 permitirá a dichas entidades aplicar este enfoque en el día a día de sus operaciones, en el *Delivery* de sus proyectos que tengan a la inteligencia artificial como su tecnología principal.

Propósitos globales de Microsoft respecto de la IA ética

Microsoft se plantea tres grandes objetivos globales en IA ética que a lo largo de presente capítulo iremos aterrizando para permitir que cualquier entidad pueda asimilarlos ateniendo a la realidad de su día a día.

Innovar de forma responsable

En Microsoft, ponen en práctica sus principios adoptando un enfoque centrado en las personas, para la investigación, desarrollo e implementación de las soluciones de IA. Para lograr esto, adoptan distintas perspectivas y se basan en el aprendizaje continuo y una respuesta ágil a la evolución de las tecnologías de IA.

Empoderar

Ayudar a las empresas y organizaciones a adoptar una cultura preparada para la IA responsable en todas sus líneas de negocio, y a implementar los principios en todas las fases del ciclo de vida de la solución, proporcionando recomendaciones, herramientas y tecnologías basadas en la investigación multidisciplinar, el aprendizaje compartido y la innovación.

Fomentar un impacto positivo

En Microsoft están comprometidos a garantizar que la tecnología de IA tenga un impacto positivo duradero, ayudando a dar forma a las leyes, contribuyendo en la industria y capacitando a trabajadores para abordar los mayores desafíos de la sociedad.

Principios éticos sobre los que Microsoft se focaliza

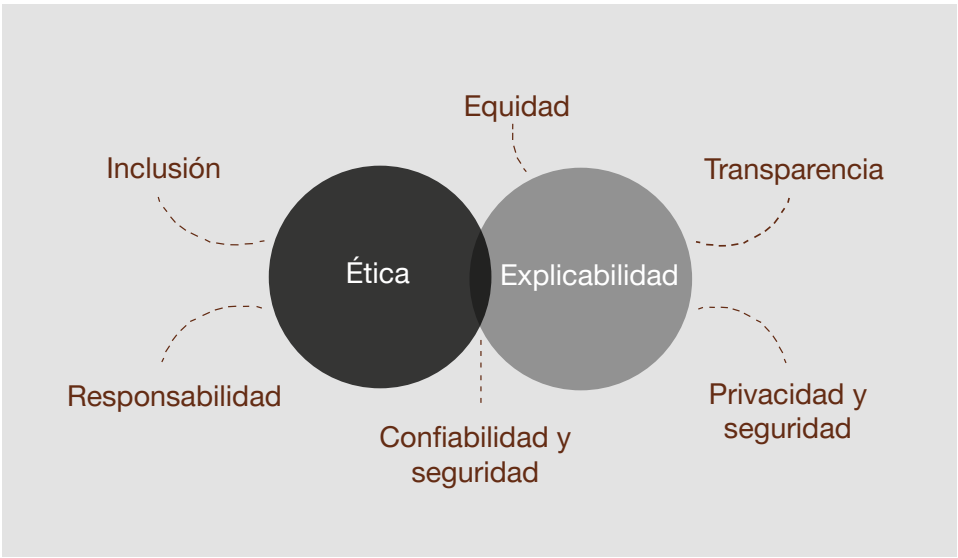
Microsoft contempla seis principios clave a la hora de abordar la IA responsable:

1. Equidad
2. Confiabilidad y Seguridad
3. Privacidad y Seguridad
4. Inclusión
5. Transparencia
6. Responsabilidad

Fuente: Microsoft

Estos principios son esenciales para garantizar una IA responsable y confiable, teniendo en cuenta que se avanza hacia la adopción de modelos de IA cada vez más cotidianos. **Microsoft engloba dichos principios desde dos perspectivas**, la ética y la explicabilidad.

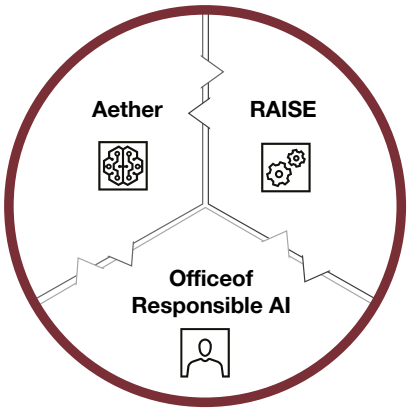
- **Ética**. Desde esta perspectiva, la IA debe ser justa e inclusiva en sus afirmaciones, además de asumir la responsabilidad de sus decisiones, sin discriminar ningún tipo de raza, discapacidad o entorno.
- **Explicabilidad**. La explicabilidad ayuda a científicos de datos, auditores y figuras responsables de las decisiones de negocio a garantizar que los sistemas de IA sean capaces de justificar de forma razonable las decisiones que toman y las conclusiones a las que llegan. Esto también asegura el cumplimiento de las políticas de empresa, estándares de la industria y regulaciones gubernamentales asociadas a cada caso de uso. Un científico de datos debería poder explicar a los distintos *stakeholders* cómo se han logrado los niveles de precisión, y qué variables han influido en el resultado. Asimismo, para cumplir con las políticas de empresa, un auditor necesita una herramienta que valide el modelo, y desde la parte de negocio, se necesita poder proporcionar un modelo transparente para generar confianza.



Fuente: Microsoft

Gobierno de Microsoft para una IA ética

El enfoque de gobernanza de Microsoft para una IA ética sigue un modelo de *hub-and-spoke* que ayuda a la empresa a integrar la privacidad, la seguridad y la accesibilidad en sus productos y servicios. Existen tres equipos desempeñan un papel fundamental en este enfoque de gobernanza.



1. En primer lugar, el comité **AETHER** (1) comprende grupos de trabajo de expertos científicos y de ingeniería para asesorar sobre cuestiones de IA responsable y la aplicación de los de los principios de IA de la empresa, así como un comité de evaluación.

2. En segundo lugar, la Office of Responsible AI (**ORA**) que gestiona las políticas internas, la gobernanza, la habilitación de equipos y la revisión de los casos de uso sensibles:

- Gobernanza. Establecer las reglas de la empresa y promulgar una IA responsable, además de definir roles y responsabilidades de los equipos involucrados
- Habilitación de los equipos. Determinar la disposición de los equipos a adoptar prácticas de IA responsable, tanto internamente como entre los clientes y *partners*
- Revisión de casos de uso sensibles. Revisar casos de uso sensibles para ayudar a garantizar que los principios de IA de Microsoft se mantienen durante el desarrollo e implementación.
- Políticas. Ayudar a dar forma a nuevas leyes, normas y estándares que serán necesarios para garantizar una IA responsable.

Fuente: Microsoft

3. **En tercer lugar el *Responsible AI Strategy in Engineering*** (RAISE) equipo que habilita a los equipos de ingeniería de Microsoft implementar procesos de 'IA responsable' de Microsoft ingeniería través de la adopción de herramientas y sistemas.

El RAISE y el Aether están formadas por equipos de trabajo multidisciplinares que gestionan las áreas de especialización en el ámbito de la ingeniería y *research* para IA responsable.

Fuente: Microsoft

(1) https://www.microsoft.com/en-us/ai/our-approach?activetab=pivot1%3aprimaryr5

El enfoque de Microsoft

Esta estructura interna de Microsoft para la definición e implementación de sus estrategias internas entra en funcionamiento cuando un *partner* de Microsoft presenta un proyecto de IA, identificado como potencialmente uso sensible conforme a los criterios establecidos por Microsoft. En tal caso, el proyecto es inicialmente revisado por el *Responsible AI Champ*, quien activa los mecanismos internos de Microsoft para la gestión interna del proyecto y dirige la petición al equipo de ORA (equipo de casos de uso sensibles). Para dar transparencia al proceso, el *Responsible AI Champ* informa al *partner* acerca de la revisión interna del proyecto e informa sobre los siguientes pasos en el proceso de revisión. En caso de requerir más información, por lo general, el siguiente paso es recopilar información adicional que ayudará en la revisión del caso. Esta información se recopila a través de una evaluación de impacto del equipo del proyecto y que puede ser mediante una reunión o correo electrónico.

En los casos considerados de uso sensibles, el *RAI Champ* y la ORA pueden definir los medios más idóneos para obtener y proporcionar orientación de sobre el caso de uso a revisar que pueden incluir:

- Orientación técnica
- Consulta con expertos en la materia
- Envío al equipo de implementación de la RAI correspondiente
- Envío al panel de usos sensibles

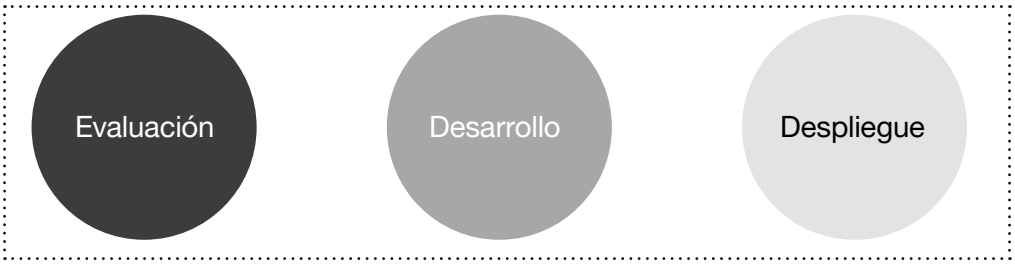
La emisión de una guía escrita por parte de Microsoft es el paso final en el proceso de revisión de usos sensibles. La orientación emitida es válida para el caso de uso en cuestión, a menos que se indique lo contrario en el documento de orientación.

La revisión no sustituye otros requisitos legales, de cumplimiento y de directivas (por ejemplo, RGPD, revisiones de seguridad).

Recomendaciones metodológicas en todo el ciclo de vida

Microsoft ofrece una aproximación metodológica que es la base a lo largo del ciclo de vida de las soluciones inteligentes y que son comunes a todos los principios éticos para conseguir una IA ética y responsable. Dichas recomendaciones son utilizadas internamente por Microsoft.

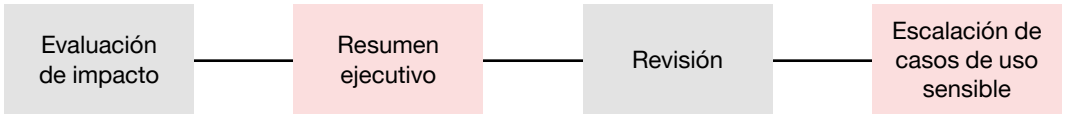
Este ciclo de vida tiene varias fases que se definen a continuación:



En cada fase se propone una serie de requisitos a analizar que dan pie a la gestión de los principios éticos propuestos por Microsoft.

Fase de Evaluación

Durante la fase de evaluación se han de cumplir una serie de requisitos que se organizan en los siguientes bloques:



Evaluación de impacto

Para acometer la evaluación de impacto de un sistema inteligente Microsoft en primer lugar se centra en **identificar el propósito de la aplicación**, quienes van a ser los **responsables del despliegue** y cuáles serán los **beneficios, daños** o **tensiones** que este sistema inteligente puede provocar. A partir de esa primera evaluación de impacto se establecen los **planes de ejecución** incluyendo su desarrollo, despliegue y posterior evolución del sistema.

Esta evaluación del impacto se documenta para que así pueda posteriormente revisar por parte de los equipos responsables.

Resumen ejecutivo

En el resumen ejecutivo se incluyen los impactos más relevantes y se busca de manera proactiva potenciales daños, tensiones y beneficios detectados a partir de los inicialmente identificados.

Revisión

A continuación se analiza la información recolectada en los dos anteriores pasos y se pone a disposición de los steakholders del proyecto: usuarios finales, integrador o socio encargado del desarrollo de la solución, y a la organización interna de Microsoft descrita anteriormente (AETHER, RAISE y ORA). Dichos *stakeholders* revisan que el enfoque del proyecto se ajusta al marco ético requerido por el caso de uso.

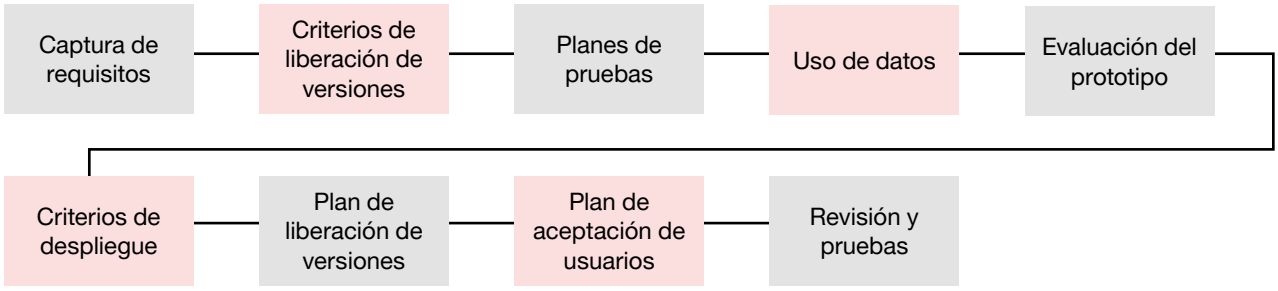
Escalación de los casos de uso sensibles

Si se identifican casos de uso sensibles se escalan y se identifican proactiva para así ofrecer una respuesta adecuada para la realización del proyecto.

Fase de Desarrollo

Durante la fase de desarrollo se realizan diferentes tareas que se apoyan en sistemas y herramientas que veremos más adelante.

Las actividades propuestas en esa fase son las tres siguientes:



Captura de requisitos

La aproximación sigue un modelo similar al modelo de desarrollo tradicional, se priorizan los requerimientos principales (P0 y P1), recogidos en la fase de Evaluación, identificando los principales modelos de IA que van a formar parte de la solución global, se establecen los principales conjuntos de datos requeridos para cubrir la funcionalidad global de la solución y, fundamental, se define el modelo de gobierno de dichos conjuntos de datos para el cumplimiento de los requisitos éticos identificados previamente.

Por último, se definen las estrategias para la gestión de los requisitos éticos, con el objetivo de mitigar en todo lo posible los daños y tensiones previamente identificados.

Criterios de liberación de versiones

Esta actividad es muy relevante, ya que en ella se analizan los criterios de equidad buscando un equilibrio con la precisión del modelo, es decir, que la equidad del modelo no penalice en exceso su precisión y su rendimiento. Además, se identifican los requisitos relacionados con anonimización de datos, y los mecanismos necesarios para gestionar la transparencia del modelo mediante la explicabilidad del mismo, etc.

Los criterios de liberación de versiones también deben incluir criterios de inclusión y accesibilidad al sistema inteligente.

Planes de pruebas

Se definen los planes de pruebas para el sistema inteligente poniendo foco en la identificación de casos de prueba que pueden colisionar con los principios éticos.

Uso de datos

Se profundiza sobre el análisis de la privacidad de los datos, identificando a los responsables de la recolección de estos, se procede al etiquetado de los datos y se limitan los casos de uso que pueden utilizar dichos datos. Aquí también se dota de transparencia al proceso en el uso de datos.

Evaluación del prototipo

En esta fase se identifican y analizan los potenciales escenarios de fallo del modelo que sustenta el sistema inteligente, se buscan vulnerabilidades con técnicas de muestra adversaria y se define el plan para mitigarlos. Ya con el detalle que proporciona la evolución en el prototipo, se evalúa de nuevo el mismo con el objetivo de identificar los casos de uso sensibles por si hubiese alguno nuevo que elevar al comité.

Criterios de despliegue

En esta actividad se evalúan las condiciones de despliegue, se toman en consideración los datos que ya se pueden utilizar en un escenario productivo, y así ya poder hacer una definición del testeo holístico de todas las funcionalidades.

Además de estas actividades en esta fase pone especial foco en la actividad de *Ring Testing*, también conocido como *friends & Family*, donde se definen las pruebas y usuarios objetivo para las pruebas de lo que será una versión cercana al producto final, con datos y escenarios reales. Es en esta actividad donde se pueden identificar nuevos posibles *gaps* respecto del uso ético y responsable de la solución inteligente.

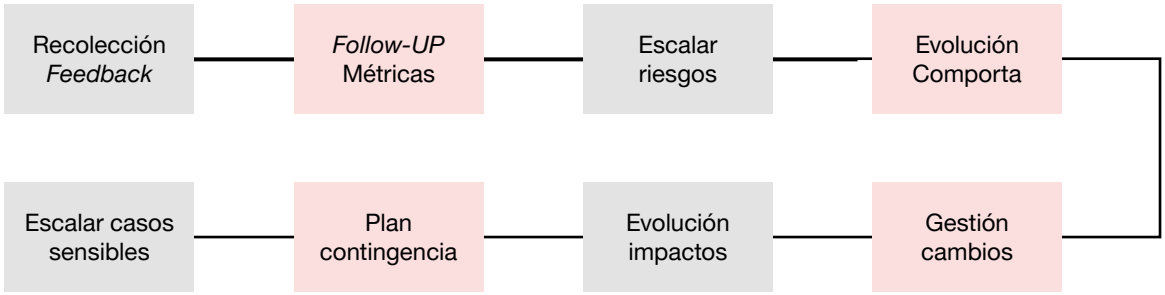
Plan de liberación de versiones

Esta actividad comprende las pruebas finales del sistema así como pruebas por parte de los usuarios incluyendo un conjunto lo más heterogéneo y diverso posible. Se monitoriza el *feedback* recibido y se libera la documentación. A menudo también se trabaja en los términos y condiciones de uso y el plan de *rollback*.

Revisión y pruebas

Durante esta actividad se ejecutan los planes de revisión y pruebas, teniendo en consideración los principios éticos definidos y finalmente se escalan los potenciales *gaps* encontrados.

Fase de Despliegue



Esta es la fase en la que la solución inteligente pasa a producción. Antes de iniciar esta fase es fundamental habilitar los canales que permitan proporcionar *feedback* acerca de los casos de uso y aspectos éticos que han sido contemplados, para garantizar su correcto funcionamiento confirme a las estrategias definidas para su gestión.

En esta fase se pueden monitorizar métricas de rendimiento y precisión real que tienen relación con el impacto ético. Pero sobre todo hacer una evaluación en base al *feedback* de los usuarios contra los principios éticos. También es posible en esta fase elevar los riesgos identificados a los diferentes comités para su estudio.

En este apartado cabe la mención especial a otras métricas que se establecen no tan sólo en la fase de despliegue (inferencia) sino en las fases de desarrollo y que tienen que ver con el impacto medioambiental visto desde la perspectiva ética. Microsoft está poniendo foco en introducir la ingeniería sostenible (1) como un elemento consustancial al desarrollo de sistemas inteligentes no ya pensando sólo en el *Green by AI* sino también en el Green in AI. (2) (3)

Adicionalmente a estas actividades en la fase de despliegue se incluye además las actividades relacionadas con la **evolución de los sistemas inteligentes**, el control y seguimiento de su funcionamiento y otros aspectos tan importantes como los planes de contingencia y las nuevas aplicaciones y el impacto que puedan tener en los principios éticos.

(1) Principios de la ingeniería de software sostenible - Learn | Microsoft Docs
(2) DeepSpeed powers 8x larger MoE model training with high performance - Microsoft Research
(3) AsyMo: Scalable and Efficient Deep-Learning Inference on Asymmetric Mobile CPUs - Microsoft Research

Seguimiento del comportamiento

Se asegura que los principios de equidad, confiabilidad y seguridad se cumplen y están en los rangos definidos en los criterios de liberación de versiones.

Gestión de cambios

Se lleva un control de las nuevas características, los parches a aplicar, la regulación y las políticas que deben cumplirse y se escala a los *stakeholders* correspondientes la documentación necesaria para evaluar dicha evolución.

Impacto de las nuevas aplicaciones

En esta actividad se evalúa el impacto que tendrá la evolución del sistema inteligente en las personas, y se evalúan los retos que se presentan, los potenciales daños que pueda ocasionar y se identifican los casos de uso sensibles para su posterior evaluación a los comités.

Planes de contingencia

Los planes de contingencia se documentan principalmente para definir un plan de *rollback*, un plan para decomisionar el sistema inteligente y un plan para la gestión de los datos.

Escalación de nuevos casos de uso sensibles

Obviamente durante la evolución de un sistema inteligente se siguen aplicando los mismos criterios de escalación de nuevos casos de uso sensible. Poniendo especial foco en aquellos que no se han podido remediar durante la conceptualización de las nuevas evoluciones.

Resumiendo durante la evolución el foco está en asegurar que los estándares éticos tenidos en cuenta en la evaluación y el desarrollo de la solución se siguen manteniendo con la evolución funcional de la misma forma.

Toolkit de Microsoft para la gestión de los principios éticos

En este apartado, para cada principio ético, profundizaremos en las **tecnologías, herramientas y *Guidelines* propuestas por Microsoft** que facilitan su gestión.

Equidad

- Qué es la equidad para Microsoft.
- *FairLearn*, tecnología para la valuación y mitigación de parcialidad.
  - Métricas para detectar la parcialidad.
  - Algoritmos para mitigar la parcialidad.
- Herramienta de gestión. *AI Fairness Checklist*.

Confiabilidad y Seguridad

- Qué es la Confiabilidad y la seguridad para Microsoft.
- Recomendaciones en el ámbito de Confiabilidad y seguridad.
- Tecnologías para aplicar la transparencia.
  - Error *Analysis*, para identificar y diagnosticar errores.
  - *Counterfit*, para la evaluación de riesgos de seguridad.

Privacidad y Seguridad

- Qué es la privacidad y la seguridad para Microsoft.
- Recomendaciones en el ámbito de privacidad y seguridad.
- Tecnologías para aplicar la privacidad y seguridad.
  - Presidio, para la protección y anonimización del dato.
  - *Smart Noise*, para la Privacidad Diferencial.
  - *SEAL Homomorphic Encryption*, para cálculos sobre datos cifrados en *cloud*.

Inclusión

- Como Microsoft entiende la inclusión en la IA.
- *Guidelines* para un diseño centrado en las personas
  - Para la interacción humano-IA (*Human-AI Experiences -HAX-*).
  - *Inclusive Design Guidelines*.
  - *Confidencial computing for ML*, infraestructura *cloud* Segura y privada.

Transparencia

- Qué es la Transparencia para Microsoft.
- Prácticas recomendadas.
- Tecnologías para aplicar la transparencia
  - *InterpetML* para explicar el funcionamiento de los modelos.
  - DiCE para gestionar la contrafactualidad.
  - *EconML* para la inferencia causal.
- Herramienta de gestión. *Datasheets for Datasets*.

Responsabilidad

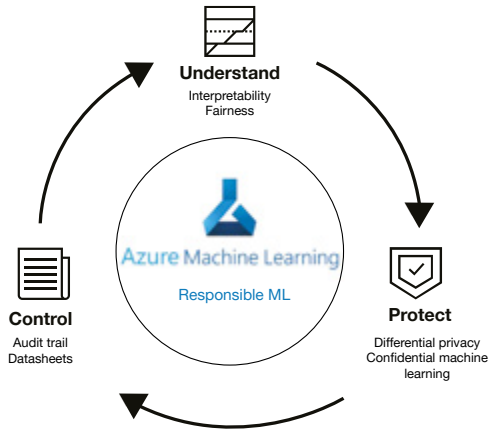
- Qué es la Responsabilidad para Microsoft.
- La involucración del ser humano en las soluciones inteligentes.

El enfoque de Microsoft

Como puede observarse, todos los principios tienen asociados unas recomendaciones o buenas prácticas y unas herramientas. Sobre estas últimas, Microsoft establece unos objetivos globales que denomina **"Principios de aprendizaje automático responsable"**.

Dichos principios exponen como la confianza debe estar en el centro de la IA permitiendo:

- Comprender los modelos de aprendizaje automático.
  - Interpretando y explicando el comportamiento del modelo.
  - Evaluando y mitigando la injusticia del modelo.
- Proteger a las personas y sus datos
  - Evitando la exposición de datos con privacidad diferencial
  - Trabajando con datos encriptados usando encriptación homomórfica
- Controlar el proceso de aprendizaje automático de extremo a extremo
  - Documentando el ciclo de vida del aprendizaje automático con hojas de datos.



Fuente: Microsoft

Principio de Equidad

En esta sección vamos a desarrollar:

- Qué es la equidad para Microsoft
- *FairLearn*, tecnología para la valuación y mitigación de parcialidad
  - Métricas para detectar la parcialidad
  - Algoritmos para mitigar la parcialidad
- Herramienta de gestión. *AI Fairness Checklist*.

Qué es la equidad para Microsoft

Los sistemas de inteligencia artificial (IA) están jugando un papel cada vez más importante en la vida diaria de las personas; por lo que conseguir que operen de manera equitativa es de vital importancia. Esto significa que los sistemas de IA deben tratar a todos de manera justa y evitar que realicen un trato distinto entre grupos de personas que se encuentran en la misma situación. Por ejemplo, cuando los sistemas de IA ofrecen orientación sobre tratamientos médicos, solicitudes de préstamos o empleos, deben hacer las mismas recomendaciones a todas las personas con síntomas, circunstancias financieras o calificaciones profesionales similares.

Como puntos destacables, es conveniente tener en cuenta que:

- Los seres humanos cargan con prejuicios culturales y esos sesgos afectan a las decisiones y acciones que se registran en los datos que son utilizados para entrenar los sistemas de IA.
- Es necesario asegurarse de que la IA sea justa y no esté programada para tomar decisiones sesgadas o discriminatorias, de igual manera que los seres humanos también tienen prohibido tomarlas.
- Conseguir una IA “justa” es actualmente uno de los campos activos de investigación en referencia a la perspectiva ética de la IA.

Por tanto, la inteligencia artificial y los sistemas de aprendizaje automático pueden mostrar un comportamiento parcial. Una manera de definir el comportamiento parcial es fijarse en el daño que hace o impacto que tiene en las personas. Existen muchos tipos de daños que los sistemas de IA pueden ocasionar. Estos son dos tipos comunes de daños que puede causar la IA:

- Daño en la asignación: Un sistema de inteligencia artificial extiende o retiene oportunidades, recursos o información para determinados grupos. Entre los ejemplos se incluyen la contratación, las admisiones escolares y los préstamos, donde un modelo podría ser mucho mejor a la hora de seleccionar buenos candidatos entre un grupo específico de personas que entre otros grupos.
- Daño en la calidad del servicio: un sistema de AI puede no funcionar tan bien para un grupo de personas como lo haría en otro. Por ejemplo, un sistema de reconocimiento de voz puede no funcionar correctamente para las mujeres, pero sí para los hombres.

Para reducir el comportamiento parcial de los sistemas de AI, tiene que evaluar y mitigar estos daños.

Necesidades para resolverlo

Por otra parte, es fundamental hacer de la equidad una prioridad en todo el ciclo de vida de desarrollo e implementación de la IA, mediante la identificación de las necesidades de los profesionales de la industria para desarrollar sistemas de IA más justos y comprender los desafíos y oportunidades organizacionales entorno a la equidad en la IA.

Mejorar la equidad en los sistemas de aprendizaje automático: ¿qué necesitan los profesionales de la industria? (1)

La capacidad de los sistemas de *Machine Learning* (ML) para aumentar las desigualdades sociales y la injusticia está recibiendo cada vez más atención popular y académica. Por lo que, recientemente, se han desarrollado herramientas basadas en algoritmos para evaluar y mitigar estos efectos. Sin embargo, para que estas herramientas tengan un impacto positivo en la práctica de la industria, es fundamental que su diseño se base en la comprensión de las necesidades del mundo real.

(1) https://www.microsoft.com/en-us/research/publication/improving-fairness-in-machine-learning-systems-what-do-industry-practitioners-need/

El enfoque de Microsoft

A través de treinta y cinco entrevistas semiestructuradas y una encuesta anónima de doscientos sesenta y siete usuarios que utilizan sistemas de ML, Microsoft ha llevado a cabo la primera investigación sistemática de los desafíos y necesidades de los equipos de las empresas para desarrollar sistemas de ML más justos. Se han identificado áreas que coinciden con los desafíos ya identificados que enfrentan los profesionales de la industria y las soluciones propuestas en la literatura de investigación de ML justa para solventarlos, y otras áreas que no estaban contempladas en investigaciones teóricas previas. Basándose en estos hallazgos, se destacan las direcciones para futuras investigaciones de ML con el objetivo de conseguir abordar mejor las necesidades de los profesionales de la industria.

Co-diseño de listas de verificación para comprender los desafíos y oportunidades organizacionales en torno a la equidad en la IA (1)

Muchas organizaciones han publicado principios destinados a guiar el desarrollo ético y la implementación correcta de sistemas de IA. Sin embargo, su naturaleza abstracta dificulta su operatividad. A menos que las listas de verificación se basen en las necesidades de los profesionales en cuestión, es posible que se utilicen de manera incorrecta. Es por lo que, para comprender el papel de las listas de verificación en la ética de la IA, Microsoft ha llevado a cabo un proceso de co-diseño iterativo con cuarenta y ocho profesionales, centrándose en la equidad.

Han creado una lista de verificación para la comprobación de la imparcialidad de la IA y se han identificado los intereses y preocupaciones de estas listas en general. También, se han discutido aspectos de la cultura organizacional que pueden afectar la eficacia de tales listas de verificación y se mencionan las direcciones de investigación futuras.

(1) https://www.microsoft.com/en-us/research/publication/co-designing-checklists-to-understand-organizational-challenges-and-opportunities-around-fairness-in-ai/

Recomendaciones

Además, desde Microsoft se recomienda tener en cuenta las siguientes cuestiones a la hora de la consecución de unos sistemas de inteligencia artificial justos y equitativos:

Equipos diversos

Es importante reunir un equipo de desarrolladores que tenga una diversidad inherente. Esta recomendación no hace referencia a crear un equipo únicamente demográficamente diverso, sino también a incluir miembros con diferentes experiencias, perspectivas, formación, etc.

Entender el alcance del sistema

- ¿Cómo está previsto que funcione el sistema?
- ¿Para quién está diseñado el sistema?
- ¿Funcionará para todos por igual?
- ¿Podría alguien resultar perjudicado?

Identificar sesgos en los conjuntos de datos

- ¿De dónde provienen los datos?
- ¿Cómo se etiquetaron (1)?
- ¿Se están introduciendo sesgos en los datos al etiquetarlos?
- ¿Cómo se puede compensar el sesgo detectado?
- ¿Son los datos representativos de la población?

Identificar sesgos en el modelo de aprendizaje automático

Los sesgos en los algoritmos ocurren cuando el sistema realiza sus suposiciones a través de un conjunto de datos erróneos que no describen de manera completa la realidad, bien porque no son completos o, simplemente, porque no son los correctos. Por ejemplo, en el caso de un sistema de reconocimiento facial que haya sido entrenado en su totalidad con imágenes de personas de raza blanca, a la hora de hacer sus predicciones discriminará a las personas de raza negra.

(1) Etiquetado o 'labeling': proceso de identificar datos sin procesar y agregar una o más etiquetas significativas e informativas para proporcionar contexto y permitir el entrenamiento del modelo.

Cinco formas de identificar el sesgo

Microsoft ha trabajado con líderes tanto intelectuales como académicos en busca de formas de identificar el sesgo y problemas con la inclusión, determinando cinco tipos principales de sesgo:

- **Sesgo por el conjunto de datos**: Cuando el conjunto de datos empleado para entrenar el algoritmo de *machine learning* no representa la diversidad actual del conjunto de las personas que se servirán del mismo. El objetivo es evitar la construcción de conjuntos de datos utilizando menos información de la necesaria o siendo esta excesivamente homogénea. Por ejemplo, las tecnologías de visión por computador entrenadas con etnias predominantemente blancas que tienen un mayor porcentaje de error con personas de piel oscura (1).
- **Sesgo por asociación**: Cuando los datos utilizados para entrenar un modelo refuerzan o multiplican un sesgo cultural. Los sesgos humanos pueden verse reflejados en la IA y afectar a la experiencia del cliente. Por ejemplo, herramientas de traducción de textos que asumen el género de la persona que las utiliza, como que los pilotos son hombres y el personal de cabina mujeres. En 2016 se descubrió que algunos de los algoritmos de LinkedIn tenían un sesgo por género, que recomendaba empleos mejor pagados a hombres en vez de a mujeres. Esto se veía reforzado porque, en la actualidad, los puestos de elevada remuneración están predominantemente ocupados por hombres (2).
- **Sesgo por automatización**: Cuando decisiones automatizadas sobrescriben consideraciones sociales y culturales. Los desarrolladores de IA deben tener en cuenta los objetivos de las personas a las que afecta el sistema que van a construir. Por ejemplo, cuando los filtros de belleza de redes sociales como Instagram refuerzan la noción de belleza europea, aclarando el tono de piel o haciendo más redondos los ojos, llegando a causar problemas de autoestima a personas con otras características (3).
- **Sesgo por interacción**: Cuando los humanos manipulan una IA existente y generan resultados sesgados. Los *Chatbots* de hoy pueden hacer bromas y engañar a la gente para que piense que son humanos, pero muchos intentos de humanizar la inteligencia artificial han contaminado involuntariamente a la misma con sesgos humanos tóxicos. Este tipo de sesgo se dará en IAs que tengan un aprendizaje continuo durante su interacción con humanos, dándose la posibilidad de que en dichas interacciones parte de los sesgos de estas personas pasen a la IA. Por ello es importante diseñar con salvaguardas contra la toxicidad, como una lista de palabras que nunca utilizar. Por ejemplo, personas que enseñen a un *Chatbot* palabras malsonantes o insultos raciales de forma deliberada.
- **Sesgo por confirmación**: Cuando una personalización demasiado simplificada asume ciertos factores de un grupo o un individuo. El sesgo de confirmación interpreta la información de una manera que confirma los prejuicios. Los algoritmos de IA devuelven contenido que ha sido elegido por otras personas, lo cual excluye los resultados de las personas que hicieron elecciones menos populares. Una persona que solo recibe *feedback* de otras personas que piensan de forma similar nunca encontrará puntos de vista opuestos y será mucho más difícil que vea caminos alternativos e ideas diferentes. Por ejemplo, anuncios de Facebook u otras páginas web que recomiendan artículos que ya hemos comprado, por haber hecho recientemente una búsqueda intensiva de los mismos intentando encontrar la mejor opción.

(1) En 2015, un desarrollador de software afroamericano advirtió a Google de que un colega suyo y él habían sido etiquetados como “gorilas” por su algoritmo.
https://www.theverge.com/2018/1/12/16882408/google-racist-gorillas-photo-recognition-algorithm-ai

(2) Noticia: https://www.seattletimes.com/business/microsoft/how-linkedins-search-engine-may-reflect-a-bias/

(3) https://www.nylon.com/beauty/instagrams-beauty-filters-perpetuate-the-industrys-ongoing-racism

El enfoque de Microsoft

FairLearn, tecnología para la valuación y mitigación de parcialidad

Fairlearn (1) es un paquete de *Python* de código abierto publicado en *GitHub* (2) que permite a los desarrolladores de sistemas de aprendizaje automático evaluar la equidad de sus sistemas y mitigar la parcialidad.

La imparcialidad es un desafío técnico y social. Muchos aspectos de la imparcialidad no se capturan en métricas cuantitativas de imparcialidad. Asimismo, muchas métricas cuantitativas de imparcialidad no se pueden satisfacer simultáneamente. El objetivo del paquete de código abierto de *Fairlearn* es permitir a los usuarios evaluar las distintas estrategias de impacto y mitigación.

El paquete de código abierto de Fairlearn tiene dos componentes:

- **Panel de evaluación**: es un *widget* de *Jupyter Notebook* para evaluar cómo afectan las predicciones de un modelo a grupos diferentes. También permite comparar varios modelos mediante métricas de imparcialidad y rendimiento.
- **Algoritmos de mitigación**: es un conjunto de algoritmos para mitigar la imparcialidad en la clasificación y la regresión binarias.

Juntos, estos componentes permiten a los científicos de datos y a los líderes empresariales administrar cualquier intercambio entre equidad y rendimiento, y seleccionar la estrategia de mitigación que mejor se adapte a sus necesidades.

En el paquete de código abierto de *Fairlearn*, la equidad se conceptualiza a través de un enfoque conocido como equidad de grupo, que pregunta: ¿Qué grupos de usuarios están en riesgo de experimentar daños? Los grupos pertinentes, también conocidos como subpoblaciónes, se definen a través de características o atributos confidenciales. Las características confidenciales se pasan a un estimador en el paquete de código abierto de *Fairlearn* como un vector o una matriz denominada *sensitive features*. El término sugiere que el diseñador del sistema debe mantener la confidencialidad de estas características al evaluar la imparcialidad del grupo.

Algo que se debe tener en cuenta es si estas características contienen implicaciones de privacidad debido a datos privados. De todos modos, la palabra "confidencial" no implica que estas características no se utilicen para realizar predicciones.

(1) https://fairlearn.org/
(2) https://github.com/fairlearn/fairlearn/

Métricas para detectar la parcialidad

Una evaluación de equidad no es un ejercicio puramente técnico. *Fairlearn* puede ayudar a evaluar la equidad de un modelo, pero no llevará a cabo la evaluación de manera automática. *Fairlearn* ayuda a identificar métricas cuantitativas para evaluar la equidad, pero los desarrolladores también deben realizar un análisis cualitativo para evaluar la equidad de sus propios modelos. Las características confidenciales indicadas anteriormente son un ejemplo de este tipo de análisis cualitativo.

La imparcialidad se cuantifica a través de las métricas de disparidad. Las métricas de disparidad pueden evaluar y comparar el comportamiento del modelo en grupos diferentes, ya sea como proporciones o diferencias. *Fairlearn* admite **dos clases de métricas de disparidad**:

- **Disparidad en el rendimiento del modelo**: estos conjuntos de métricas calculan la disparidad (diferencia) en los valores de la métrica de rendimiento seleccionada en los distintos subgrupos. Estos son algunos ejemplos:
  - Disparidad en la tasa de precisión.
  - Disparidad en la tasa de error.
  - Disparidad en la precisión.
  - Disparidad en la coincidencia.
  - Disparidad en la métrica MAE.

- **Disparidad en la tasa de selección**: esta métrica contiene la diferencia de la tasa de selección entre distintos subgrupos. Un ejemplo de esto es la disparidad en la tasa de aprobación de préstamos. La tasa de selección indica la fracción de los puntos de referencia de cada clase clasificada como 1 (en la clasificación binaria) o la distribución de los valores de predicción (en la regresión).

Algoritmos para mitigar la parcialidad

Fairlearn incluye una variedad de algoritmos de mitigación de la parcialidad. Estos algoritmos admiten un conjunto de restricciones en el comportamiento del predictor denominadas **restricciones o criterios de paridad**. Las restricciones de paridad requieren que algunos aspectos del comportamiento de la predicción sean comparables en los grupos que definen las características confidenciales (por ejemplo, razas diferentes). Los algoritmos de mitigación en *Fairlearn* usan estas restricciones de paridad para mitigar los problemas de equidad observados.

Los algoritmos de mitigación de la parcialidad en *Fairlearn* proporcionan estrategias de mitigación sugeridas para reducir la parcialidad en un modelo de aprendizaje automático, pero no son soluciones para eliminar la parcialidad por completo. Puede que haya otras restricciones o criterios de paridad que se deben tener en cuenta para cada modelo de aprendizaje automático. Los desarrolladores y responsables de la solución deben determinar por sí mismos si la mitigación elimina suficientemente cualquier injusticia en el uso previsto.

Fairlearn admite los siguientes tipos de restricciones de paridad:

- **Paridad demográfica**. Su propósito es mitigar los daños en la asignación, y aplica a clasificaciones binarias y regresión.
- **Probabilidades igualadas e igualdad de oportunidades**. Su objetivo es diagnosticar los daños en la asignación y la calidad del servicio, y aplica a clasificaciones binarias.
- **Pérdida de grupos limitada**. Su función es mitigar los daños en la calidad del servicio, y aplica a regresiones.

Y estos son los tipos de algoritmos de mitigación de la parcialidad:

- **Reducción**. Estos algoritmos toman un estimador de aprendizaje automático estándar de caja negra (por ejemplo, un modelo de *LightGBM*) y generan un conjunto de modelos que se vuelven a entrenar con una secuencia de conjuntos de datos de entrenamiento que se han vuelto a ponderar. Por ejemplo, los solicitantes de un sexo determinado podrían estar ponderados en mayor o menor medida para volver a entrenar los modelos y reducir así las disparidades entre los diferentes sexos. A continuación, los usuarios pueden elegir un modelo que les ofrezca el mejor equilibrio entre la precisión (u otra métrica de rendimiento) y la disparidad, lo que generalmente debería basarse en reglas de negocios y cálculos de costos.

Algoritmo	Descripción	Uso en	Características confidenciales	Restricciones de paridad
ExponentiatedGradient	Enfoque de la caja negra para la clasificación imparcial	Clasificación binaria	Categorías	Paridad demográfica, Probabilidades igualadas
GridSearch	Enfoque de caja negra para la clasificación imparcial	Clasificación binaria	*Binary*	Paridad demográfica, Probabilidades igualadas
GridSearch	Enfoque de la caja negra que implementa una variante de la búsqueda de cuadrícula de la regresión imparcial con el algoritmo, para la pérdida de grupos limitada.	Regresión	*Binary*	Pérdida de grupos limitada

Postprocesamiento: estos algoritmos usan un clasificador existente y la característica confidencial como entrada. A continuación, derivan una transformación de la predicción del clasificador para aplicar las restricciones de imparcialidad especificadas. La mayor ventaja de la optimización del umbral es su simplicidad y flexibilidad, ya que no es necesario volver a entrenar el modelo.

Algoritmo	Descripción	Uso en	Características confidenciales	Restricciones de paridad
ThresholdOptimizer	Esta técnica toma como entrada un clasificador existente y la característica confidencial, y deriva una transformación monótona de la predicción del clasificador para aplicar las restricciones de paridad especificadas.	Clasificación	Categorías	Paridad demográfica, Probabilidades igualadas

AI Fairness Checklist, herramienta para la gestión de la Equidad

Muchas organizaciones han publicado principios para guiar el desarrollo y la implementación responsable de los sistemas de IA, pero en gran medida se dejan en manos de los profesionales para ponerlos en práctica. Por lo tanto, otras organizaciones han producido listas de verificación de ética de IA, incluidas listas de verificación para conceptos específicos, como la equidad.

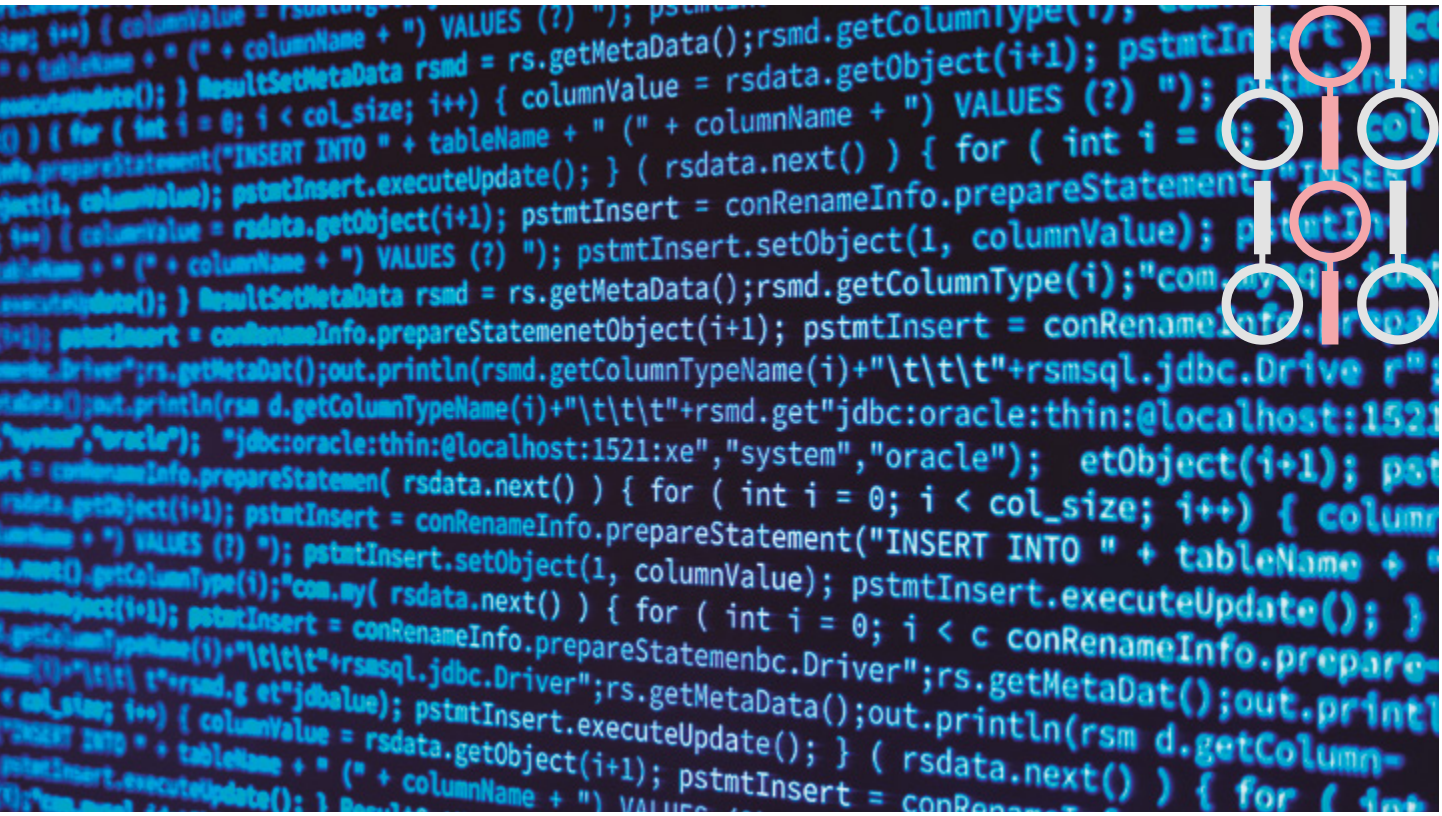
Las listas de verificación en otros dominios, como la aviación, la medicina y la ingeniería estructural, han tenido un éxito bien documentado en salvar vidas y mejorar las prácticas profesionales. Pero a menos que las listas de verificación se basen en las necesidades de los profesionales, pueden ser mal utilizadas o ignoradas.

El proyecto de investigación *AI Fairness Checklist* (1) explora cómo se pueden diseñar las listas de verificación para respaldar el desarrollo de productos y servicios de IA más justos.

La *checklist* incluye una serie de recomendaciones destinadas a ser utilizadas como punto de partida para que sea personalizada en función del caso de uso. Por tanto, no todos los apartados son aplicables a todos los sistemas de IA, y es probable que en cada caso se necesite agregar, revisar o eliminar elementos para que se ajusten mejor a sus necesidades específicas.

Las recomendaciones están agrupadas en un ciclo de vida que, tal y como hemos visto en otros momentos de este apartado, Microsoft divide en fases definidas como *Envision - Define- Prototype - Launch y Evolve*.

(1) https://www.microsoft.com/en-us/research/project/ai-fairness-checklist/



Principio de Confiabilidad y seguridad

En esta sección vamos a desarrollar:

- Qué es la Confiabilidad y la seguridad para Microsoft.
- Recomendaciones en el ámbito de Confiabilidad y seguridad.
- Tecnologías para aplicar la transparencia
  - *Error Analysis*, para identificar y diagnosticar errores.
  - *Counterfit*, para la evaluación de riesgos de seguridad.

Qué es la Confiabilidad y la seguridad para Microsoft

La Confiabilidad y la seguridad son unos de los principios más importantes para Microsoft a la hora de diseñar sistemas de inteligencia artificial (IA). Cuando Microsoft diseña un sistema de IA se asegura que estén enfocados de manera acorde a sus principios éticos y morales. Para ello, los sistemas que son diseñados por Microsoft no deben crear ningún tipo de daño al mundo, pero esto no es una situación real, ya que es imposible evitar que en el 100% de los casos los sistemas de IA no tengan ningún defecto. Por lo tanto, Microsoft se encarga de que todos los usuarios que vayan a utilizar sus productos sean conscientes de los riesgos que ello implica, por muy pequeños que sean. Este es el principio fundamental que Microsoft tiene respecto a la Confiabilidad y seguridad, y es un concepto que se aplica en cada uno de los productos relacionados con la inteligencia artificial.

Cuando hablamos de seguridad, uno de los primeros ejemplos que se viene a la cabeza son los coches inteligentes, los cuales son capaces de conducir de manera autónoma, pero también existen otro tipo de sistemas de IA que tienen una gran influencia en nuestras vidas, como por ejemplo aquellos que se emplean en los hospitales y tienen un impacto directo sobre la salud de las personas, ya que son capaces de realizar predicciones sobre diagnósticos. En estos casos, un error en los sistemas se puede transformar en un daño considerable hacia la persona. Por esta razón, Microsoft tiene puesto el foco en este tipo de situaciones.

Para generar Confiabilidad, es crucial que los sistemas de inteligencia artificial funcionen de forma fiable, segura y coherente tanto en circunstancias normales como en situaciones inesperadas. Estos sistemas deben ser capaces de funcionar según su diseño original, responder de manera segura a situaciones inesperadas y resistir los posibles ataques de usuarios maliciosos con fines perjudiciales. También es importante poder comprobar que estos sistemas se comportan según lo previsto en condiciones de funcionamiento reales, es decir, efectuando pruebas en entornos que simulen situaciones del mundo real. La manera en la que se comportan y la variedad de situaciones que un sistema de IA puede controlar de forma fiable y segura, reflejan, en gran medida, la gama de situaciones y circunstancias que los desarrolladores han sido capaces de anticipar y preparar durante la etapa de diseño y el desarrollo de las pruebas. Si no se mantienen correctamente, los sistemas de inteligencia artificial pueden convertirse en poco fiables o imprecisos, por lo que es fundamental tener en cuenta las operaciones a largo plazo y la supervisión en cada implementación de inteligencia artificial. El juicio humano es la clave para identificar posibles puntos débiles y sesgos en los sistemas de inteligencia artificial.

La Confiabilidad es un principio que se debe aplicar a todos los sistemas de IA y es necesario para crear sistemas seguros. Un sistema se considera fiable cuando funciona de manera coherente y según lo previsto, no solo en las condiciones del laboratorio en las que se entrena, sino también en las condiciones del mundo real o cuando son atacados por adversarios. Cuando un sistema no es fiable sus deficiencias pueden suponer riesgos para las vidas humanas, en este tipo de casos los problemas de Confiabilidad se traducen en riesgos para la seguridad.

El enfoque de Microsoft

Para entender cómo se producen los problemas de seguridad y Confiabilidad de los sistemas de IA, los investigadores de Microsoft han estudiado:

- **Puntos débiles de los *datasets*** (1) : Los modelos predictivos basados en IA desplegados en el mundo real pueden asignar etiquetas incorrectas a instancias, con un intervalo de confianza alto. Este tipo de errores suelen surgir debido a un desajuste entre los datos de entrenamiento y los datos captados en el mundo real.
- **Desajustes entre los entornos de entrenamiento y los entornos de ejecución** (2) : Los sistemas de IA entrenados en entornos de simulación pueden cometer errores en el mundo real debido a los desajustes entre los entornos de entrenamiento y de ejecución. Este tipo de errores son peligrosos y difíciles de descubrir, ya que el propio sistema es incapaz de identificarlos. Los investigadores de Microsoft proponen, para abordar este problema, el uso del Aprendizaje por Refuerzo, conocido como *Reinforced Learning* (3) , que se caracteriza por premiar al sistema cuando toma las decisiones correctas.
- **Problemas en las especificaciones**: Los errores en las especificaciones de los entrenamientos, pruebas, etc., pueden provocar errores en los sistemas de IA.

Debido a la diversidad de los orígenes de los fallos, la clave para garantizar la Confiabilidad de los sistemas es una evaluación rigurosa durante el desarrollo y la implantación de estos, de esta forma se podrán minimizar los fallos de rendimiento inesperados y se podrá guiar a los desarrolladores de los sistemas para que los vayan mejorando de manera continua. Por esta razón, los investigadores de Microsoft han desarrollado nuevas técnicas de depuración de modelos y análisis de los errores (4) que son capaces de revelar patrones en el modo en el que estos errores se presentan. Los esfuerzos que Microsoft está realizando en este ámbito incluyen convertir todas estas investigaciones en herramientas que puedan ser utilizadas por los desarrolladores.

Microsoft reconoce que cuando se emplean los sistemas de IA en aplicaciones críticas para nuestra sociedad, por ejemplo, los coches autónomos o los sistemas empleados en hospitales para realizar diagnósticos, la precisión de estos sistemas no puede ser la única métrica para cuantificar el rendimiento de las máquinas. Los investigadores han demostrado que las actualizaciones de los modelos pueden dar lugar a problemas de retrocompatibilidad (5), es decir, que se produzcan nuevos errores debido a la incorporación de una nueva actualización en el sistema. En este tipo de casos, se puede dar la situación en la que a pesar de que aparezcan los problemas de retrocompatibilidad, la precisión del sistema aumente, lo que pone de manifiesto que el rendimiento de los modelos no debe basarse tan solo en la precisión del sistema, sino que debe tener en cuenta más consideraciones que estén centradas en el ser humano.

(1) https://www.microsoft.com/en-us/research/wp-content/uploads/2017/02/unknown_unknowns_identify_algo.pdf

(2) https://arxiv.org/pdf/1805.08966.pdf

(3) https://arxiv.org/pdf/cs/9605103.pdf

(4) https://www.microsoft.com/en-us/research/uploads/prod/2018/07/accountable_AI_hcomp_2018.pdf

(5) https://www.microsoft.com/en-us/research/blog/creating-better-ai-partners-a-case-for-backward-compatibility/

Recomendaciones en el ámbito de Confiabilidad y seguridad

Con el objetivo de cumplir con los principios éticos del área de la Confiabilidad y seguridad, Microsoft ha propuesto una serie de recomendaciones:

- Comprender la madurez de la inteligencia artificial de la organización. Identificar qué tecnologías de inteligencia artificial se adaptan al nivel de madurez actual de la organización.
- Desarrollar procesos de auditoría de los sistemas de inteligencia artificial. Se deberá evaluar la calidad y la idoneidad de los datos y los modelos, supervisar el rendimiento continuo y comprobar que los sistemas se comportan según lo previsto en función de las medidas de rendimiento establecidas. Además, habrá que tener en cuenta aspectos como la trazabilidad de los datos, que consiste en conocer todas las modificaciones que han sufrido los datos desde que han sido recogidos hasta que se han introducido al sistema de IA, y utilizar marcos de trabajo de gobernanza para administrar, almacenar y garantizar la seguridad de los datos.
- Proporcionar una explicación detallada del funcionamiento del sistema. Se deberá incluir información sobre el diseño del sistema, los datos usados en la etapa de entrenamiento, las posibles deficiencias que puedan tener los datos, y los errores de aprendizaje detectados en las diferentes pruebas realizadas.
- Realizar diseños para situaciones imprevistas. Con el objetivo de adelantarse a los usuarios maliciosos que quieran atacar nuestro sistema, los desarrolladores se deberán anticipar a estos ataques. Para ello podrán generar interacciones accidentales que provoquen fallos en el sistema, incorporar datos malintencionados o simular ciberataques.
- Incorporar a expertos en el dominio de los procesos de diseño e implementación.
- Ejecutar pruebas durante el desarrollo y la implementación del sistema IA. Los sistemas deberán responder de forma segura ante situaciones imprevistas y no evolucionar de forma inesperada. Los sistemas de inteligencia artificial involucrados en escenarios de gran importancia que afectan a la seguridad humana o a grandes poblaciones deben probarse tanto en entornos de laboratorio como del mundo real.
- Evaluar cuándo y cómo un sistema de inteligencia artificial debe buscar aportación humana para decisiones de alto impacto o durante situaciones críticas.
- Desarrollar un mecanismo de comentarios sólido. Proporcionando de esta manera la posibilidad de que los usuarios notifiquen los problemas de rendimiento y que estos puedan resolverse rápidamente.

Tecnologías para aplicar la Confiabilidad y seguridad

Microsoft dispone de dos herramientas para la gestión automatizada de la Confiabilidad y seguridad:

- *Error Analysis*, para identificar y diagnosticar errores.
- *Counterfit*, para la evaluación de riesgos de seguridad.

Error Analysis, tecnología para identificar y diagnosticar errores

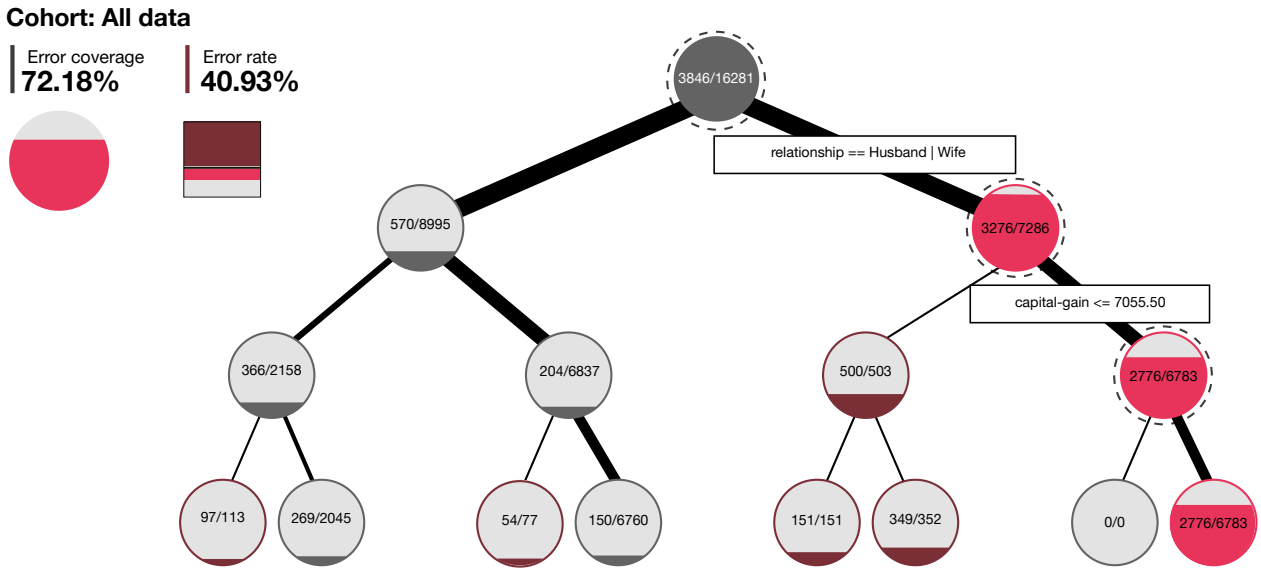
- *Error Analysis* (1) es un conjunto de herramientas de IA responsable que permite obtener una comprensión más profunda de los errores del modelo de aprendizaje automático. Al evaluar un modelo de aprendizaje automático, la precisión agregada no es suficiente puede ocultar importantes inexactitudes. *Error Analysis* permite identificar muestras de datos con tasas de error altas y diagnosticar las causas fundamentales de estos errores. En resumen, permite:
- Evaluar muestras, mostrando cómo se distribuyen los errores entre diferentes muestras y diferentes niveles de granularidad.
- Explorar predicciones, utilizando funciones de interpretación integradas o combinándolas con *InterpretML* para mejorar aún más el detalle obtenido.
- Visualizar de manera sencilla los errores y diagnosticar el origen de los mismos.

(1) https://erroranalysis.ai/

Identificación de errores

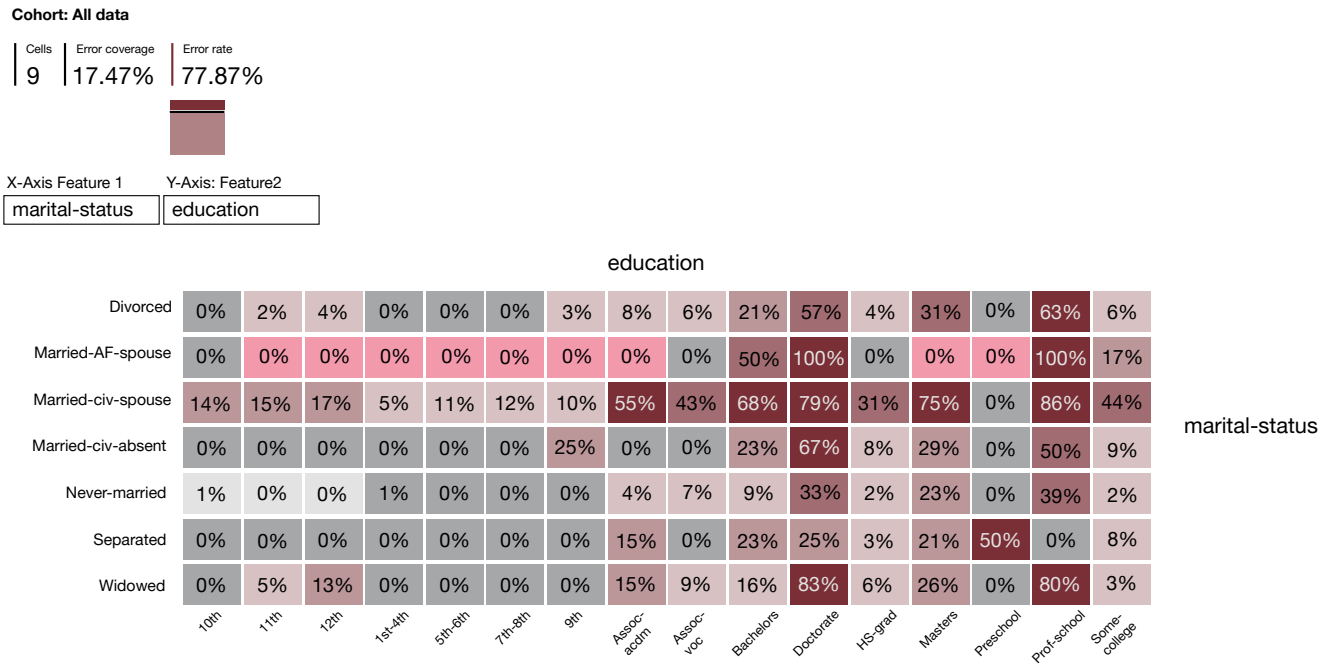
Identifica conjuntos de datos con una tasa de error más alta que la referencia general. Estas discrepancias pueden ocurrir cuando el sistema o modelo tiene un rendimiento inferior con grupos demográficos específicos o con condiciones de entrada observadas que han sido poco contempladas en los datos de entrenamiento. Utiliza diferentes métodos para la identificación de los errores:

- **Mediante árboles de decisión**, permite descubrir muestras con altos índices de error utilizando una visualización de árbol binario.



Fuente: Microsoft

Mediante mapas de calor de errores, que permiten investigar más a fondo cómo una o dos características de entrada afectan la tasa de error en las muestras.



Fuente: Microsoft

Diagnóstico de errores

Después de identificar muestras con altas tasas de error, el análisis de errores permite comprender las razones detrás de las tasas de error para poder tomar medidas correctivas. Permite utilizar diferentes métodos para el diagnóstico de errores:

- **Exploración de datos**, permitiendo explorar estadísticas del conjuntos de datos, comparar estadísticas de la muestra con otras muestras o con los valores de referencia, investigando si ciertas muestras están subrepresentadas o si su distribución de características es significativamente diferente de los datos generales.
- **Explicación global**, permitiendo explorar las características más importantes que impactan en la explicación global del modelo de una muestra seleccionada, comprendiendo cómo los valores de las características afectan las predicciones del modelo, y comparando las explicaciones de las muestras seleccionadas con otras muestras o puntos de referencia.
- **Explicación local**, permitiendo analizar datos individuales de la muestra seleccionada, divididos por predicciones correctas o incorrectas, pudiendo examinar visualmente los potenciales ruidos causantes de predicciones incorrectas. Además, permite comprender qué características tienen el mayor impacto en las predicciones.
- ***What-if Analysis***, permite aplicar cambios en las entradas para observar los cambios resultantes en la predicción.

Counterfit, tecnología para la evaluación de riesgos de seguridad

Counterfit es una herramienta open *source* que permite automatizar pruebas de seguridad de soluciones inteligentes. *Counterfit* ayuda a las organizaciones a realizar evaluaciones de riesgos de seguridad de IA para garantizar que los algoritmos utilizados en sus negocios sean sólidos y confiables.

Los sistemas de IA se utilizan cada vez más en áreas críticas de todo tipo. Los consumidores deben tener confianza en que los sistemas de IA son fiables y están protegidos contra la manipulación por parte de adversarios. Por ejemplo, una de las recomendaciones de las 5 prioridades principales de *Gartner* para gestionar el riesgo de IA dentro del marco *MOST* de *Gartner* publicado en enero de 2021 (1) es que las organizaciones “adopten medidas de seguridad de IA específicas contra los ataques de los adversarios para garantizar la resistencia y la resiliencia”, señalando que “para 2024, Las organizaciones que implementan controles de gestión de riesgos de IA dedicados evitarán con éxito resultados negativos de IA con el doble de frecuencia que aquellas que no lo hacen”.

Realizar evaluaciones de seguridad de los sistemas de IA no es trivial. Microsoft encuestó a 28 organizaciones, que abarcan empresas de *Fortune* 500, gobiernos, organizaciones sin fines de lucro y pequeñas y medianas empresas (PYMES), para comprender los procesos actuales implementados para proteger los sistemas de IA. Descubrieron que 25 de 28 empresas indicaron que no cuentan con las herramientas adecuadas para proteger sus sistemas de inteligencia artificial y que los profesionales de la seguridad buscan orientación específica en este espacio, y lo documentaron en un extenso documento (2).

Counterfit nació de la propia necesidad interna de Microsoft de evaluar sus sistemas de inteligencia artificial en busca de vulnerabilidades con el objetivo de proteger de manera proactiva sus servicios de inteligencia artificial, de acuerdo con los principios éticos de inteligencia artificial responsable y la iniciativa Estrategia de inteligencia artificial responsable en ingeniería (RAISE) de Microsoft. *Counterfit* comenzó como un corpus de *scripts* de ataque escritos específicamente para apuntar a modelos de IA individuales y luego se transformó en una herramienta de automatización genérica para atacar múltiples sistemas de IA a escala.

Para garantizar que *Counterfit* aborda un conjunto amplio de necesidades de seguridad, Microsoft está contando con grandes organizaciones, pymes y organizaciones gubernamentales para utilizar la herramienta con sus modelos de ML.

(1) https://blogs.gartner.com/avivah-litan/2021/01/21/top-5-priorities-for-managing-ai-risk-within-gartners-most-framework/

(2) https://arxiv.org/pdf/2002.05646.pdf

El enfoque de Microsoft

Independiente de infraestructura, tipo de modelo y tipos de datos

Counterfit es flexible, permitiendo adaptarse en tres dimensiones fundamentales:

- **Es independiente del entorno**: puede ayudar a evaluar los modelos de IA alojados en cualquier entorno de nube, en las instalaciones o en el perímetro.
- **Es independiente del modelo**: la herramienta abstrae el funcionamiento interno de sus modelos de IA para que los profesionales de la seguridad puedan concentrarse en la evaluación de la seguridad.
- **Es independiente de los datos**: funciona en modelos de IA que utilizan texto, imágenes o entrada genérica.

Counterfit es una herramienta de línea de comandos que proporciona una capa de automatización genérica para marcos de trabajo de seguridad en IA como *Adversarial Robustness Toolbox y TextAttack* . La herramienta dispone de algoritmos de ataque accesibles para la comunidad opensource y ayuda a proporcionar una interfaz extensible desde la cual construir, administrar y lanzar ataques a modelos de IA.

Microsoft recomienda usar *Counterfit* junto con *Adversarial ML Threat Matrix* (ATLAS) (1), que es un marco de estilo ATT&CK lanzado por MITRE (2) y Microsoft para gestionar las amenazas contra los sistemas de IA.

(1) https://github.com/mitre/advmlthreatmatrix
(2) https://www.mitre.org/

Principio de Privacidad y seguridad

En esta sección vamos a desarrollar:

- Qué es la privacidad y la seguridad para Microsoft
- Recomendaciones en el ámbito de privacidad y seguridad
- Tecnologías para aplicar la la privacidad y seguridad
  - Presidio, para la protección y anonimización del dato
  - *Smart Noise*, para la Privacidad Diferencial
  - *SEAL Homomorphic* Encryption, para cálculos sobre datos cifrados en *cloud*.
  - Confidencial *computing for ML*, infraestructura *cloud* privada y segura.

Qué es la privacidad y la seguridad para Microsoft

La privacidad es un derecho fundamental y Microsoft tiene un gran compromiso con ella y con la seguridad de los datos que se usan para desarrollar sus productos y sistemas. A partir de la aparición de la inteligencia artificial (IA), el *Machine Learning* (ML) y el *Deep Learning* (DL), ha habido un incremento tanto en la complejidad de los sistemas de IA que Microsoft desarrolla, como en la dependencia del uso de datos para desarrollar estos sistemas de IA, debido a que el acceso a los datos es esencial para que los sistemas de IA realicen predicciones y tomen decisiones precisas y fundamentadas.

Como consecuencia de que parte de estas predicciones y decisiones que toman los sistemas de IA tienen influencia sobre las personas, los problemas de seguridad y privacidad son de una gran relevancia. Es destacable también la facilidad que existe para engañar a los sistemas de IA mediante el envenenamiento de los datos de entrada, es decir, la modificación de los datos que se introducen al sistema para realizar las predicciones. Este envenenamiento se puede realizar de diferentes modos, como por ejemplo añadiendo un filtro de color a las imágenes que se introducen en un detector de imágenes, añadiendo ruido en los audios que utiliza un reconocedor de audio, etc. Estos diferentes tipos de envenenamientos vienen recogidos en el documento ***Failure* Modes *in Machine Learning***.

Los principales objetivos que tienen los usuarios maliciosos que se encargan del envenenamiento de los datos de nuestro sistema son:

1. **Atacar la disponibilidad del modelo**: Mediante la modificación de los datos del sistema, tanto la efectividad como la precisión de este disminuirá considerablemente provocando que el modelo quede inutilizable.

2. **Atacar la integridad del modelo**: Consiste en modificar la forma en la que aprende nuestro sistema a partir del envenenamiento de los datos de entrenamiento generando de esta forma que nuestro sistema tome decisiones erróneas. Un posible escenario de este tipo de ataque se podría dar con un sistema de clasificación de imágenes, el cual sea capaz de diferenciar entre dos tipos de clases, perros y peces. Al envenenar los datos de entrenamiento se podría provocar que el clasificador dejará de funcionar correctamente y que siempre clasifique las imágenes como peces, independientemente de que tipo de imagen analice nuestro clasificador (1).

Toda esta situación ha provocado que se necesiten nuevos requisitos para conseguir que los sistemas de IA sean seguros. Además, los sistemas de IA deben cumplir las leyes de privacidad, como por ejemplo GDPR (2), LOPD (3), que exigen transparencia sobre la recopilación, el uso y el almacenamiento de datos y obligan a los consumidores a establecer controles adecuados para elegir cómo se usan sus datos.

(1) https://arxiv.org/abs/1804.00792
(2) https://gdpr-info.eu/
(3) https://www.boe.es/buscar/pdf/2018/BOE-A-2018-16673-consolidado.pdf

El enfoque de Microsoft

Recomendaciones en el ámbito de privacidad y seguridad

A continuación, vamos a revisar un conjunto de recomendaciones para cumplir una serie de principios éticos del área de la privacidad y la seguridad propuestos por Microsoft.

- **Cumplir las leyes de protección privacidad y transparencia de los datos relevantes (GDPR, LOPD)**. Se deberán desarrollar procesos para comprobar de forma continua que los sistemas de inteligencia artificial satisfacen todos los aspectos de estas leyes.
- **Diseñar sistemas de IA para mantener la integridad de los datos personales**. Los datos personales sólo se usarán durante el tiempo que sea necesario y para los fines definidos con los clientes. Además, los datos personales que hayan sido recopilados de forma involuntaria serán eliminados.
- **Proteger los sistemas IA de usuarios no válidos**. Se deberán desarrollar sistemas IA mediante acceso basado en roles (RBAC (1)), que es un sistema de autorización que ayuda a administrar quién tiene acceso a los recursos de *Azure*, qué pueden hacer con esos recursos y a qué áreas y servicios pueden acceder los diferentes usuarios. Además, los sistemas de IA deberán ser capaces de identificar comportamientos anómalos para evitar la manipulación de los datos y los ataques malintencionados.
- **Diseñar sistemas de IA con controles adecuados**. Se les brindará a los clientes la oportunidad de que tomen decisiones sobre cómo y para qué se recopilan y se van a usar sus datos.
- **Garantizar que el sistema de IA mantiene el anonimato**. Se deberá suprimir la identificación de los datos personales.
- **Realizar revisiones de seguridad y privacidad de todos los sistemas IA**.
- **Investigar e implementar procedimientos recomendados del sector**.

Por otra parte, según afirma *Gartner* en el artículo *Market Guide for AI Trust, Risk and Security Management (2)*, "los sistemas de IA plantean nuevos requisitos para la gestión de la confianza, el riesgo y la seguridad que los sistemas tradicionales no abordan". Sin embargo, pese a esta carencia, Microsoft ha decidido no crear un sistema de defensa nuevo. Un sistema de defensa, dentro del ámbito de la IA y el ML, son el conjunto de procesos y sistemas que se encargan de proteger a otros sistemas que utilizan la IA, por ejemplo, evitando la manipulación de los datos que dichos sistemas utilizan, denegando el acceso a usuarios maliciosos, etc. Lo que proponen consiste en modificar los sistemas de defensa actuales de manera que se puedan adaptar a los nuevos tipos de riesgos que los sistemas de IA generan. Para ello, ha diseñado su propia evaluación de riesgos de seguridad de la IA siguiendo los principios que se utilizan en los marcos de evaluación de riesgos de seguridad ya existentes para los sistemas tradicionales (3) (4); aquellos que no utilizan la IA.

Microsoft tiene la creencia de que, para evaluar de manera exhaustiva el riesgo de seguridad de un sistema de IA, es necesario examinar el ciclo de vida del desarrollo y la implantación de dicho sistema. Además, Microsoft define la tarea de asegurar los sistemas de IA como una tarea en equipo, en el que cada persona o grupo de trabajo desempeña una labor diferente. Los investigadores de IA diseñan la arquitectura de los modelos; los ingenieros de ML se encargan de la recepción de los datos, el entrenamiento del sistema y la generación de un *Deployment Pipeline*, que se define como la cadena de despliegue de un sistema IA, la cual puede resultar muy útil a la hora de identificar los puntos de nuestro sistema que son más vulnerables a un posible ataque y asegurarlos; los arquitectos de seguridad establecen las políticas de seguridad; y por último, los analistas de seguridad se encargan de responder a las diferentes amenazas. Con esta información el marco de trabajo desarrollado por Microsoft cuenta con tareas que implican la participación de cada uno de los diferentes grupos.

(1) https://docs.microsoft.com/en-us/azure/role-based-access-control/

(2) https://www.gartner.com/en/documents/4005344/market-guide-for-ai-trust-risk-and-security-management

(3) https://docs.microsoft.com/en-us/compliance/assurance/assurance-risk-management-program

(4) Tweneboah-Koduah, Samuel & Buchanan, William. (2018). Security Risk Assessment of Critical Infrastructure Systems: A Comparative Study. The Computer Journal. 61. 10.1093/comjnl/bxy002.

Como resultado, el marco de trabajo para evaluar los riesgos de seguridad de los sistemas de IA desarrollado por Microsoft presenta las siguientes características:

- **Ofrece una visión completa de la seguridad de los sistemas de IA**: Cada elemento del ciclo de vida del sistema de IA en un entorno de producción es examinado: comenzando con la recopilación de los datos, continuando con el procesado de estos y finalizando con el desarrollo de los modelos. También se tiene en cuenta las cadenas de suministro de IA, que son los algoritmos y procesos que componen un sistema de IA; así como los controles y las políticas con respecto a las copias de seguridad, la recuperación de datos y la planificación de posibles imprevistos relacionados con los sistemas.
- **Describe las amenazas del ML y las recomendaciones para eliminarlas**: Con el objetivo de ayudar a los ingenieros y profesionales de la seguridad, Microsoft ha enumerado las posibles amenazas que pueden aparecer en cada uno de los pasos que constituyen el proceso de desarrollo de los sistemas de IA. En la sección "MITRE ATLAS", se describe un documento que contiene las mejores prácticas recomendadas por Microsoft que refuerzan las prácticas de seguridad de los *softwares* ya existentes en el contexto de la seguridad de los sistemas de IA.
- **Permite a las organizaciones realizar evaluaciones de riesgos**: El marco de trabajo desarrollado por Microsoft proporciona la capacidad de reunir información sobre el estado actual de los sistemas de IA en una organización, realizar un análisis de las deficiencias y hacer un seguimiento del progreso de la situación de la seguridad.

Tecnologías para aplicar la privacidad y seguridad

Microsoft proporciona las siguientes tecnologías para gestión de la privacidad y seguridad:

Smart Noise, para la Privacidad Diferencial	Presidio, para la protección y anonimización del dato	*SEAL Homomorphic Encryption*, para cálculos sobre datos cifrados en *cloud*	*Confidencial computing for ML*, infraestructura *cloud* privada y segura

Smart Noise, para la Privacidad Diferencial

La pandemia de COVID-19 ha demostrado la importancia de contar con datos suficientes y relevantes para la investigación, el análisis causal, la acción gubernamental y el progreso médico. Sin embargo, por consideraciones de protección de datos, los responsables de la toma de decisiones suelen ser reacios a compartir datos personales o confidenciales. Necesitamos obtener información los datos personales y, al mismo tiempo, proteger la privacidad de las personas de manera confiable.

La privacidad diferencial es un estándar para proteger datos en aplicaciones que preparan y publican análisis estadísticos que pueden ser gestionados mediante aprendizaje automático. El concepto tiene su origen en un estudio realizado por *Microsoft Research*, la división de investigación y desarrollo de la compañía en *Redmond* publicado en 2006.

La privacidad diferencial proporciona una garantía de privacidad matemáticamente medible para las personas al agregar una cantidad cuidadosamente ajustada de ruido estadístico a datos o cálculos confidenciales. Ofrece niveles de protección de privacidad significativamente más altos que las prácticas de limitación de divulgación comúnmente utilizadas, como la anonimización de datos. Este último muestra cada vez más vulnerabilidad a los ataques de reidentificación, especialmente a medida que se hacen públicos más datos sobre las personas.

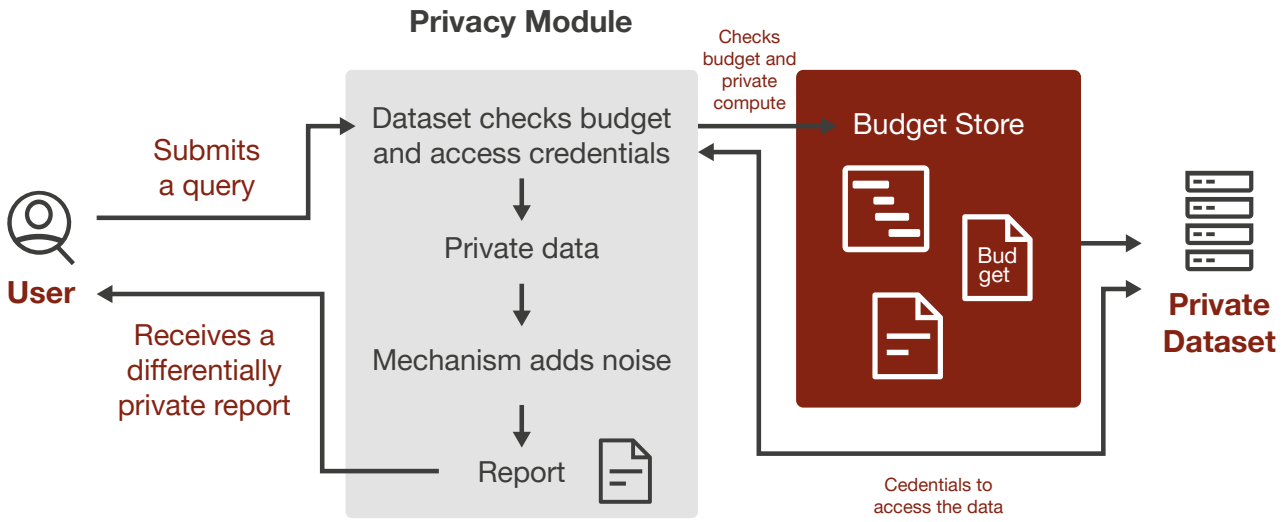
SmartNoise (1) es una plataforma *opensource* publicada en *GitHub* (2) desarrollado conjuntamente por Microsoft, el Instituto de Ciencias Sociales Cuantitativas (IQSS) de *Harvard* y la Escuela de Ingeniería y Ciencias Aplicadas (SEAS) como parte de la iniciativa *Open Differential Privacy* (OpenDP) . La plataforma dispone de mecanismos para proporcionar resultados privados diferenciales a los usuarios de consultas analíticas para proteger el conjunto de datos subyacente. *SmartNoise* incluye algoritmos diferencialmente privados, técnicas para administrar presupuestos de privacidad para consultas posteriores y otras capacidades (3).

(1) https://smartnoise.org/

(2) https://github.com/opendifferentialprivacy

(3) https://azure.microsoft.com/mediahandler/files/resourcefiles/microsoft-smartnoisedifferential-privacy-machine-learning-case-studies/SmartNoise%20Whitepaper%20Final%203.8.21.pdf

El enfoque de Microsoft



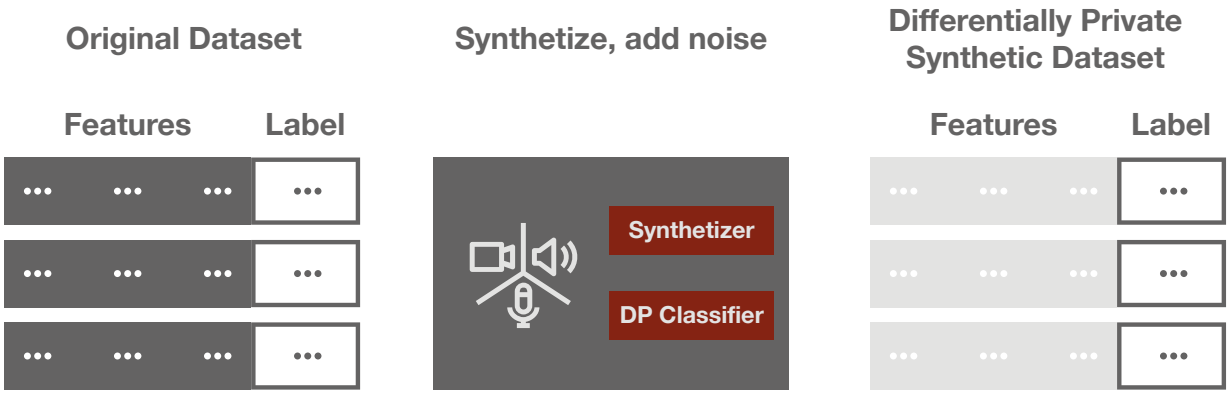
Fuente: Microsoft

Datos sintéticos que preservan la privacidad

SmartNoise, permite crear conjuntos de datos privados diferencialmente derivados de datos desprotegidos. Genera un conjunto de datos sintéticos diferencialmente privados a partir de un modelo estadístico basado en el conjunto de datos original. El conjunto de datos sintéticos representa una muestra "falsa" derivada de los datos originales y conserva tantas características estadísticas como sea posible. La ventaja esencial es que el conjunto de datos privados diferencialmente se puede analizar cualquier número de veces sin aumentar el riesgo de privacidad. Por lo tanto, permite la colaboración entre varias partes, democratizando el conocimiento o iniciativas de conjuntos de datos abiertos.

Si bien el conjunto de datos sintético incorpora las propiedades esenciales de los datos originales, es matemáticamente imposible preservar el valor total de los datos y garantizar la privacidad a nivel de registro al mismo tiempo. Por lo general, no es posible realizar análisis estadísticos y tareas de aprendizaje automático en el conjunto de datos sintetizados en la misma medida en que es posible con los datos originales. Por lo tanto, se debe considerar el tipo de trabajo a realizar antes de sintetizar los datos.

Por ejemplo, el flujo de trabajo para generar un conjunto de datos sintéticos para el aprendizaje automático supervisado con *SmartNoise* tiene el siguiente aspecto:



Fuente: Microsoft

Existen varias técnicas para generar datos sintéticos con provaciad diferencial diferencialmente privados, incluidos enfoques basados en redes neuronales profundas, codificadores automáticos y modelos de confrontación generativos. *SmartNoise* estos son algunos de los sintetizadores de datos incluídos:

- **Mecanismo Exponencial de Pesos Multiplicativos (MWEM)**. Logra privacidad diferencial al combinar pesos multiplicativos y técnicas de mecanismo exponencial de manera simple pero efectiva, requiriendo menos recursos computacionales y menor tiempo de ejecución.
- **Red Adversaria Generativa Diferencialmente Privada (DPGAN)**. Agrega ruido al discriminador de la GAN (Red Generativa Antagónica) para hacer cumplir la privacidad diferencial, y es especialmente utilizado con datos de imagen y registros de salud electrónicos (HER).
- **Agrupación Privada de Conjuntos Docentes Red Generativa Adversaria (PATEGAN)**. Es una modificación que se aplica a las GAN para preservar la privacidad diferencial de los datos sintéticos, mejorando a las DPGAN especialmente para tareas de clasificación.
- **DP-CTGAN**. Sintetiza datos tabulares y aplica DPSGD (el mismo método para garantizar la privacidad diferencial que usa DPGAN).

Más información sobre la privacidad diferencial

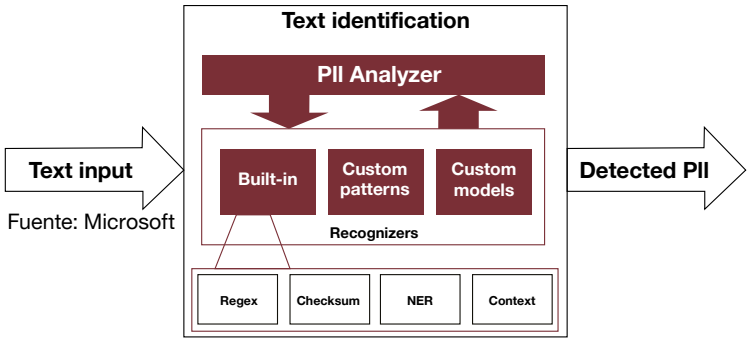
La protección de datos en empresas, autoridades gubernamentales, instituciones de investigación y otras organizaciones es un esfuerzo conjunto que involucra varios roles, incluidos analistas, científicos de datos, oficiales de privacidad de datos, tomadores de decisiones, reguladores y abogados.

Para hacer que el concepto altamente efectivo, pero no siempre intuitivo de privacidad diferencial sea accesible a una amplia audiencia, Microsoft ha publicado un documento técnico completo sobre la técnica y sus aplicaciones prácticas (1). En el documento, se puede obtener información sobre los riesgos subestimados de las prácticas comunes de anonimización de datos, la idea detrás de la privacidad diferencial y cómo usar *SmartNoise* en la práctica. Además, se evalúan diferentes niveles de protección de la privacidad y su impacto en los resultados estadísticos.

(1) https://azure.microsoft.com/en-us/resources/microsoft-smartnoisedifferential-privacy-machine-learning-case-studies/

Presidio, para la protección y anonimización del dato

Presidio (1) es una biblioteca de código abierto publicada en *GitHub* (2) para la protección de datos y la anonimización de texto e imágenes. Proporciona módulos de identificación y anonimización rápidos para texto e imágenes, como números de tarjetas de crédito, nombres, ubicaciones, números de seguro social, billeteras bitcoin, números de teléfono, datos financieros y más.



Fuente: Microsoft

El analizador de texto de Presidio es un servicio basado en *Python* para detectar entidades PII (Información de Identificación Personal) en texto, soportando varios idiomas y entidades distintas en función del país (por ejemplo, el NIF Español). Durante el análisis, ejecuta un conjunto de diferentes reconocedores de PII, cada uno de los cuales se encarga de detectar una o más entidades de PII utilizando diferentes mecanismos.

El analizador Presidio viene con un conjunto de reconocedores predefinidos, pero se puede ampliar fácilmente con otros tipos de reconocedores personalizados. Los reconocedores predefinidos y personalizados aprovechan las expresiones regulares, el reconocimiento de entidades con nombre y otros tipos de lógica para detectar PII en texto no estructurado.

Este mismo proceso puede realizarlo sobre texto incluido en imágenes, mediante la integración de OCR (Reconocimiento óptico de caracteres), aunque esta capacidad está en versión beta ahora mismo.

(1) https://microsoft.github.io/presidio/
(2) https://github.com/Microsoft/presidio
(3) https://presidio-demo.azurewebsites.net/

SEAL Homomorphic Encryption para cálculos sobre datos cifrados en cloud

Microsoft SEAL (1), con tecnología de cifrado homomórfico, es una herramienta de código abierto publicado en GitHub (2), proporciona un conjunto de bibliotecas de cifrado que permiten realizar cálculos directamente en datos cifrados. Esto permite a los ingenieros de software crear servicios de cómputo y almacenamiento de datos cifrados de extremo a extremo en los que el cliente nunca necesita compartir su clave con el servicio en cloud donde se almacenan los datos.

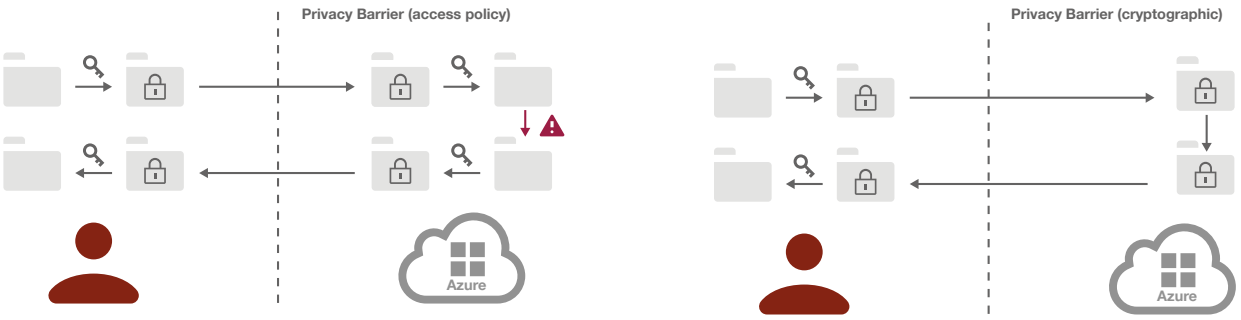
El cifrado homomórfico se refiere a un nuevo tipo de tecnología de cifrado que permite que el cómputo se realice directamente sobre datos cifrados, sin necesidad de descifrado en el proceso. El primer esquema de cifrado homomórfico se inventó en 2009 y se crearon varios esquemas mejorados en los años siguientes, pero su uso requería una amplia comprensión de las matemáticas complicadas que subyacen al cifrado homomórfico y no eran fáciles de usar por desarrolladores de software.

En la solución tradicional de computación y almacenamiento en la nube, los clientes deben confiar en el proveedor de servicios para almacenar y administrar sus datos de manera adecuada, por ejemplo, para no compartirlos con terceros sin el consentimiento del cliente. Microsoft SEAL reemplaza esta confianza con criptografía que permite que los servicios en la nube brinden capacidades de cómputo y almacenamiento encriptado, al mismo tiempo que garantiza que los datos de sus clientes nunca estarán expuestos a nadie en forma no encriptada.

(1) https://www.microsoft.com/en-us/research/project/microsoft-seal/
(2) https://github.com/Microsoft/SEAL

En las soluciones tradicionales de computación y almacenamiento en la nube, la nube debe tener acceso sin cifrar a los datos del cliente para calcularlos, lo que necesariamente expone los datos a los operadores de la nube. La privacidad de los datos se basa en las políticas de control de acceso implementadas por la nube y en las que confía el cliente.

Con *Microsoft SEAL*, los operadores de la nube nunca tendrán acceso sin cifrar a los datos que están almacenando y computando. Esto es posible gracias a la tecnología de encriptación homomórfica, que permite que los cálculos se realicen directamente en los datos encriptados. La privacidad de los datos se basa en criptografía (matemáticas) de última generación y toda la divulgación de información será controlada por el cliente.



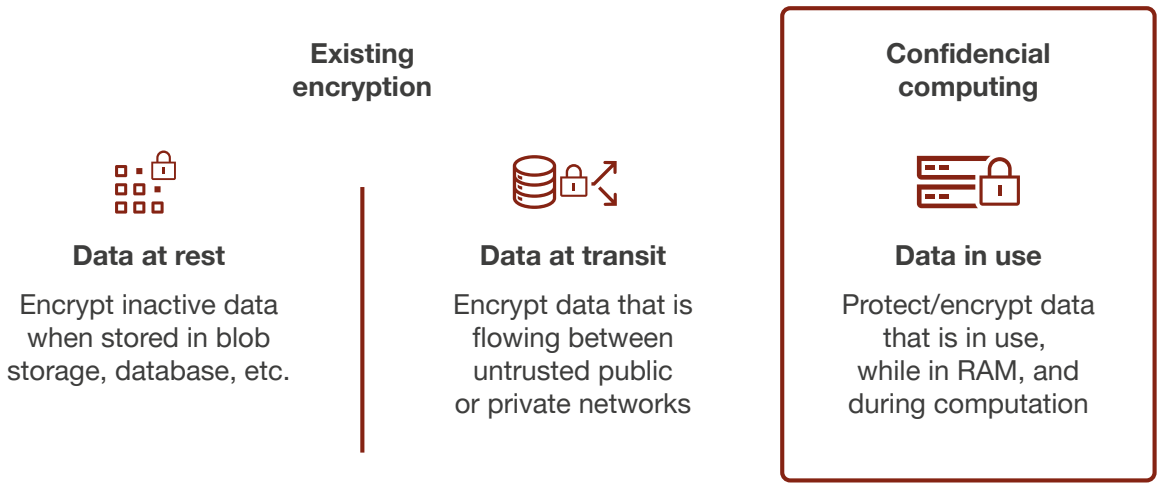
Fuente: Microsoft

Confidencial computing for ML, infraestructura *cloud* privada y segura

Un TEE es un entorno que aplica la ejecución de solo código autorizado. Los datos de TEE no se pueden leer ni alterar con ningún código ajeno a ese entorno.

En la actualidad, los clientes cifran tanto sus datos en reposo como los que están en tránsito, pero no mientras estén en uso en la memoria. El Consorcio de Computación Confidencial (CCC), del que Microsoft es cofundador, define la computación confidencial como la protección de los datos en uso mediante entornos de ejecución de confianza (TEE), basados en hardware.

Microsoft dispone de *Confidential computing for ML* (1), un servicio de infraestructura en Azure basado en hardware de Intel.



Fuente: Microsoft

(1) https://docs.microsoft.com/es-es/azure/confidential-computing/

El enfoque de Microsoft

Estos entornos impiden el acceso o la modificación no autorizados de aplicaciones y datos mientras están en uso, lo que aumenta el nivel de seguridad de las organizaciones que administran datos confidenciales y regulados. Los TEE son un entorno de confianza que proporciona un nivel de garantía de la integridad y confidencialidad de los datos, y de la integridad del código. El objetivo del modelo de amenazas de la computación confidencial es eliminar o reducir la capacidad de un operador de proveedor de nube y de otros actores del dominio del inquilino para acceder al código y a los datos mientras se ejecutan.

Cuando se usa con el cifrado de datos en reposo y en tránsito, la computación confidencial elimina la única barrera grande de cifrado (cifrado durante el uso) mediante la protección de conjuntos de datos confidenciales o altamente regulados y cargas de trabajo de aplicaciones en una plataforma de nube pública segura. La computación confidencial va más allá de la protección de datos genérica.

La computación confidencial permite aislar los datos confidenciales mientras se procesan, permitiendo el uso de la nube para soluciones inteligentes donde es necesario, por ejemplo:

- Proteger datos financieros.
- Proteger información médica de pacientes
- Ejecutar procesos de aprendizaje automático sobre información confidencial.
- Realizar algoritmos en conjuntos de datos cifrados de varios orígenes.

Principio de Inclusión

En esta sección vamos a desarrollar:

- Como Microsoft enfoca la inclusión en la IA.
- *Guidelines* para un diseño centrado en las personas.
  - Para la interacción humano-IA en general (*Human-AI Experiences -HAX-*).
  - Para la Inclusión en particular.

Como Microsoft entiende la Inclusión en la IA

Microsoft ha desarrollado herramientas de diseño inclusivas y procesos para reconocer a las personas con discapacidades durante la etapa de diseño de la solución, para que se tengan en cuenta a la hora de tomar las decisiones. A medida que evolucionan sus prácticas, han ampliado su abanico de diseño a otras áreas de inclusión, como los problemas cognitivos, prejuicios sociales o déficits en el aprendizaje.

En la actualidad, las máquinas suelen aprender con un modelo de datos que sería como si a una persona se le enseñase mediante ejemplos. Una programación compleja que intenta simular el método de aprendizaje humano; las personas pueden reconocer una cara en una fiesta o ir en bicicleta pese a llevar años sin hacerlo. Las máquinas pueden alcanzar cierto grado de conocimiento tácito, pero aún les falta el matiz humano que permite a las personas realizar estas acciones (1). Es ese matiz que falta el que ha dado lugar a llevar el foco a una IA inclusiva en Microsoft. El campo de la IA se encuentra en constante crecimiento y los equipos están aprendiendo constantemente, analizando los resultados obtenidos de los experimentos y probando nuevas soluciones. Con el objetivo de mantener el ritmo de estos cambios y mantenerse a la vanguardia de la innovación, el equipo de 'Diseño inclusivo' de Microsoft se ha aliado con desarrolladores, investigadores y grupos encargados de la parte legal para tender un puente entre las mentes más técnicas de Microsoft y aquellas más centradas en el día a día de las personas. Esta unión ha permitido formar cinco ideas clave para identificar la exclusión y crear una IA más inclusiva:

- **Redefinir el sesgo como un rango**: Muchas conversaciones sobre la IA se encuentran polarizadas entre IAs "buenas" y "malas", pero los equipos encargados del diseño y desarrollo de estas a veces tienen dificultades para ver en sus diseños algunos de los casos más extremos de sesgo que aparecen en los medios. En vez de enfocarse en los casos más extremos, los equipos de Microsoft han aprendido a atacar todo un espectro de pequeños sesgos que se acumulan en el día a día. Situaciones como no poder introducir el nombre real del cliente en la aplicación porque incluye algún carácter externo al alfabeto inglés, en España, por ejemplo, la 'ñ', como en el caso de 'Iñigo'.
- **Ayudarse de los clientes para corregir el sesgo**: El aprendizaje del modelo lo es todo a la hora de diseñar una IA más inclusiva. Desafortunadamente, el desarrollo de IA se suele hacer a puerta cerrada, restringido al *input* de equipos que pueden no representar la totalidad del espectro del cliente para el que se diseña. Estos últimos años ha habido un crecimiento exponencial de conversaciones en internet acerca de la ética de la IA, esto ha dado lugar a proyectos de código libre a gran escala como simuladores para coches autónomos y mejoras en el *Text-To-Speech* (2), para acercarse más al lenguaje natural. Microsoft considera que es importante tener en cuenta estas 'voces' de la comunidad para crear una IA lo más inclusiva posible.

(1) Artículo: https://hms.harvard.edu/magazine/artificial-intelligence/importance-nuance

(2) *Text-To-Speech*: Es el método de generar un habla similar a la humana, de forma artificial, con la ayuda de máquinas. Un sistema *Text-to-Speech* acepta el lenguaje humano en forma de texto como entrada y lo convierte en voz como salida.

El enfoque de Microsoft

- **Cultivar la diversidad con privacidad y consentimiento**: Una de las partes más importantes del crecimiento de una IA es que su conjunto de datos sea diverso, esté correctamente etiquetado y se use de una forma que represente a la totalidad de los clientes. Si ya existe un sesgo en este conjunto de datos, lo único que hará será agravarlo. El tratamiento de estos datos debe además adoptar el RGPD (1) ; la privacidad por diseño es uno de los pilares de Microsoft, y no actúa de forma reactiva respecto a esta, sino de manera proactiva.
- **Equilibrar inteligencia con descubrimiento**: El algoritmo de un sistema IA asume datos de nuestro presente basándose en nuestro comportamiento pasado, a veces con poca flexibilidad, y provocando muchas veces situaciones incómodas. Los clientes han de sentir que tienen la opción de cambiar de rumbo en cualquier momento para alcanzar sus nuevos objetivos sin sentirse limitados por la IA.
- **Construir equipos de IA inclusivos**: Por mucho que se quiera creer en la neutralidad de la IA, no deja de ser, hasta cierto punto, un reflejo del equipo que se encargó de su diseño. Por ello, es primordial la contratación de personal con diversas experiencias, especializaciones, géneros, razas, culturas e intereses. Equipos diversos encuentran posibles sesgos de manera más sencilla, de forma que se pueda entrenar a la IA para ser lo más inclusiva posible. Estos equipos han de ser abiertos de mente y tomar responsabilidad por sus errores accidentales, puesto que sus acciones tendrán eco en los medios y en la opinión de la sociedad, además, siempre teniendo en mente el sesgo inherente de sus diseños.

(1) RGPD: El Reglamento General de Protección de Datos (o GDPR en inglés) es el reglamento europeo relativo a la protección de las personas físicas en lo que respecta al tratamiento de sus datos personales y a la libre circulación de estos datos.

Una solución para uno, extensible a muchos

Microsoft mantiene el foco en lo que es universalmente importante para todas las personas. Todas ellas tienen motivaciones y construyen relaciones, y todas tienen habilidades y límites a ellas. Microsoft es consciente de que todo el mundo puede experimentar exclusión cuando interactúa con un diseño en particular, pero, también todas las personas pueden beneficiarse de un diseño inclusivo.

Diseñar para personas con discapacidades permanentes puede parecer como una gran limitación, pero los diseños resultantes pueden beneficiar a un gran número de personas. Por ejemplo, algo diseñado para una persona con un solo brazo puede beneficiar a una persona que sólo tiene una limitación temporal en el mismo, como una escayola o incluso a un padre que sostiene a su hijo recién nacido. Microsoft ha definido esto como el *Persona Spectrum*.

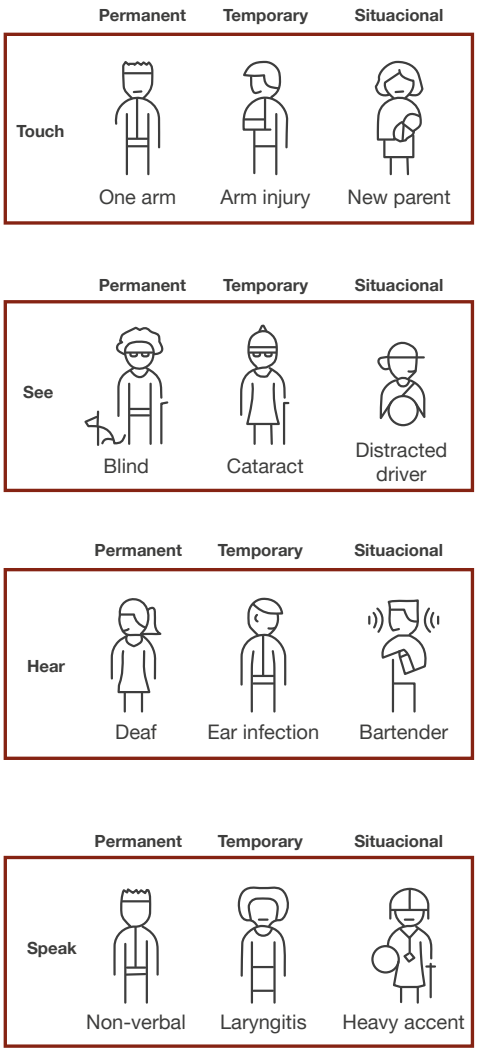
Tener en consideración el rango de personas con discapacidades desde las permanentes hasta casos más situacionales permite a Microsoft escalar sus diseños a más personas. Por ejemplo, se estima que en los Estados Unidos, cada año, 26.000 personas sufren la pérdida de una extremidad superior, pero si se incluye a este cálculo el número de personas con este tipo de problemas de forma temporal o situacional, el número asciende a más de 20 millones.

Reconocer la exclusión

Diseñar para la inclusión no solo abre nuestros productos y servicios a más personas, sino que también refleja cómo son realmente las personas. Todos los humanos crecen y se adaptan al mundo que los rodea y queremos que nuestros diseños reflejen eso.

Aprender de la diversidad

Los seres humanos somos los verdaderos expertos en adaptarnos a la diversidad. El diseño inclusivo pone a las personas en el centro desde el comienzo del proceso, y esas perspectivas frescas y diversas son la clave para una visión real.



Fuente: Microsoft

Guidelines para un diseño centrado en las personas

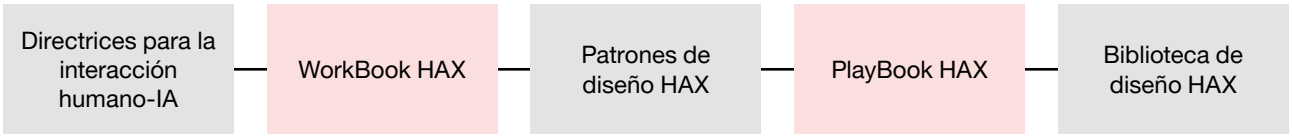
Si el objetivo es construir una IA que realmente ayude y entienda a las personas, debe ser concebida desde un punto de vista humano. No se puede depositar toda la confianza en que todos los desarrollos futuros vayan a ser inclusivos únicamente porque hasta el presente lo hayan sido. Para Microsoft es importante reconocer que errar es humano y que en ocasiones es necesario pararse a pensar y considerar el por qué se está moviendo en esa dirección el diseño. Puede, por ejemplo, que sea necesario incluir a más personas en el proceso de creación, para que estas aporten nuevos puntos de vista no sesgados y así seguir mejorando.

La naturaleza humana es imperfecta por defecto, pero también es increíblemente compleja. Nos sentimos empujados a conectar, interactuar, resolver problemas, buscar perspectivas nuevas y seguir avanzando. Microsoft sigue trabajando día a día para llevar estos mismos principios al diseño de IA.

Para ayudar a dicho diseño inclusivo, Microsoft propone el uso de algunas Guidelines desde el comienzo de la concepción de la solución, recomendando tener en cuenta a grupos minoritarios y vulnerables. Además, siempre cumpliendo las leyes sobre accesibilidad e inclusión que afecten a la organización.

Para la interacción humano-IA (*Human-AI Experiences -HAX-*)

El proyecto *HAX Toolkit* es una colaboración entre *Microsoft Research y Aether*, el organismo asesor de Microsoft sobre ética y efectos de la IA en ingeniería e investigación. El uso de este Toolkit permite gestionar correctamente muchos de los principios éticos, y especialmente el de Transparencia. El kit de herramientas ofrece un conjunto de herramientas prácticas para crear experiencias de inteligencia artificial humana:



Directrices para la interacción humano-IA

Human-AI Interaction Guidelines (1) incluye dieciocho directrices recomendadas para diseñar sistemas de IA a lo largo de la interacción del usuario y el ciclo de vida de la solución, brindando recomendaciones sobre cómo crear experiencias significativas con IA que transmitan al usuario que la sensación de control, y que sus valores y objetivos son respetados.

¿Por qué necesitamos pautas para la interacción humano-IA?

Necesitamos pautas para la interacción humano-IA porque los sistemas de IA pueden mostrar comportamientos impredecibles que pueden ser perjudiciales, confusos, ofensivos o incluso peligrosos. Por estas razones, los sistemas de IA a menudo violan los principios tradicionales de diseño de interacción humano-computadora.

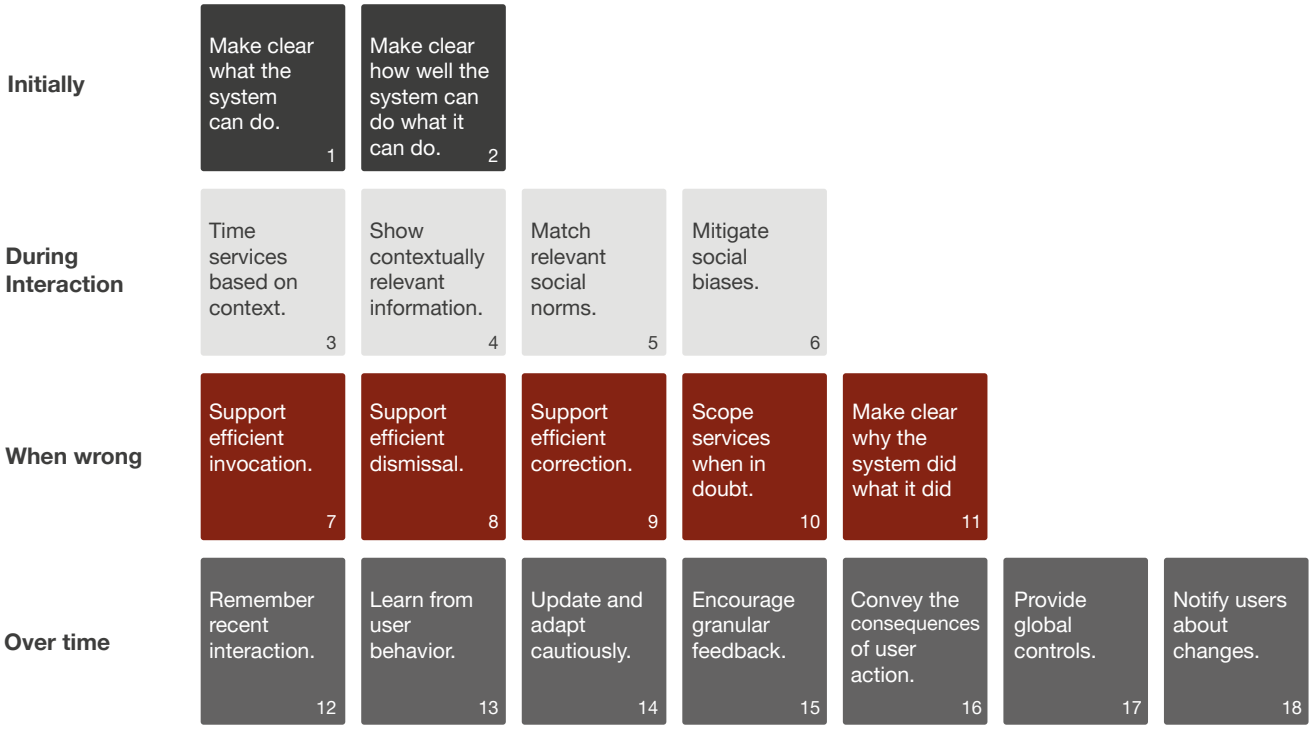
Cuando una aplicación o producto tradicional no se comporta de manera consistente, se considera que tiene una deficiencia o un error de diseño. Sin embargo, la inconsistencia y la incertidumbre son inherentes a los sistemas infundidos con IA debido a su naturaleza probabilística y porque cambian con el tiempo a medida que aprenden con nuevos datos.

(1) https://docs.microsoft.com/en-us/ai/guidelines-human-ai-interaction/

El enfoque de Microsoft

¿Cuáles son las pautas?

Microsoft propone 18 pautas de diseño de aplicación general para la interacción humano-IA. Estas pautas sintetizan más de dos décadas de pensamiento e investigación sobre cómo hacer que la IA sea fácil de usar. Las pautas se agrupan en cuatro categorías, según la fase de interacción del usuario a la que se aplican.



Fuente: Microsoft

WorkBook HAX

El libro de trabajo HAX (1) es una herramienta para estructurar conversaciones tempranas a través de los múltiples roles necesarios para implementar las pautas para la interacción humano-IA vistas anteriormente. Está disponible como una hoja de cálculo de Excel (2). Permite abstraerse de todo el tecnicismo pudiendo hablar del mismo, pero desde una perspectiva de usuario final.

¿Por qué el utilizarlo?

Microsoft considera que la creación de una solución inteligente tiene más éxito cuando se aúnan todas las disciplinas involucradas en el desarrollo del mismo, permitiendo empoderar a los equipos para que colaboren y planifiquen todas las disciplinas necesarias.

Muchas de las directrices se basan por debajo en la tecnología, proporcionando una visión holística de toda la solución que permite que las especificaciones de la solución sean flexibles y orientadas al usuario final.

(1) https://www.microsoft.com/en-us/haxtoolkit/workbook/
(2) https://www.microsoft.com/en-us/haxtoolkit/uploads/prod/2021/05/2-HAX_Workbook.xlsx

¿Cómo usarlo?

Es fundamental utilizar el libro de trabajo al principio del proyecto, en la de la etapa de definición de requisitos, o al rediseñar uno existente. Debido a que la implementación de las mejores prácticas de interacción humano-IA afecta a toda la solución, permite combinar UX, IA, gestión de proyectos y tecnología guiando al equipo a través de cinco pasos:

- Seleccionar qué pautas son relevantes para un sistema o característica del mismo.
- Estimar los requisitos de IU, IA, datos e ingeniería necesarios para implementar las directrices de alto impacto. Los patrones de diseño HAX que se muestran más adelante pueden ayudar con este paso.
- Comenzando con pautas de alto impacto, se describen los requisitos de IU, IA, datos e ingeniería para implementar cada pauta.
- Priorización de las Directrices para implementar, considerando las ventajas y desventajas entre el impacto del usuario y el costo.
- Seguimiento de su progreso.

Patrones de diseño HAX

Los patrones de diseñoHAX permiten describir soluciones flexibles para problemas recurrentes de interacción humano-IA. Cada patrón sigue la misma estructura que guía desde el problema que el patrón puede resolver hasta su solución.

¿Por qué diseñar patrones?

Existen múltiples formas de implementar cada una de las directrices vistas anteriormente, por lo que en ocasiones cuesta decidir cuál es el mejor enfoque para el escenario de la solución que estamos implementando. Los equipos suelen pedir algo que sea confiable y que no esté sesgado hacia un punto de vista u otro de todas las alternativas posibles. Estos patrones permiten ahorrar tiempo a los equipos y brindar experiencias de usuario de alta calidad, ya que los patrones retutilizan soluciones establecidas en experiencias previas.

¿Cómo usarlos?

Mirosoft recomienda consultar los patrones cuando el equipo esté definiendo requisitos y directrices para la solución o en las primeras etapas de creación de prototipos. Cuando se haya seleccionado una pauta para implementar, hay que revisar los patrones y los ejemplos correspondientes para esa pauta en la Biblioteca de diseño HAX que veremos posteriormente.

Los patrones son independientes de la interfaz de usuario y se pueden implementar en una variedad de sistemas e interfaces.

Debido a que la mayoría de los patrones de diseño afectan a muchos componentes de la solución (interfaz, datos, modelos, etc.) este trabajo debe ser realizando con la colaboración de los equipos de UX, AI, gestión de proyectos e ingeniería.

PlayBook HAX

Es una herramienta interactiva (1) que permite explorar de manera proactiva y sistemática las fallas comunes de interacción entre humanos y IA en soluciones inteligentes de varias tipologías, todas ellas sobre datos no estructurados (texto):

- Búsqueda.
- Recomendación.
- IA conversacional.
- Predicción y asistencia de texto.
- Clasificación de texto.

El Playbook permite enumerar las fallas relevantes para poder diseñar las formas para que los usuarios finales se recuperen de manera eficiente. El *PlayBook HAX* también proporciona orientación práctica y ejemplos sobre cómo simular de manera sencilla los comportamientos del sistema para las primeras pruebas de los usuarios.

(1) https://microsoft.github.io/HAXPlaybook/

El enfoque de Microsoft

Biblioteca de diseño HAX (1)

Es una extensa librería que aúna las Directrices y Patrones descritos previamente, permitiendo buscar combinaciones y ejemplos para doce categorías de producto y otros tantos tipos de aplicación de inteligencia artificial.

Guidelines

- ☐ G1: Make clear what the system can do.
- ☐ G2: Make clear how well the system can do what it can do.
- ☐ G3: Time services based on context.
- ☐ G4: Show contextually relevant information.
- ☐ G5: Match relevant social norms.
- ☐ G6: Mitigate social biases.
- ☐ G7: Support efficient invocation.
- ☐ G8: Support efficient dismissal.
- ☐ G9: Support efficient correction.
- ☐ G10: Scope services when in doubt.
- ☐ G11: Make clear why the system did what it did.
- ☐ G12: Remember recent interactions.
- ☐ G13: Learn from user behavior.
- ☐ G14: Update and adapt cautiously.
- ☐ G15: Encourage granular feedback.
- ☐ G16: Convey the consequences of user actions.
- ☐ G17: Provide global controls.
- ☐ G18: Notify users about changes.

Products Categories

- ☐ Advertising
- ☐ Chatbot
- ☐ E-commerce
- ☐ Email
- ☐ Health and wellness
- ☐ Maps and navigation
- ☐ News, media, and entertainment
- ☐ Productivity
- ☐ Search engine
- ☐ Social networking
- ☐ Voice assistants
- ☐ Writing and editing

AI Application Types

- ☐ Classification
- ☐ Facial recognition
- ☐ Filtering and ranking
- ☐ Image recognition
- ☐ Natural language processing (Text)
- ☐ Natural language processing (Voice)
- ☐ Prediction
- ☐ Recommender systems
- ☐ Route planning
- ☐ Search
- ☐ Text generation

(1) https://www.microsoft.com/en-us/haxtoolkit/library/

Inclusive Design Guidelines

Las pautas de diseño inclusivo de Microsoft definen 3 principios de inclusión. Reconocer la exclusión, diseñando inclusivamente y abriendo los productos y servicios a más gente, de manera que nuestros diseños reflejen crecimiento y adaptación. Resuelve para uno y aplica al resto, diseñando para personas con discapacidades permanentes y obteniendo un resultado que beneficia a todo el mundo. Y, por último, aprender de la diversidad, ya que, si nos enfocamos en la gente desde el primer momento, podremos utilizar diversas perspectivas que nos llevarán a una mejor percepción.

Microsoft proporciona unas *Guidelines* y recursos (1) entre las que están:

- **Inclusivo 101**. Una introducción completa al mundo del diseño inclusivo, para cambiar el pensamiento en el diseño hacia soluciones universales.
- **Actividades inclusivas**. Es un juego compuesto de tarjetas diseñadas para integrarse en el proceso de diseño poniendo en marcha el pensamiento creativo y los conceptos de prueba de estrés a través de una mente inclusiva.
- **Un manual** para comprender cómo el sesgo afecta la inteligencia artificial.
- **El ejemplo** de cómo un equipo de producto integró el diseño inclusivo en su trabajo.

(1) https://www.microsoft.com/design/inclusive/

Principio de Transparencia

En esta sección vamos a desarrollar:

- Qué es la transparencia para Microsoft.
- Cuáles son las prácticas recomendadas por Microsoft.
- Tecnologías para aplicar la transparencia (*InterpretML, DiCE y EconML*)
- Herramienta de gestión (*Datasheets for Datasets*)

Qué es la Transparencia para Microsoft

La Transparencia o interpretabilidad del modelo es fundamental para los científicos de datos, los auditores y los responsables de la toma de decisiones comerciales para garantizar el cumplimiento de las políticas de la empresa, los estándares de la industria y las regulaciones gubernamentales:

- Los científicos de datos necesitan la capacidad de explicar sus modelos a los ejecutivos y las partes interesadas, para que puedan comprender el valor y la precisión de sus hallazgos. También requieren interpretabilidad para depurar sus modelos y tomar decisiones informadas sobre cómo mejorarlos.
- Los auditores legales requieren herramientas para validar modelos con respecto al cumplimiento normativo y monitorear cómo las decisiones de los modelos están impactando a los humanos.
- Los tomadores de decisiones comerciales necesitan tranquilidad al tener la capacidad de brindar transparencia a los usuarios finales. Esto les permite ganar y mantener la confianza.

Habilitar la capacidad de explicar un modelo de aprendizaje automático es importante durante dos fases principales del desarrollo del modelo:

- Durante la fase de capacitación, los diseñadores y evaluadores de modelos pueden utilizar los resultados de interpretación de un modelo para verificar hipótesis y generar confianza con las partes interesadas. También utilizan los conocimientos sobre el modelo para depurar, validar que el comportamiento del modelo coincida con sus objetivos y comprobar si el modelo no es justo o tiene características insignificantes.
- Durante la fase de inferencia, tener transparencia en torno a los modelos implementados permite a los ejecutivos comprender "cuando se implementa" cómo funciona el modelo y cómo sus decisiones tratan e impactan a las personas en la vida real.

Prácticas recomendadas

- Dejar claro cuál es el alcance y rendimiento del modelo. Esto ayuda a los usuarios a gestionar expectativas y a no especular en torno a los resultados obtenidos.
- Compartir las características principales de los conjuntos de datos para ayudar a los desarrolladores a comprender si un conjunto de datos concreto es adecuado para su caso de uso.
- Mejorar la inteligibilidad del modelo al aprovechar modelos más sencillos y generar explicaciones inteligibles del comportamiento del modelo.
- Simplificar los modelos mediante *feature selection* (1), no solo potenciando así su explicabilidad sino reduciendo también el coste computacional del modelo. Sin embargo, se ha de ser cauteloso para evitar introducir sesgos introducidos por el desarrollador.
- Aplicar diferentes métodos tales como LIME (2) y SHAP (3) (métodos disponibles dentro de la herramienta InterpretML) sobre los modelos, para analizar la contribución de cada una de las variables a la predicción, tanto de forma global a todo el modelo como local a ca da una de las predicciones.
- Formar a los empleados, empoderándolos para que puedan interpretar los resultados de la inteligencia artificial.

(1) Proceso por el cual se reduce el número de variables de entrada cuando se desarrolla un modelo predictivo.

(2) https://interpret.ml/docs/lime.html

(3) https://interpret.ml/docs/shap.html

Tecnologías para aplicar la transparencia

Microsoft dispone de tres herramientas tecnológicas para gestionar el principio de transparencia:

- ***InterpretML***, para explicar el **funcionamiento interno** de modelos de caja blanca y de caja negra.
- ***DiCE (Diverse Counterfactual Explanations for Machine Learning Classifiers)***, para gestionar la **contrafactualidad**, "interrogando" al modelo para encontrar los cambios necesarios que cambiarían su decisión y así ofrecer una explicación complementaria a la decisión tomada.
- ***EconML***, para gestionar la **inferencia causal** y conocer la causa de predicciones concretas realizadas por un modelo de *machine learning*.

InterpetML para explicar el funcionamiento de los modelos

Microsoft ha desarrollado *InterpretML* (1), un conjunto de herramientas para ayudar a comprender los modelos y habilitar así la Transparencia sobre los mismos. Es un paquete *open-source* en *Python* disponible en *GitHub* (2).

Permite explicar modelos de caja blanca y de caja negra, y hacerlo de distintas maneras:

- De manera global, permitiendo explorar el comportamiento general del modelo y encontrar las características principales que afectan a las predicciones del mismo.
- De manera local, permitiendo explicar una predicción individual y encontrar las características que contribuyan a realizarla.
- En un subconjunto de predicciones, explicando las características del grupo utilizadas para las decisiones sobre el mismo.
- Permitiendo utilizar técnicas como el análisis hipotético, analizando cómo los cambios en las entradas afectan las predicciones.

¿Quién puede beneficiarse de *InterpretML*?

- Científicos de datos, permitiéndoles comprender los modelos, depurarlos, descubrir errores y explicar el modelo a otras partes interesadas.
- Auditores externos, ya que permite validar un modelo antes de implementarlo y auditarlo después de la implementación.
- Responsables de negocio, a los que les permite comprender cómo se comportan los modelos para brindar transparencia sobre las predicciones a los clientes.
- Investigadores, ya que permite integrar con nuevas técnicas de interpretabilidad y comparar algoritmos entre sí.

(1) https://interpret.ml/
(2) https://github.com/interpretml/interpret/

DiCE para gestionar la contrafactualidad

Es un proyecto *open-source* accesible en *GitHub* (1) cuyo objetivo principal es explicar las predicciones de los sistemas basados en ML que se utilizan para informar las decisiones en dominios críticos para la sociedad. En estos dominios, es importante brindar explicaciones a todas las partes interesadas clave que interactúan con el modelo ML.

Se basa en el concepto de explicación contrafactual, "interrogando" al modelo para encontrar los cambios necesarios que cambiarían su decisión. Específicamente, DiCE proporciona esta información al mostrar las versiones perturbadas de la misma entrada que habrían obtenido un resultado diferente.

Por ejemplo, una persona solicitó un préstamo y fue rechazada por el algoritmo de distribución de préstamos de una empresa financiera. DiCE mostraría un conjunto de versiones perturbadas de características de la misma persona que habría recibido el préstamo por el mismo modelo ML. Es decir, explicaría que dicha persona, por ejemplo, "Habría recibido el préstamo si su ingreso fuera superior en $ 10k". En otras palabras, una explicación contrafáctica ayuda a un sujeto de decisión a decidir qué debe hacer a continuación para obtener un resultado deseado en lugar de proporcionarle solo las características importantes que contribuyeron a la predicción. Además, las explicaciones de CF de DiCE también son útiles para quienes toman las decisiones, ya que pueden usarlas para evaluar la confiabilidad de una predicción particular del modelo de ML. Del mismo modo, las explicaciones de CF sobre múltiples entradas pueden ser útiles para que los evaluadores de decisiones evalúen criterios como la equidad y los desarrolladores de modelos para depurar sus modelos y evitar errores en nuevos datos.

(1) https://github.com/interpretml/DiCE

EconML para la inferencia causal

EconML (1) es un paquete open source desarrollado por *Microsoft Research* publicado en *GitHub* (2) y desarrollado en Python que aplica las técnicas de aprendizaje automático para estimar respuestas causales individualizadas a partir de datos experimentales. Al incorporar pasos individuales de aprendizaje automático en modelos causales interpretables, estos métodos mejoran la confiabilidad de las predicciones hipotéticas y hacen que el análisis causal sea más rápido y sencillo. Veamos este concepto de inferencia causal a través de **dos sencillos ejemplos**:

- **Precios personalizados**. Para establecer la política de descuento personalizada óptima, una empresa necesita comprender cuál es el efecto de una caída en el precio sobre la demanda de un producto por parte de un cliente en función de las características del mismo. Por lo tanto, la gestión e la inferencia causal permite explicar de manera transparente las políticas de fijación de precios.
- **Estratificación en ensayos clínicos**. ¿Qué pacientes deben ser seleccionados para un ensayo clínico? Si queremos demostrar que un tratamiento clínico tiene un efecto en al menos un subconjunto de una población, entonces los ensayos clínicos completamente aleatorios son inapropiados ya que solo estimarán los efectos promedio. Usando inferencia causal de pueden obtener estimaciones de estos efectos e identificar buenos pacientes candidatos para un ensayo clínico.

(1) https://www.microsoft.com/en-us/research/project/econml/
(2) https://github.com/Microsoft/EconML

Herramienta de gestión. Datasheets for Datasets (1)

Datasheets for Datasets es una herramienta para documentar los *datasets* usados a la hora de entrenar y evaluar modelos de *machine learning*. El objetivo de los *datasheets* es aumentar la transparencia de los conjuntos de datos y facilitar una mejor comunicación entre los creadores y los usuarios de dichos datos. Un entendimiento exhaustivo de las características y orígenes de los datos usados en la fase de entrenamiento es fundamental a la hora de construir un modelo de IA más responsable.

Estas hojas de datos alientan a los creadores de conjuntos de datos a reflexionar detenidamente sobre el proceso de creación de los mismos, lo que les permite descubrir posibles fuentes de sesgo en sus datos o suposiciones no intencionales que han hecho.

Por otra parte, para los consumidores de los datos, la información contenida en dichas hojas puede ayudar a garantizar que el conjunto de datos sea la opción correcta para la tarea en cuestión. Adicionalmente, estas hojas de datos se pueden exponer opcionalmente a los usuarios finales para aumentar la transparencia y la confianza.

La *Datasheets for Datasets* contiene más de 50 preguntas que ayudan a comprender la motivación, composición, recopilación, procesamiento previo, etiquetado, usos previstos, distribución y mantenimiento de los datos.

(1) https://www.microsoft.com/en-us/research/publication/datasheets-for-datasets/

Principio de Responsabilidad

En esta sección vamos a desarrollar:

- Qué es la Responsabilidad para Microsoft.
- La involucración del ser humano en las soluciones inteligentes.

Qué es la responsabilidad para Microsoft

Resulta complejo traducir al castellano el término *Accountability*, ya que no se tiene ninguna palabra que haga referencia en concreto a esta acción. *Accountability* puede entenderse como el acto de ser responsable de lo que se exige y en el caso que compete de la inteligencia artificial, hace referencia a introducir al ser humano en todo el proceso de desarrollo de este tipo de soluciones inteligentes para hacerle responsable en las distintas etapas. Es decir, que el *Accountability* en este contexto **se encuentra más próximo a considerarse una práctica que un principio implementable mediante herramientas y tecnologías**.

La Responsabilidad es un pilar fundamental de la IA responsable, válgase la redundancia. Las personas que diseñan e implementan los sistemas de IA deben ser responsables de sus acciones y decisiones, especialmente a medida que se avanza hacia sistemas más autónomos. Por ello, las organizaciones deben considerar establecer un organismo de revisión interno que brinde supervisión, información y orientación sobre el desarrollo y la implementación de sistemas de IA. Este organismo puede colaborar en tareas como la definición de procedimientos recomendados para la documentación y prueba de los sistemas de IA durante el desarrollo o incluso proporcionar orientación dado el caso de que se vaya a utilizar un sistema de IA en situaciones delicadas (como, por ejemplo, aquellos que pueden denegar servicios consecuentes como asistencia sanitaria o empleo, crear riesgo de daño físico o emocional o infringir los derechos humanos).

Desde Microsoft se considera que, hoy por hoy, la inteligencia artificial no va a sustituir de manera completa al ser humano en la toma de decisiones, a no ser que se trate de decisiones muy básicas. Lo realmente importante es lo que esta tecnología puede hacer por las personas y los problemas que puede resolver. Es aquí donde surge un notable conflicto, si no se incorpora al ser humano en todo el proceso (entrenamiento, datos, tecnología, ingeniería, temas legales, temas éticos, etc.) de los sistemas inteligentes.

Otra cuestión a tener en cuenta es que estos sistemas se encuentran en constante evolución, por lo que resulta importante conseguir una visión conjunta de toda la tecnología para llegar a su control completo. Debido a esta continua evolución, surge un nuevo problema con la regulación, ya que lo que se controla hoy en día el día de mañana está obsoleto. Es por ello por lo que, Microsoft, considera una idea inteligente el aproximar los modelos de regulación a modelos basados en riesgo, para de esta manera lograr una regulación viva, entendiéndose viva como algo continuamente actualizado, que a medida que vayan surgiendo nuevos escenarios o casos de uso de alto riesgo se traten como tal y en cada caso se definan unas salvaguardas adicionales para evitar preocupaciones o tener en cuenta consideraciones relativas a la privacidad de los sistemas.

La involucración del ser humano en las soluciones inteligentes

Cuando se requiere la automatización de decisiones para la transformación digital de las organizaciones, los seres humanos deben ser capaces de identificar y resolver los problemas a través de los resultados y la ejecución de los modelos de inteligencia artificial.

Para poder lograr lo anterior, Microsoft considera importante la consecución de los dos siguientes aspectos: la introducción del ser humano en el ciclo de los sistemas y el control humano significativo.

Humanos involucrados en el diseño de los modelos de IA

Desde Microsoft se proponen el siguiente conjunto de acciones para alcanzar el objetivo de integrar la figura del ser humano en todo el recorrido de los sistemas basados en inteligencia artificial:

- Capacitar a la persona/personas responsables de la herramienta en el uso y mantenimiento de la solución de manera responsable y ética. Hacerle comprender el nivel de madurez requerido para diseñar, entregar y operar una solución, así como su propia inteligencia artificial.

- Proporcionar materiales de formación y apoyo que permitan a la persona responsable:
  - Comprender el contexto ético en la que se desarrolló la solución
  - Garantizar que la tecnología continúe funcionando de acuerdo con las mismas pautas éticas con las que fue desarrollada
  - Comprender cuándo la solución puede requerir soporte técnico adicional
- En el caso que se determine que es posible que el cliente no pueda adoptar o mantener la tecnología de manera ética y responsable, se debe plantear esta cuestión para una revisión
- El cliente es responsable de cumplir con todas las leyes aplicables en la implementación y uso de las soluciones

Control humano significativo

- Se debe conseguir un entendimiento claro de quién es el responsable último del sistema inteligente, para ello es importante:
  - Otorgar más importancia a la aplicación final de la solución desarrollada
  - Asegurar que cualquier acción que involucre a niños u otras poblaciones sensibles debe ser tratada con especial cuidado e incluir consideraciones adicionales

Resumen en un vistazo del Toolkit de Microsoft para la gestión de los principios éticos

	Tecnologías	*Guidelines*	Herramientas
Transparencia	InterpetML DICE EconML		Datasheets for Datassets
Responsabilidad			
Inclusividad		Duman-AI Experiences (HAX) Inclusive Design	
Fiabilidad y seguridad	Error Analysis Counterfit		
Privacidad y seguridad	Presidio Smart Noise SEAL Homomorphic Encryption Confidencial computing for ML		
Equidad	FairLearn		AI Fair Checklist

3. Framework GuIA

Cómo aterrizar cada principio ético

El enfoque de IBM

El enfoque de IBM para gestionar los principios éticos

En esta sección dedicada a los planteamientos de IBM vamos a desarrollar el marco de trabajo de dicha compañía para conseguir una IA ética. Para ello profundizaremos sobre los siguientes contenidos:

1. **Propósitos globales** de IBM respecto de la IA ética.
2. **Principios éticos** sobre los que IBM se focaliza.
3. **Organización y recursos internos** de IBM para asegurar que sus productos utilizados por terceros cumplen con sus propósitos globales y principios éticos sobre los que IBM se focaliza.
4. Para cada principio ético, **el *toolkit* de IBM con las tecnologías y herramientas de gobierno** que permiten automatizar su gestión.

De esta manera iremos profundizando **desde una visión global a una aterrizada**. Los tres primeros puntos permitirán a cualquier entidad entender planteamientos globales que poder adoptar en su estrategia, su organización y procesos de gestión. El punto 4 permitirá a dichas entidades aplicar este enfoque en el día a día de sus operaciones, en el *Delivery* de sus proyectos que tengan a la inteligencia artificial como su tecnología principal.

Propósitos globales de IBM respecto de la IA ética

IBM se plantea tres grandes objetivos globales en IA ética que a lo largo de presente capitulo iremos aterrizando para permitir que cualquier entidad pueda asimilarlos ateniendo a la realidad de su día a día.

El propósito de la IA es aumentar la inteligencia humana, no sustituirla

En IBM, creen que la IA debería hacernos a todos mejores en nuestros trabajos, y que los beneficios de la era de la IA deberían llegar a muchos, no solo a unos pocos.

Los datos y el conocimiento alrededor de los mismos pertenecen a su creador

Teniendo en cuenta que las políticas de gobierno del dato deben ser justas, equitativas y priorizar la apertura, pero también su seguridad.

La tecnología debe ser transparente y explicable

Las empresas deben ser claras sobre cómo y con qué datos fueron entrenados sus sistemas basados en IA y, lo que es más importante, qué criterios se incluyeron en sus algoritmos de predicción. Aquellas empresas que no sean capaces de explicar las decisiones que toma la inteligencia artificial de sus productos no deberían operar en el mercado.

Principios éticos sobre los que IBM se focaliza

IBM se focaliza en los siguientes cinco principios éticos, entendidos con las siguientes definiciones:

1. **Explicabilidad**. La capacidad del sistema de IA para proporcionar acceso a una explicación interpretable por humanos para sus predicciones y conocimientos.
2. **Equidad**. Trato equitativo de individuos o grupos de individuos por un sistema de IA, teniendo presente que la equidad para un sistema de IA depende del contexto en el que se utiliza.
3. **Robustez**. La capacidad del sistema de IA para manejar condiciones excepcionales, como anomalías en la entrada, de manera eficaz.
4. **Transparencia**. La capacidad del sistema de IA para incluir y compartir información sobre cómo ha sido diseñado y desarrollado.
5. **Privacidad**. Capacidad para priorizar y salvaguardar la privacidad y los derechos de las personas cuyos datos maneja el sistema de IA.

Organización y recursos internos de IBM

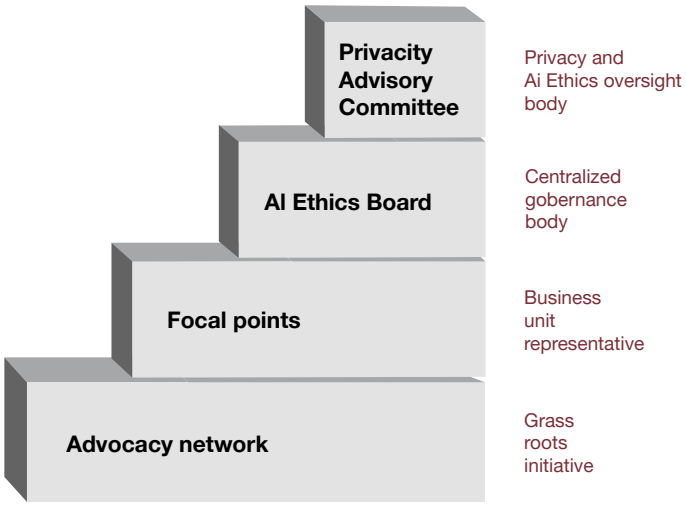
IBM dispone internamente de cuatro grandes pilares para asegurar que sus productos ofrecidos a las empresas cumplen con sus propósitos globales y principios éticos sobre los que se focaliza.

1. **Estructura de Gobierno interno.** IBM dispone desde 2018 de una junta de IA Ética presidida desde 2019 por Francesca Rossi (AI Líder Global de Ética) y Christina Montgomery (directora de Privacidad). Su misión es apoyar una gobernanza centralizada en el proceso de elaboración de los planteamientos éticos de IBM sobre políticas, prácticas, comunicaciones, investigación, productos y servicios. Además, por cada unidad de negocio existe un *focal point*, un embajador de la IA ética que tiene como objetivo crear una red que permita hacer permear el mensaje y trabajar con dichas unidades para evaluar si sus productos y servicios se adhieren a los principios definidos.

   Cuando las unidades de negocio locales y los *focal points* tienen preguntas o inquietudes durante sus evaluaciones éticas, la Junta ofrece asesoramiento o determina el plan de acción apropiado. Por último, y coordinados por esos focal points, está la red de personas interesadas en la IA ética.

2. **Formación interna**. IBM considera que la formación a sus empleados es un facilitador clave para la implementación y operacionalización de la IA ética. Para ello habilita programas formativos agrupados en un marco denominado *Everyday ethics for Artificial Intelligence* que son de obligatorio cumplimiento por parte de los empleados de IBM.

3. **Investigación**. *Trusted AI* es el área de *IBM Research* que trabaja en los enfoques para garantizar que los sistemas de IA que se creen en el futuro sean justos, robustos, explicables y alineados con los valores de la sociedad para la que están diseñados. Su objetivo es asegurar que en el futuro las aplicaciones de IA sean tan justas como eficientes a lo largo de todo su ciclo de vida. Este área trabaja sobre los propósitos globales y principios éticos concretos descritos anteriormente generando, entre otros *outputs*, los modelos de gobierno y tecnologías que permiten la gestión de los mismos y que veremos posteriormente en este capítulo.



Fuente: IBM

4. **Ethics by design**. IBM es consciente de que es fundamental contemplar los conceptos éticos desde el principio del proyecto, desde la conceptualización y el diseño inicial de cualquier sistema inteligente, contemplando las necesidades del usuario final de dicho sistema. Por eso focaliza un esfuerzo particular en esta fase mediante una metodología denominada *Ethics by Design* cuyos principales pilares veremos posteriormente en este capítulo.

El *output* de los dos últimos puntos son los orientados a las entidades que utilizan los productos y servicios de IBM, y son sobre los cuales **pondremos foco** para ayudar a las organizaciones en su objetivo: adoptar la IA de una manera ética.

Toolkit de IBM para la gestión de los principios éticos

En este apartado profundizaremos en las **tecnologías y herramientas de gobernanza de IBM** que permiten la **gestión automatizada de los principios éticos**.

Equidad

- Qué es la Equidad para IBM, y su relación con el sesgo.
- *IBM AI Fairness* 360, *opensource* para gestionar el principio ético de Equidad.
  - Cuáles son las métricas que nos permiten identificar automáticamente el sesgo.
  - Cuáles son los algoritmos proporcionados para mitigar el sesgo.

Explicabilidad

- Qué es la Explicabilidad para IBM.
- *AI Explainability* 360, *opensource* para gestionar el principio ético de Explicabilidad.
  - Qué algoritmo de Explicabilidad elegir en función de qué se necesita explicar.
  - Cuál es el algoritmo de Explicabilidad más adecuando para cada usuario.
  - El algoritmo de Explicabilidad más adecuando en cada fase del ciclo de vida.
  - Métricas para medir la calidad de la Explicabilidad.

Equidad y Explicabilidad

- *IBM Watson Openscale*, producto para los principios de Equidad y Explicabilidad.
- *IBM Causal Inference* 360, *opensource* para gestionar la inferencia Causal, un concepto relacionado con la Equidad y la Explicabilidad, aunque el fin último de este *toolkit* no sea gestionarlas.

Robustez

- Qué es la robustez para IBM.
- *IBM Adversarial Robustness* 360, *opensource* para gestionar la Robustez.
  - Qué es una muestra adversaria, el elemento clave para poder realizar un ataque a un modelo inteligente que ponga en jaque su Robustez.
  - Cómo gestionar ataques y defensas según el momento en el que se ataca o prueba el modelo.

Transparencia

- Qué es la transparencia para IBM.
- *Uncertainty Quantification* 360, para gestionar el principio ético de Transparencia.
  - Cómo estimar la incertidumbre en las predicciones de un modelo de ML.
  - Cómo evaluar la calidad de esas incertidumbres y, si es necesario, mejorarla.
  - Cómo comunicar esas incertidumbres a las personas que hacen uso del modelo.
- La *FactSheet*, herramienta de gobernanza para una IA transparente y confiable.

Privacidad

- Qué es la privacidad para IBM.
- *IBM AI Privacy* 360, para gestionar el principio ético de Privacidad.
  - Gestión de datos cifrados
  - Privacidad diferencial.
  - Anonimización.
  - Minimización de datos.
  - Evaluación del riesgo de privacidad.
  - Privacidad en el aprendizaje federado.

Principio de Equidad

En esta sección vamos a desarrollar:

- Qué es la Equidad para IBM, y su relación con el sesgo.
- *IBM AI Fairness* 360, *opensource* para gestionar el principio ético de Equidad.
  - Cuáles son las métricas que nos permiten identificar automáticamente el sesgo.
  - Cuáles son los algoritmos proporcionados para mitigar el sesgo.

Qué es la Equidad para IBM, y su relación con el sesgo

La equidad se refiere al trato equitativo de personas o grupo de personas por parte de un sistema de inteligencia artificial. **En caso de no cumplirse** es cuando se produce un **sesgo**. El **origen del sesgo está en los datos** y puede ser debido, entre otros, a **dos motivos principales**:

- **Datos desbalanceados**. Es decir, que no se hayan contemplado todos los posibles escenarios relacionados con las entidades de datos principales que forman parte de la muestra. Veámoslo con un ejemplo muy sencillo: si queremos crear un sistema que, en función de os datos históricos de los que disponemos, gestione nuestro stock de medicamentos y nos ayude a predecir cual es la provisión de inventario a realizar, no podremos contar únicamente con los datos de venta de los meses de verano, ya que en dicha predicción no incluirá `por ejemplo los antigripales. Este tipo de escenarios se gestionan de manera sencilla, abriendo la muestra a un conjunto de datos que incluya todas los casos posibles.

  Otro ejemplo sencillo de entender, y que tiene una mayor sensibilidad social, puede ser en el caso de un sistema de reconocimiento facial que es entrenado únicamente con imágenes de personas de raza blanca. Esto provocará un sesgo, ya que el sistema tendrá problemas a la hora de reconocer a personas de otras razas.

- **Prejuicios históricos y sociales**. Puede ser que, pese a tener una muestra amplia de datos que contemple todos los posibles casos, dichos datos lleven implícitos una serie de prejuicios históricos y/o sociales. Por ejemplo, Medicare (sistema público de salud en EEUU), desarrolló un sistema de diagnóstico para la depresión post-parto en mujeres. En base a los datos pasados, dicho sistema emitía pre-diagnósticos de esta patología con mayor frecuencia a mujeres de raza blanca. El origen del sesgo venía dado porque las mujeres de otras razas no suelen utilizar el sistema de salud para patologías de esta índole, lo que provoca que se dispongan de menos datos de mujeres de razas distintas a la blanca.

La propia naturaleza de un modelo de IA consiste en el concepto de discriminación estadística. Un concepto que busca patrones con los que diferenciar, agrupar y clasificar cada ejemplo de entrenamiento para etiquetar correctamente los nuevos ejemplos cuando el modelo está en producción. Por lo que, si nuestros conjuntos de datos tienen sesgos por motivos como los anteriormente expuestos, estos serán aprendidos y perpetuados.

Estas diferencias aprendidas pueden llegar a favorecer sistemáticamente ciertos grupos privilegiados y poner en desventaja sistemática a otros grupos no privilegiados, lo que supone un problema de igualdad y no es considerado ético ni legal en muchos contextos.

En resumen, hay que tener en cuenta que los datos de los que disponemos no tienen por qué representar el 100% de la “verdad”.

IBM AI Fairness 360, *opensource* para gestionar el principio ético de Equidad

IBM proporciona un *toolkit opensource* y un producto licenciado para gestionar la Equidad:

- *IBM AI Fairness* 360. Se puede utilizar en todo el ciclo de vida de la solución inteligente que se desea implantar, desde que está en desarrollo hasta que está en producción. Orientada a perfiles técnicos, es *opensource* y está desarrollado en *Python y R*.
- *IBM Watson Openscale*. Es un producto licenciado que tiene como base las librerías de *IBM AI Fairness* 360 a las que se ha añadido, entre otras capacidades, una interfaz visual que permite su uso por perfiles no necesariamente técnicos.

IBM AI Fairness 360 (IBM AIF 360)

Accesible en https://aif360.mybluemix.net/, esta es parte de la información y recursos relevantes que contiene:

- Una librería *OpenSource* en *Python* y R, publicada en *Github* (1) que incluye:
  - 75 métricas que sirven para detectar la falta de equidad (o el sesgo).
  - 13 algoritmos que permiten mitigar el sesgo, incluyendo taxonomías que ayudan a decidir porqué, cuándo y para quién utilizarlos.
- Vídeos, papers, glosario de términos.
- Tutoriales de casos de uso.
- Una demo que permite escoger un conjunto de datos, chequear mediante métricas los posibles sesgos sobre datos sensibles, y mitigar el sesgo mediante algoritmos.
- Solicitar acceso a una comunidad *slack* formada por más de 1.000 personas.

IBM Watson Openscale

Accesible públicamente en (2).

Es un producto licenciado que tiene como base las librerías de *IBM AI Fairness* 360 a las que, aparte de otras características, se ha añadido una interfaz visual que permite su uso por perfiles no necesariamente técnicos.

Permite analizar sesgos y aplicar soluciones para mitigarlo, y sus capacidades son explicadas posteriormente en este documento ya que, además del principio de Equidad, también permite la gestión del principio de Explicabilidad, que veremos posteriormente.

(1) https://github.com/Trusted-AI/AIX360/tree/master/aix360/algorithms)

(2) https://www.ibm.com/docs/en/cloud-paks/cp-data/3.5.0?topic=services-watson-openscale

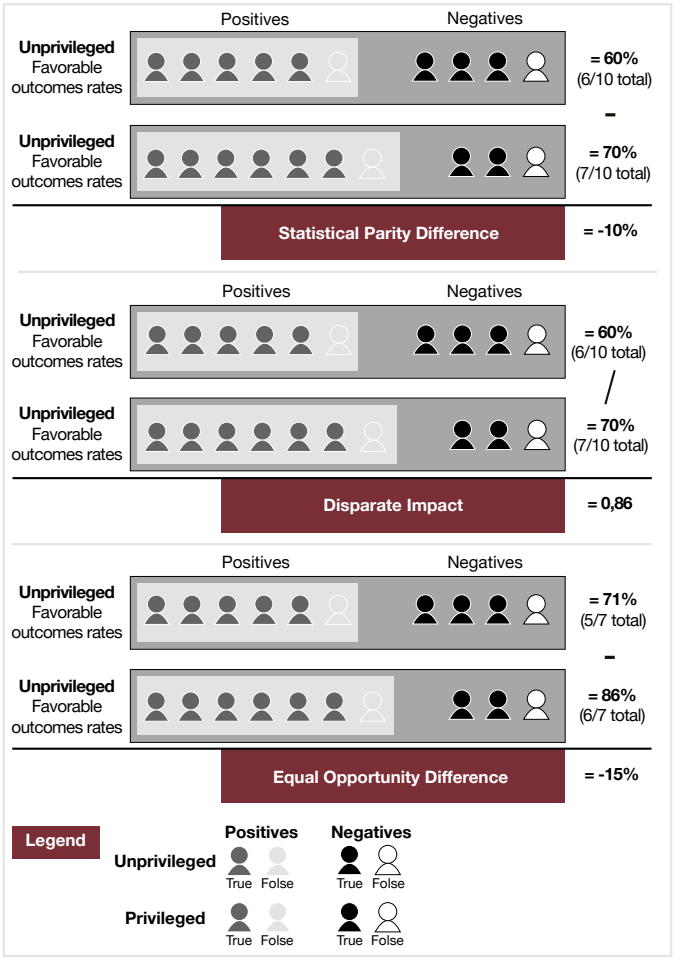
Métricas que permiten identificar que existe sesgo

Las métricas permiten identificar si, potencialmente, nuestro sistema inteligente está incurriendo en algún tipo de sesgo. Como hemos visto anteriormente, el sesgo puede tener dos tipos de origen generales (datos desbalanceados y prejuicios sociales), cuya manifestación particular en cada caso de uso puede ser muy diferente dentro de ambas tipologías.

Por tanto, es fundamental elegir la métrica más apropiada para cada contexto, y para ello es fundamental contar con expertos de negocio en el sistema inteligente cuyo proceso se desea automatizar.

El gráfico es un sencillo ejemplo de tres métricas de *Fairness*, que puede aplicar en cualquier caso de uso. Dichas métricas permiten identificar cuál es el ratio de equidad entre dos colectivos, uno que a priori es privilegiado y otro que no lo es.

- ***Statistical Parity Difference***. Refleja la diferencia de decisiones positivas del modelo para cada uno de los colectivos.
- ***Disparate impact***. Refleja de manera porcentual la diferencia de decisiones positivas del modelo para cada uno de los colectivos.
- ***Equal Opportunity Difference***. Refleja la diferencia en el ratio de acierto en la decisión del modelo sobre ambos colectivos.

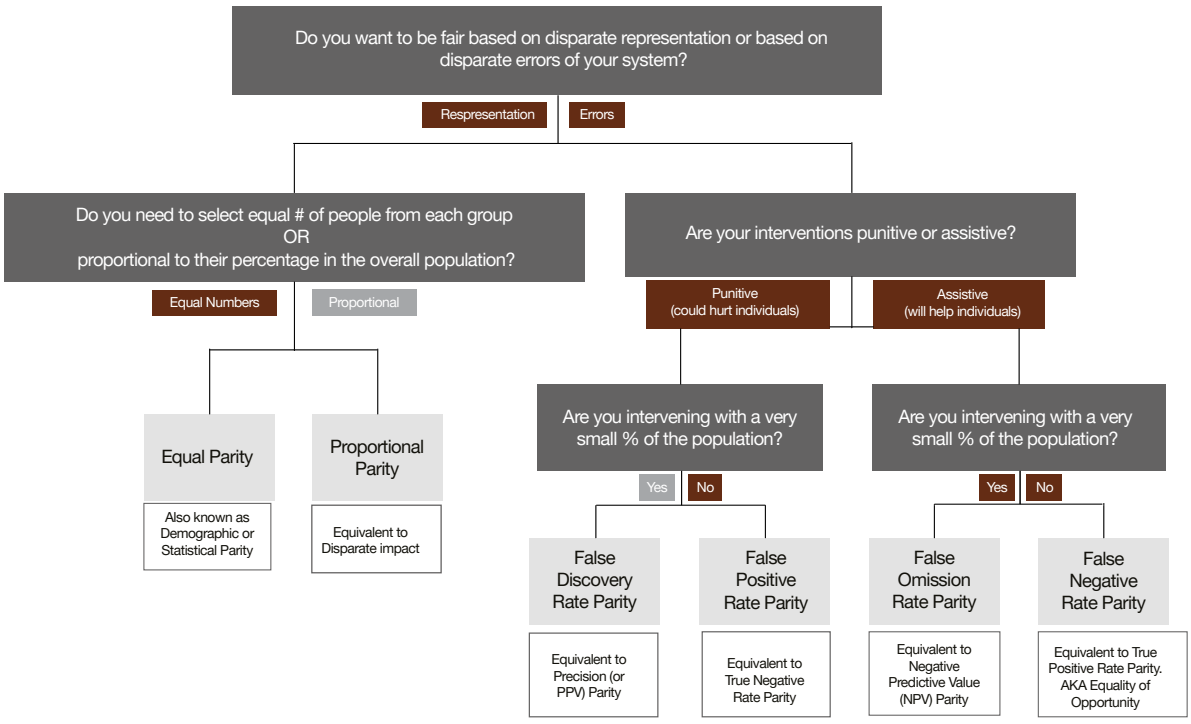


Fuente: IBM

El enfoque de IBM

IBM proporciona más de 70 métricas que permiten identificar si el modelo está incurriendo en algún tipo de sesgo. Es una cantidad importante de métricas, así que también proporciona un árbol de decisión que permite elegir las seis métricas más apropiadas para comenzar el análisis.

Dicho árbol de decisión pertenece al *framework Aequitas*, un conjunto de herramientas de auditoría de sesgos de código abierto desarrollado por el Centro de Ciencia de Datos y Políticas Públicas de la Universidad de Chicago.



Fuente: IBM

Este árbol de decisión permite elegir la métrica más adecuada a cada caso de uso. Las métricas no son excluyentes entre sí y están basadas principalmente en cuatro decisiones:

Centrarse en la representación de los distintos grupos que están incluidos en los datos:

- Buscar una igualdad cuantitativa => métrica *Equal parity*
- Buscar una igualdad proporcional => métrica *Proportional Parity*

Centrarse en lo justas que son las predicciones del modelo:

- Si la predicción o decisión del modelo pudiera provocar un daño a las personas, o privarla de algo que necesita. Por ejemplo, un crédito.
  - Si los datos incluyen a un % pequeño de la población => *False discovery rate parity*
  - Si los datos incluyen a un % alto de la población => *False Positive Rate Parity*
- Si la decisión del modelo supone una ayuda a las personas o su inclusión en un grupo con beneficios especiales (por ejemplo, pertenecer al grupo de clientes VIP de un banco es algo que no puede perjudicar, está orientado al beneficio).
  - Si los datos incluyen a un % pequeño de la población => *False Omission Rate Parity*
  - Si los datos incluyen a un % alto de la población => *False Negative Rate Parity*

Una vez seleccionada algunas de las seis métricas indicadas para comenzar, se podrá utilizar alguna de las restantes para comprobar si se ajusta mejor a la circunstancia que se estén evaluando.

Otra forma de analizar las métricas es dependiendo del contexto y los objetivos del modelo de IA atendiendo a los siguientes criterios:

Equidad de grupo vs. equidad individual, o ambas

La **equidad de grupo** divide a una población en grupos definidos por atributos protegidos y busca que alguna medida estadística sea igual entre las divisiones; por ejemplo, un banco utiliza un sistema de aprendizaje automático para decidir si otorgar un crédito a un cliente o no, por lo que para que no existiese ningún problema de equidad de grupo, las mujeres deberían tener las mismas probabilidades que los hombres de ser aptas para el préstamo. Por otro lado, la **equidad individual** busca que individuos similares sean tratados de manera similar independientemente del grupo al que pertenezcan. Es el caso de un algoritmo de reconocimiento facial para la identificación de delincuentes a través de imágenes, en el que por una falta de diversidad de imágenes en los datos de entrenamiento puede funcionar de manera menos precisa en el caso de datos demográficos distintos a los de raza blanca y traducirse en la detención indebida de alguien inocente por un delito que no ha cometido.

Si el modelo se preocupa por la equidad individual, se deben utilizar las métricas de la clase *SampleDistortionMetric*. En cambio, si el modelo se preocupa por la equidad de grupo, entonces de deben utilizar las métricas de la clase *DatasetMetric*, y su conjunto de subclases (como la clase *BinaryLabelDatasetMetric*), y las de la clase *ClassificationMetric*, con excepción de la que se indica a continuación. Pero si se da el caso de que la aplicación está relacionada con la equidad individual y grupal de manera simultánea, y requiere el uso de una única métrica, entonces se debe usar el índice de entropía generalizada con las métricas *Theil index y coefficient of variation* de la clase *ClassificationMetric*.

Equidad de grupo: datos vs. modelos

La equidad también puede medirse en los diferentes puntos en que se encuentren los datos de un modelo de aprendizaje automático (datos de entrenamiento y modelos). Si la aplicación requiere métricas sobre los datos de entrenamiento, se deben usar las métricas de la clase *DatasetMetric*, y su conjunto de subclases (como la clase *BinaryLabelDatasetMetric*). Y si, por el contrario, la búsqueda de sesgos requiere realizarse en la parte de los modelos, se deben utilizar las de la clase *ClassificationMetric*.

Equidad de grupo: todos somos iguales (*WAE – We're All Equals*) vs. lo que ves es lo que obtienes (*WYSIWYG – What You See Is What You Get*)

En el mundo existen dos visiones opuestas sobre le equidad grupal: todos somos iguales (WAE) y lo que ves es lo que obtienes (WYSIWYG). El enfoque WAE defiende que todos los grupos tienen habilidades similares con respecto a una tarea y el enfoque WYSIWYG considera que las observaciones reflejan la habilidad del individuo respecto a esa tarea. Por ejemplo, en el caso de las admisiones universitarias en Estados Unidos, utilizando la puntuación SAT (prueba estandarizada que se utiliza en las admisiones universitarias en EE. UU.) como una característica para predecir el éxito en la universidad. En el caso del enfoque WYSIWYG, este defiende que la puntuación se correlaciona directamente con el éxito futuro del solicitante y que hay una manera de usar dicha puntuación para comparar correctamente las habilidades de cada uno. Por el contrario, el enfoque WAE dice que la puntuación del SAT puede contener sesgos estructurales, por lo que su distribución, que es diferente entre los grupos, no debe confundirse con una diferencia en la distribución de las capacidades de los futuros universitarios.

Si el modelo sigue la visión WAE, entonces deben utilizarse las métricas de paridad demográfica: *disparate_impact* y *statistics_parity_difference*. Sin embargo, si la aplicación defiende la visión WYSIWYG, entonces se debe usar la métrica de igualdad de probabilidades: *average_odds_difference y averageabs_odds_difference*.

Equidad de grupo: ratios vs. diferencias

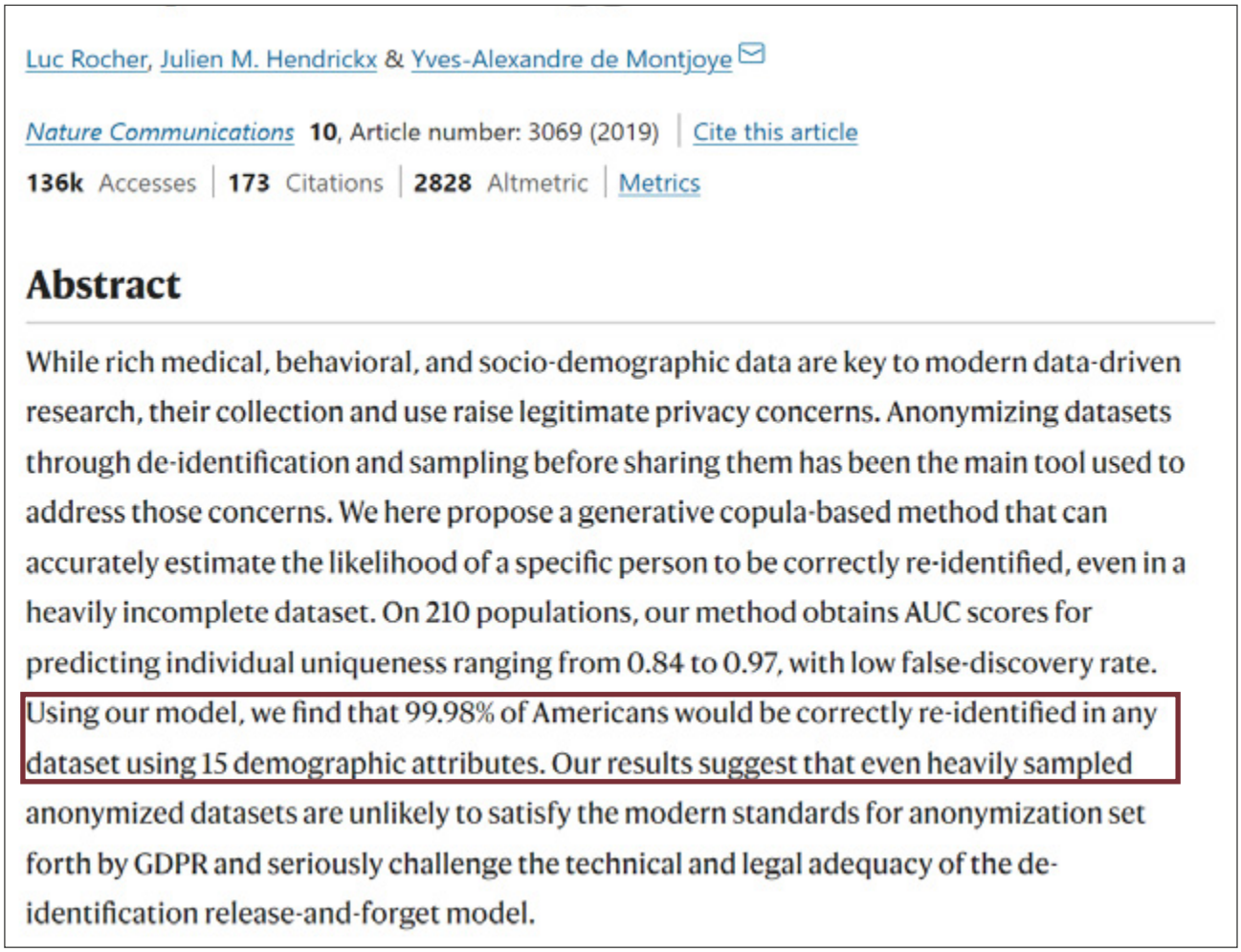
En la herramienta AIF360 todas las métricas nombradas anteriormente tienen tanto la versión de métricas de diferencia como la de métricas de proporción. La elección entre un tipo u otro debe basarse en la comodidad de los usuarios que examinan los resultados, ya que ambos tipos transmiten la misma información.

Algoritmos proporcionados para mitigar el sesgo

Las métricas nos ayudan a identificar los potenciales sesgos en nuestro conjunto de datos. Una vez detectado el potencial sesgo, el siguiente paso es tomar la decisión acerca de qué hacer para minimizar dicho sesgo.

Una alternativa planteada en estos escenarios es la de eliminar los atributos protegidos de nuestro conjunto de datos, los atributos sensibles que pueden ser origen del sesgo. Por ejemplo, si nuestro modelo está basado en personas, eliminar atributos como edad, sexo, raza, ubicación geográfica, etc.

Pero en muchos casos no es suficiente. Existen los denominados atributos *proxy*. Estos atributos permiten deducir información que no es explicita en nuestro conjunto de datos. En un artículo de 2019 se demostró que con 15 atributos demográficos anonimizados se podía acabar identificando individualmente a un 99,98% de la población norteamericana.



Luc Rocher, Julien M. Hendrickx & Yves-Alexandre de Montjoye

Nature Communications **10**, Article number: 3069 (2019) | Cite this article

136k Accesses | **173** Citations | **2828** Altmetric | Metrics

Abstract

While rich medical, behavioral, and socio-demographic data are key to modern data-driven research, their collection and use raise legitimate privacy concerns. Anonymizing datasets through de-identification and sampling before sharing them has been the main tool used to address those concerns. We here propose a generative copula-based method that can accurately estimate the likelihood of a specific person to be correctly re-identified, even in a heavily incomplete dataset. On 210 populations, our method obtains AUC scores for predicting individual uniqueness ranging from 0.84 to 0.97, with low false-discovery rate. Using our model, we find that 99.98% of Americans would be correctly re-identified in any dataset using 15 demographic attributes. Our results suggest that even heavily sampled anonymized datasets are unlikely to satisfy the modern standards for anonymization set forth by GDPR and seriously challenge the technical and legal adequacy of the de-identification release-and-forget model.

https://www.nature.com/articles/s41467-019-10933-3

Por tanto, eliminar atributos protegidos no es la solución. El objetivo es crear independencia estadística en los algoritmos de nuestro modelo que permitan minimizar el sesgo identificado. Para ello, IBM proporciona **trece algoritmos clasificados en tres categorías** que pueden ser utilizados en distintos momentos del ciclo de vida de las soluciones inteligentes (por orden):

Algoritmos de **pre-procesamiento** (*Pre-processing*). Orientados a la gestión de los datos de entrenamiento.	Algoritmos de **procesamiento interno** *(In-processing*). Orientados a la gestión del algoritmo de aprendizaje.	Algoritmos de **post-procesamiento** (*Post-processing*). Orientados a la gestión de la decisiones tomadas por el modelo.

La **elección de la categoría** se basa en la capacidad de la que disponga el equipo de desarrollo para intervenir en las fases de implementación del modelo.

- Si el equipo puede modificar los datos de entrenamiento, se puede utilizar el preprocesamiento.
- Si puede cambiar el algoritmo de aprendizaje, se puede utilizar el procesamiento interno.
- Si solo se puede tratar el modelo como una caja negra sin ninguna capacidad de modificarlo en las fases previas, solo podrá utilizar los de post-procesamiento.

IBM recomienda utilizar la categoría de algoritmo de mitigación más temprano para conseguir así una mayor flexibilidad y oportunidad para corregir el sesgo tan pronto como sea posible. Y si es posible, utilizar todos los algoritmos de todas las categorías permitidas, ya que el rendimiento final depende de la calidad del conjunto de datos.

Algoritmos de pre-procesamiento

- ***DisparateImpactRemover***. Edita los valores de las características, aumenta la equidad del grupo y conserva el orden de clasificación dentro de los grupos.
- **LFR**. Encuentra una representación latente que codifica bien los datos pero que confunde la información sobre los atributos protegidos.
- ***OptimPreproc***. Genera una transformación probabilística que edita las características y etiquetas en los datos con equidad de grupo, distorsión individual y restricciones, y objetivos de fidelidad de datos.
- ***Reweighing***. Pondera los ejemplos en cada combinación (grupo, etiqueta) de forma diferente para garantizar la equidad antes de realizar la clasificación.

Algoritmo de procesamiento interno

- ***AdversarialDebiasing***. Es una técnica que utiliza un clasificador para maximizar la precisión de la predicción y, al mismo tiempo, reducir la capacidad del adversario para determinar el atributo protegido a partir de las predicciones. Este enfoque conduce a un clasificador justo, ya que las predicciones no pueden llevar ninguna información de discriminación de grupo que el adversario pueda explotar.
- ***GerryFairClassifier***. Es un algoritmo para utilizar clasificadores que son justos con respecto a subgrupos ricos. Los subgrupos ricos se definen mediante funciones (lineales) sobre los atributos sensibles y las nociones de equidad son estadísticas: falsos positivos, falsos negativos y tasas de paridad estadística. Esta implementación usa un máximo de dos regresiones como un oráculo de clasificación sensible al costo y admite regresión lineal, máquinas de vectores de soporte, árboles de decisión y regresión del kernel.
- ***MetaFairClassifier***. El meta algoritmo aquí toma la métrica de equidad como parte de la entrada y devuelve un clasificador optimizado wrt.
- ***PrejudiceRemover***. El eliminador de prejuicios es una técnica en proceso que agrega un término de regularización consciente de la discriminación al objetivo de aprendizaje.

- ***ExponentiatedGradientReduction***. Reducción de gradiente exponencial para una clasificación justa. La reducción de gradiente exponencial es una técnica en proceso que reduce la clasificación justa a una secuencia de problemas de clasificación sensibles al costo, devolviendo un clasificador aleatorio con el error empírico más bajo sujeto a restricciones de clasificación justa.
- ***GridSearchReduction***. Es una técnica que se puede utilizar para una clasificación justa o una regresión justa. Para la clasificación, reduce la clasificación justa a una secuencia de problemas de clasificación sensibles al costo, devolviendo el clasificador determinista con el error empírico más bajo sujeto a restricciones de clasificación justa entre los candidatos buscados. Para la regresión, utiliza el mismo principio para devolver un regresor determinista con el error empírico más bajo sujeto a la restricción de la pérdida de grupo acotado.

Algoritmos de post-procesamiento

- ***CalibratedEqOddsPostprocessing***. Es una técnica que optimiza las salidas de puntuación del clasificador sobre calibrado para encontrar probabilidades con las que cambiar las etiquetas de salida con un objetivo de probabilidades igualadas.
- ***EqOddsPostprocessing***. Técnica que resuelve un programa lineal para encontrar probabilidades con las que cambiar las etiquetas de salida para optimizar las probabilidades igualadas.
- ***RejectOptionClassification***. Es una técnica que da resultados favorables a los grupos y resultados desfavorables a los grupos privilegiados en una banda de confianza alrededor del límite de decisión con la mayor incertidumbre.

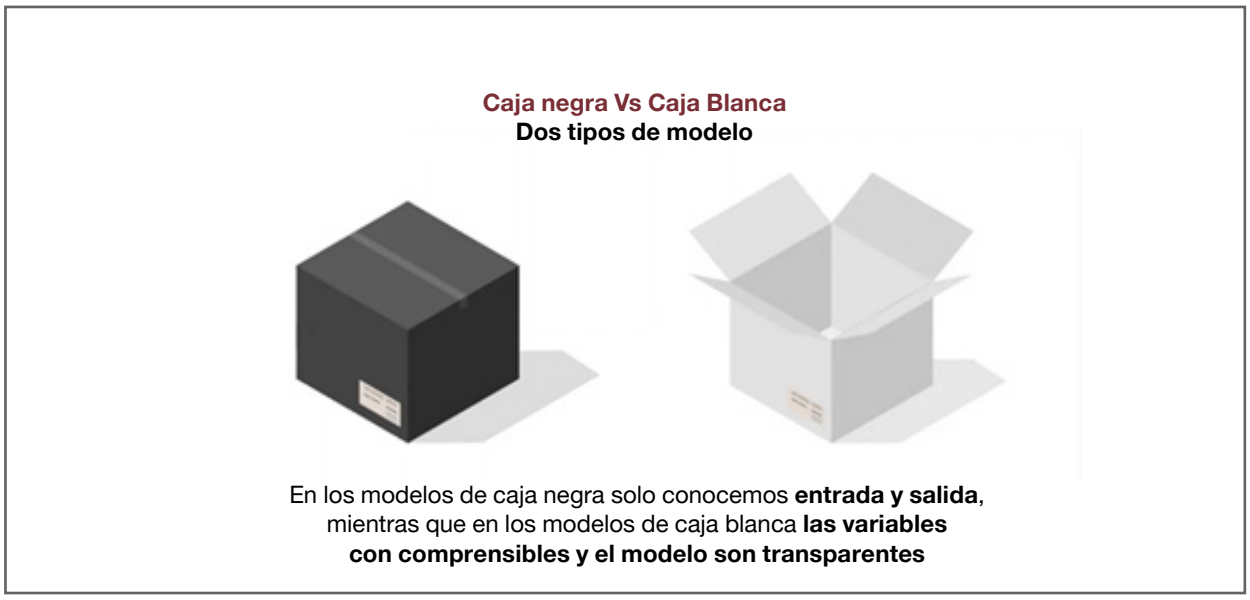
Principio de Explicabilidad

En esta sección vamos a desarrollar:

- Qué es la Explicabilidad para IBM.
- *IBM AI Explainability* 360, *opensource* para gestionar el principio ético de Explicabilidad.
  - Qué algoritmo de Explicabilidad elegir en función de qué se necesita explicar.
  - Cuál es el algoritmo de Explicabilidad más adecuando para cada usuario.
  - El algoritmo de Explicabilidad más adecuando en cada fase del ciclo de vida del modelo.
  - Métricas para medir la calidad de la Explicabilidad.

Qué es la Explicabilidad para IBM

Los modelos de aprendizaje automático (tanto los de caja blanca como los de caja negra -redes neuronales) están logrando una precisión impresionante en varias tareas. Sin embargo, a medida que el aprendizaje automático se usa cada vez más para informar decisiones de alto riesgo, la explicabilidad y la interpretabilidad de los modelos se vuelven esenciales.



Fuente: IBM

Le explicabilidad es la capacidad de comprender cómo y porqué un modelo de inteligencia artificial realiza una predicción y puede llegar a tomar una decisión a partir de la misma.

IBM AI Explainability 360, *opensource* para gestionar el principio ético de Explicabilidad

IBM proporciona las siguientes tecnologías para gestionar la Explicabilidad:

- ***IBM AI Explainability* 360**. Se puede utilizar en todo el ciclo de vida de la solución inteligente que se desea implantar, desde que está en desarrollo hasta que está en producción. Orientada a perfiles técnicos, es *opensource* y está desarrollado en *Python*.
- ***IBM Watson Openscale***. Es un producto licenciado que tiene como base las librerías de *IBM AI Explainability* 360 a las que se ha añadido, entre otras funcionalidades, una interfaz visual que permite su uso por perfiles no necesariamente técnicos.

El enfoque de IBM

IBM AI Explainability 360 (IBM AIE 360)

Accesible en https://aix360.mybluemix.net/

Esta es parte de la información y recursos relevantes que contiene:

- Una librería *OpenSource* publicada en *GitHub* (1) que incluye:
  - Acceso al API con toda su funcionalidad (2).
  - Ocho algoritmos que permiten explicar datos y modelos, incluyendo taxonomías que ayudan a decidir porqué, cuándo y para quién utilizarlos.
  - Dos métricas que permiten medir la calidad de la explicabilidad.
- Vídeos, papers, glosario de términos.
- Tutoriales de casos de uso.
- Una demo que permite explorar la explicabilidad de un modelo de concesión de crédito desde la perspectiva de diferentes usuarios del mismo (el *Data Scientist* responsable de la implementación del modelo, el representante del banco que gestiona la concesión del crédito, y el usuario final que requiere el crédito).
- Solicitar acceso a una comunidad *slack* formada por más de 1.000 personas.

(1) https://github.com/Trusted-AI/AIX360

(2) https://aix360.readthedocs.io/en/latest/

IBM Watson Openscale

Este producto está orientado para cuando la solución está en producción. Es un producto licenciado que tiene como base las librerías de *IBM AI Explainability* 360, a las que se ha añadido una interfaz visual que permite su uso por perfiles no necesariamente técnicos.

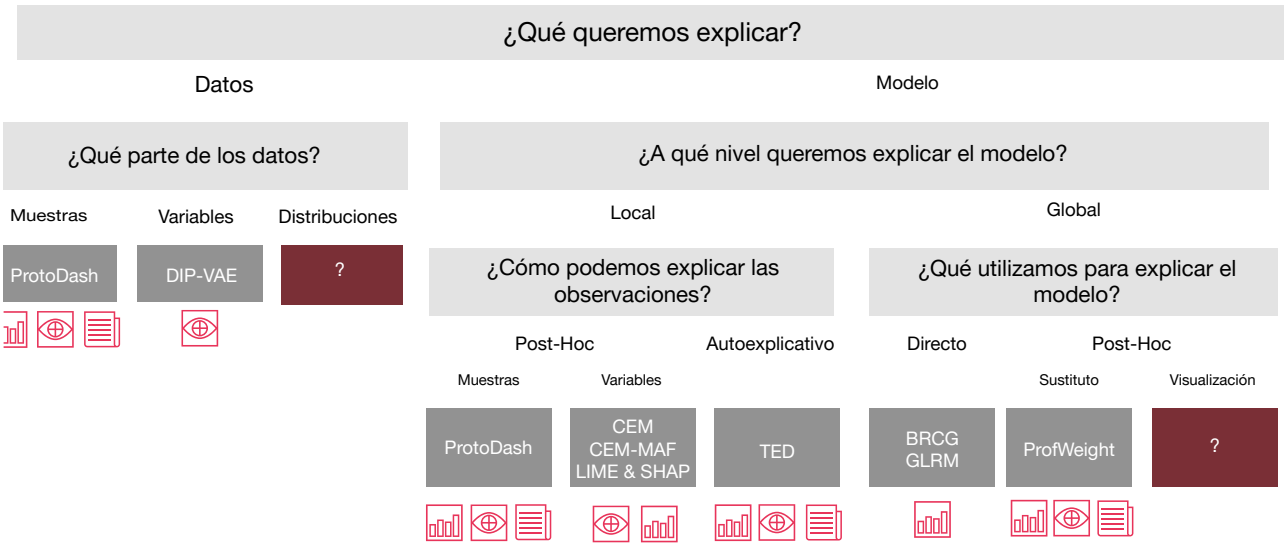
Permite rastrear y auditar predicciones de inteligencia artificial realizadas en aplicaciones de producción. La confianza en los pronósticos de modelo y los factores que contribuyen a los resultados finales, o a resultados distintos, se muestran en términos empresariales comprensibles.

Además de para el principio de explicabilidad aquí explicado, proporciona funcionalidades para la gestión del principio ético de Fairness (equidad) explicado anteriormente en este documento, y así detectar posibles sesgos y la mitigación de los mismos.

Sus capacidades serán explicadas en este mismo documento, a continuación del principio ético de Explicabilidad.

Qué algoritmo de explicabilidad elegir en función de qué se quiere explicar

Para la elección del algoritmo, IBM proporciona la siguiente taxonomía, que en un primer nivel permite diferenciar entre algoritmos orientados a la explicación de los datos y los orientados a la explicación del modelo. En el caso de querer explicar el modelo, diferencia entre explicabilidad del modelo en su globalidad y la explicabilidad de un caso concreto.



Fuente: IBM

Explicar los datos

El aprendizaje automático comienza con los datos. A menudo es útil comprender las características de los datos antes de que el modelo lleve a cabo cualquier aprendizaje.

Estos algoritmos están por tanto orientados a los datos de entrenamiento, antes de haber implementado el modelo.

No aporta de manera directa a la explicabilidad del modelo, que es el principio ético que estamos analizando, pero el beneficio de utilizarla está en que, **si se entienden bien los datos**, se pueden explicar mejor los resultados que posteriormente proporcione dicho modelo.

El *toolkit IBM AIE* 360 proporciona varios algoritmos que **permiten entender los datos de diferentes maneras**:

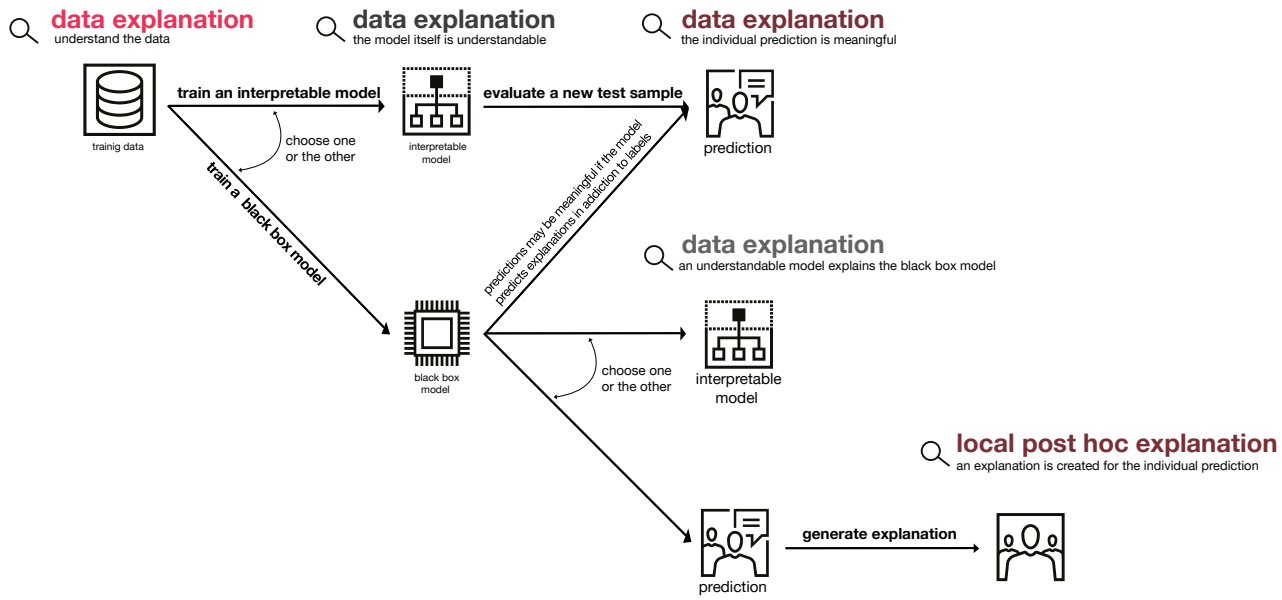
- **Analizando muestras**. El análisis de las muestras de un conjunto de datos permite conocer la esencia de un conjunto de datos (prototipo) y los valores críticos o atípicos del mismo. El algoritmo proporcionado es *ProtoDash*. Segmenta los datos, generando agrupaciones de datos equivalentes. Se puede aplicar sobre datos estructurados, imágenes y lenguaje natural.
- **Analizando variables**. A veces, las características en un conjunto de datos dado son visibles a primera vista, pero otras veces están entrelazadas. Es decir múltiples atributos significativos se combinan en una sola característica. El ***Autoencoder Variacional* Previo Inferido Desenredado (DIP-VAE)** es un algoritmo de aprendizaje de representación no supervisado que tomará las características dadas y aprenderá una nueva representación que se desenredará de tal manera que las características resultantes sean comprensibles. Es utilizado para extraer características de imágenes.

El enfoque de IBM

Explicar el modelo

A la hora de explicar un modelo debemos tener en cuenta que dicho modelo puede ser:

- Un modelo de caja blanca, directamente interpretable a priori (regresiones o árboles de decisión, por ejemplo).
- Un modelo de caja negra (típicamente una red neuronal), interpretable únicamente a posteriori (*post-hoc*) a través de la observación de las decisiones que haya tomado.



Fuente: IBM

Explicar modelos interpretables.

Son los denominados modelos de caja blanca. Son modelos basados en regresiones o árboles de decisión, por ejemplo. Este tipo de modelos son utilizados cuando es necesario comprender internamente en la organización todo el proceso de toma de decisiones y garantizar su seguridad y confiabilidad, ya sea por la criticidad del proceso y/o por la necesidad de explicar el funcionamiento de dicho modelo a un auditor externo.

Para este tipo de modelos, puede ser necesario explicar cómo se comporta globalmente el modelo a la hora de realizar predicciones, o como se ha comportado para una predicción concreta:

- **Explicar el modelo de manera global**. Se proporcionan los dos siguientes algoritmos:
  - **BRCG** (*Boolean Decision Rules via Column Generation*). Utilizado en modelos con lógica booleana.
  - **GLRM** (*Generalized Linear Rule Models*). Se puede utilizar con modelos de regresión.

  Ambos pueden ser utilizados para explicar modelos de *clustering*
- **Explicar una predicción concreta** (local). Se proporciona el algoritmo TED (*Teaching AI to Explain Its Decisions*). Este algoritmo permite explicar la decisión mediante su clasificación en una categoría que previamente ha sido descrita con lenguaje natural.

Explicar modelos de caja negra, mediante algoritmos post-hoc.

Los modelos de caja negra son habitualmente implementados mediante redes neuronales, cuya explicación no es directa ya que no se sabe con precisión cuál es el proceso que ha seguido el modelo para tomar la decisión.

La explicación de este tipo de modelos se realiza construyendo un modelo adicional que se encarga de realizar la explicación sobre el modelo de caja negra.

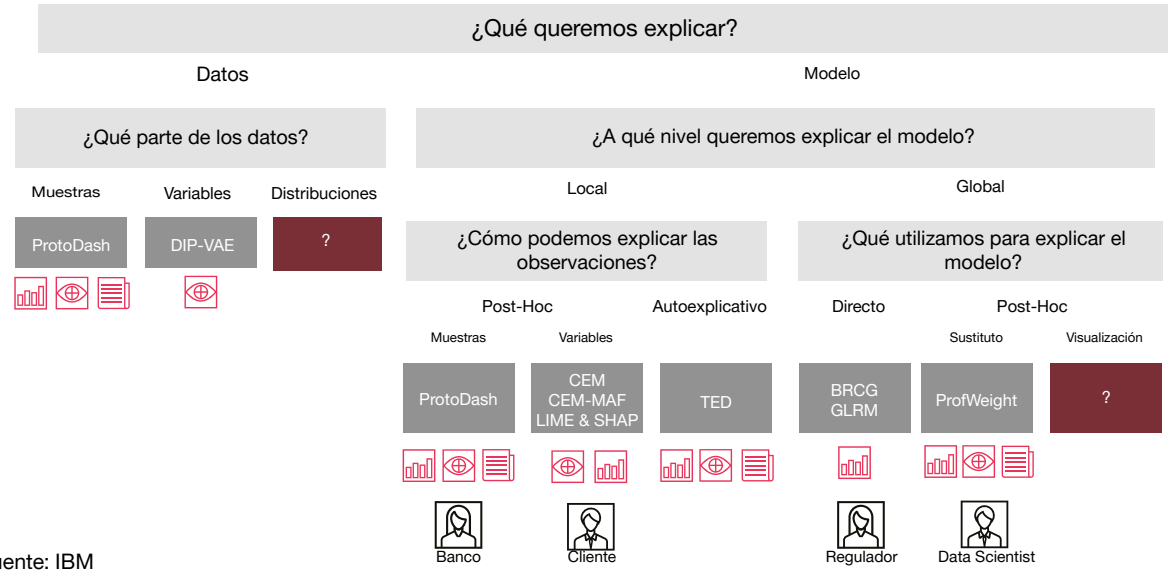
Al igual que para los modelos directamente explicables, para este tipo de modelos puede ser necesario explicar cómo se comporta globalmente el modelo a la hora de realizar predicciones, o como se ha comportado para una predicción concreta:

- **Explicar el modelo de manera global**. Se proporciona el algoritmo ProfWeight. Este algoritmo sondea la red neuronal y produce pesos de instancia que luego se aplican a los datos de entrenamiento para generar un modelo directamente interpretable.
- **Explicar una predicción concreta** (local). Hay tres algoritmos disponibles:
  - **ProtoDash** permite explicar a partir de muestras, comparando la predicción concreta con otras realizadas en la misma muestra a la que pertenece. Permite identificar anomalías en la predicción y así analizar posteriormente el motivo por el que haya podido ocurrir.
  - **CEM** (*Contrastive Explanations Method*), permite introducir perturbaciones en las variables de las mediciones y así explicar cuándo se produce el cambio en la predicción.
  - **CEM-MAF** (*Contrastive Explanations Method with Monotonic Attribute Functions*), es exactamente igual que CEM, pero orientado a imágenes.
  - **LIME & SHAP**. Es un estándar que ofrece las mismas capacidades de los dos anteriores, y también incluido por IBM en su *toolkit*.

El enfoque de IBM

Cuál es el algoritmo de explicabilidad más adecuando para cada usuario

Utilizaremos como ejemplo una solución de *Machine Learning* que predice el nivel de riesgo de un cliente a la hora de concederle o denegarle un préstamo. En este ejemplo definiremos varios perfiles que tienen diferentes objetivos.



Fuente: IBM

El Data Scientist

Como creador y responsable técnico del despliegue de la solución, quiere obtener la máxima precisión posible, y necesita conocer qué es lo que está haciendo el modelo, cuál es la explicabilidad de las decisiones que está tomando. Utilizará por tanto todos los algoritmos de explicabilidad posibles, tanto los globales para sus pruebas unitarias como los locales para las pruebas integradas y de validación final del modelo.

El director de la sucursal bancaria

Da por hecho que la precisión del modelo es buena. Quiere asegurar que las predicciones que le está proporcionando la solución son homogéneas en clientes con características similares. Utilizará por tanto algoritmos de explicabilidad locales.

El cliente que ha solicitado el préstamo

Necesita una explicación que le permita conocer cuáles son los motivos por los cuales no se le ha concedido el préstamo y saber cuáles son los cambios necesarios en el valor de sus variables analizadas para que el sistema le conceda el préstamo. Los algoritmos necesarios para darle esa explicación son los locales.

Un regulador

Los bancos necesitan explicar las decisiones tomadas por sus soluciones inteligentes No solo internamente, si no también cara a reguladores que requieren dicha explicación. Esto ocurre en todos los sectores cuando se tratan datos personales, pero especialmente en aquellos sectores globalmente regulados (banca, seguros, etc.). El regulador requiere información acerca de la precisión del sistema, así como una explicación acerca de todas las decisiones tomadas.

Por tanto, cuando implementamos una solución inteligente que va a estar sujeta a regulación, como en este caso:

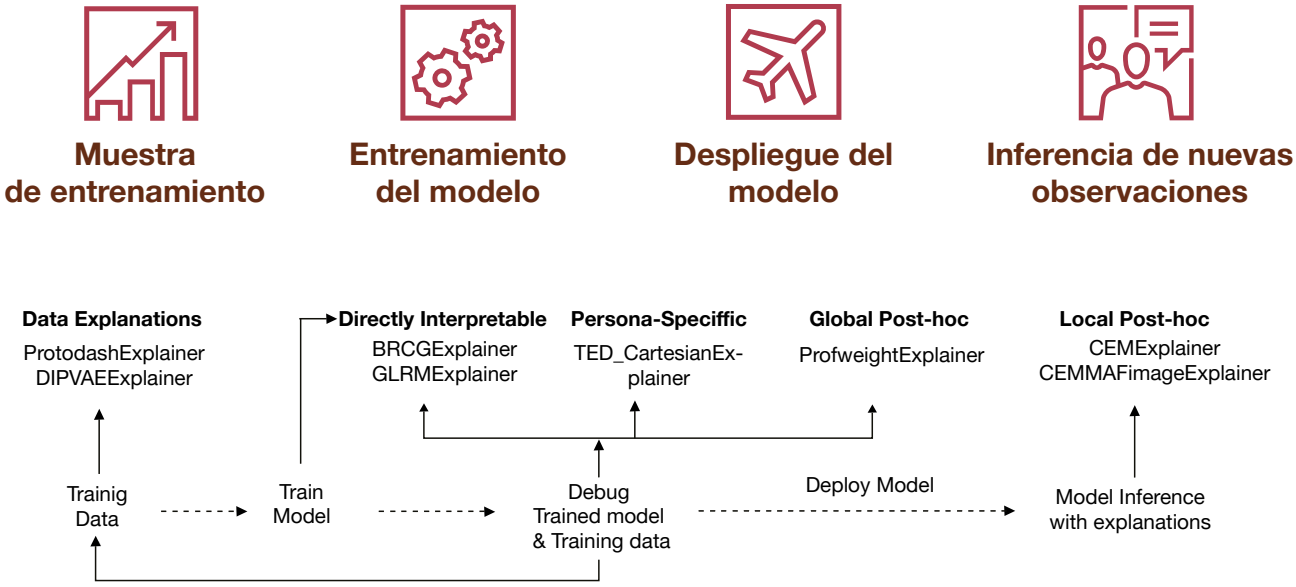
- Tendremos que utilizar modelos directamente interpretables.
- Serán necesarios tanto algoritmos globales para explicar el comportamiento general del modelo, como algoritmos locales para explicar cada decisión tomada de manera automatizada.

En función de esta descripción de objetivos, y las capacidades de los algoritmos de explicabilidad detallados en el anterior punto, podemos exponer que esta es la distribución de algoritmos de explicabilidad más adecuados para cada usuario.

Cuál es el algoritmo de explicabilidad más adecuando para cada fase del ciclo de vida del modelo.

Los algoritmos abordados anteriormente se pueden también explicar desde el punto de vista del ciclo de vida de las soluciones inteligentes, desde el inicial análisis de los datos, pasando por el entrenamiento del modelo y el despliegue del mismo, hasta que dicho modelo está en producción y genera nuevas observaciones y decisiones.

El siguiente esquema es un resumen del uso de algunos de dichos algoritmos en función de la fase del ciclo de vida en el que nos encontremos.



Fuente: IBM

Métricas para medir la calidad de la explicabilidad

La explicablidad no es un principio ético cuya calidad se pueda medir de manera cuantificada y objetiva, ya que depende mucho del receptor de dicha explicación. Esto lo diferencia de otros principios éticos como el *Fairness*, para cuya gestión IBM proporciona más de 70 métricas que permiten su identificación y mitigación.

Para el principio de Explicabilidad que ahora nos ocupa, IBM define dos métricas:

Fidelidad.

Calcula la correlación entre la importancia de las variables definidas en el algoritmo de explicabilidad y la importancia de dichas variables en el modelo que se está intentando explicar.

Monotonocidad

Valida que la precisión del modelo aumenta al introducir las variables por orden de importancia.

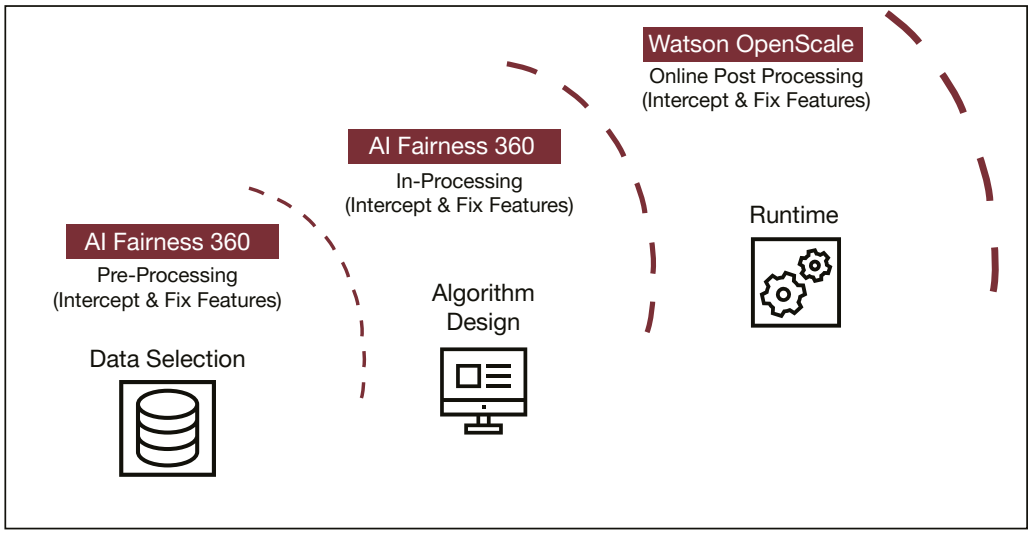
Equidad y Explicabilidad

IBM dispone de dos herramientas que facilitan la gestión de los principios de Equidad y Explicabilidad:

- *IBM Watson Openscale*, producto para los principios de Equidad y Explicabilidad.
- *IBM Causal Inference* 360, *opensource* para gestionar la inferencia Causal, un concepto relacionado con la Equidad y la Explicabilidad, aunque el fin último de este *toolkit* no sea gestionarlas.

IBM Watson Openscale

Dispone de funcionalidades para la gestión de los principios de Equidad y Explicabilidad. Es un producto licenciado que tiene como base las librerías *opensource* de *IBM AI Fairness* 360 e *IBM AI Explainability* 360 a las que se ha añadido, entre otras funcionalidades, una interfaz visual que permite su uso por perfiles no necesariamente técnicos (1).



Fuente: IBM

Su motor *opensource* le capacita para ser desplegado en casi cualquier infraestructura, permitiendo monitorizar modelos:

- Desarrollados en IDEs de terceros.
- Bajo un *framework opensource* alojado en la infraestructura de un tercero.
- Por un motor inteligente alojado en infraestructura privada.

Model build/train frameworks

TensorFlow · scikit-learn · Keras · PYTORCH · Caffe2 · Spark

Model serving environments

kubernetes · aws · Azure ML

(1) https://www.ibm.com/blogs/journey-to-ai/2020/04/comparing-ibm-watson-openscale- to-open-source-on-ai-explainability/

Equidad

OpenScale permite cumplir la equidad en un modelo mediante el análisis de transacciones en producción, encontrando potenciales comportamientos sesgados. Identifica la fuente del sesgo y mitiga activamente los sesgos encontrados en el entorno de producción aportando el siguiente valor añadido:

- Recomienda automáticamente atributos protegidos para monitorizar en producción.
- Detecta sesgos en preproducción y en tiempo de ejecución para analizar su impacto, pudiendo mitigarlos.
- Proporciona métricas y datos para ayudar a los científicos de datos a solucionar problemas en conjuntos de datos o modelos.

Explicabilidad

OpenScale registra cada transacción individual y profundiza en su funcionamiento para explicar cómo el modelo toma decisiones, proporcionando una explicación simple que es interactiva y fácil de usar, y aportando el siguiente valor añadido:

- Explicar las decisiones a nivel de transacciones individuales tomadas por el modelo en tiempo de ejecución, incluidos detalles sobre los atributos importantes y sus valores para ayudar en situaciones de cumplimiento normativo y de atención al cliente.
- Analizar transacciones individuales de manera hipotética para comprender cómo cambiará el comportamiento del modelo en diferentes situaciones.

IBM Causal Inference 360

Comprender la causa y el efecto es el objetivo último de cualquier investigación científica. Solo mediante tal comprensión se puede explicar verdaderamente un fenómeno y garantizar que las acciones produzcan los resultados previstos. El modelado causal es crucial para gestionar un modelo que genere confianza, sea equitativo, y garantice una toma de decisiones sólida y explicable.

La inferencia causal es una disciplina que consiste en estimar el efecto de una acción, en lugar de identificar la causa de algo. Un ejemplo de tal pregunta puede ser "¿cómo afectaría la decisión de dejar de fumar a mi peso en 10 años?". En el mundo real la inferencia causal es complicada porque solo podemos observar uno de dos futuros posibles: un objeto recibió un tratamiento o no recibió un tratamiento. El resultado que se observa es solo el resultado dado el tratamiento observado. El resultado que se habría obtenido si se le hubiera dado el tratamiento contrario se denomina resultado contra fáctico.

Por tanto, la Inferencia Causal es un concepto que tiene relación con los principios éticos definidos por IBM como Equidad y Explicabilidad, aunque el fin último de IBM *Causal Inference* 360 no sea directamente gestionarlas

Una complicación adicional surge del hecho de que la decisión de que un objeto reciba un tratamiento a menudo depende de muchos factores diferentes que también están relacionados con el resultado, lo cual puede generar confusión acerca del resultado obtenido. En el ejemplo anterior la decisión de dejar de fumar puede ser impulsada por un cambio a un estilo de vida más saludable, cambiando no solo el hábito de fumar, sino también su dieta y sus perfiles de actividad física. Esto confundiría nuestra estimación de cuánto al peso el hecho de fumar, en comparación con cómo lo hacen estos otros factores.

¿Cuáles son las alternativas?

Los experimentos aleatorios a menudo se describen como un estándar para estimar el efecto causal de los datos. Sin embargo, los experimentos aleatorios pueden ser prohibitivamente costosos, poco éticos o simplemente inviables. Por lo tanto, se deben aplicar algoritmos sofisticados a los datos de observación. La inferencia causal es diferente al problema del aprendizaje automático supervisado, pero muchas de las herramientas del aprendizaje supervisado, como el aprendizaje profundo, pueden ser componentes útiles de los algoritmos de inferencia causal debido a su capacidad para manejar datos de alta dimensión a gran escala.

Cabe señalar que la inferencia causal solo es válida bajo un conjunto de suposiciones que no pueden evaluarse a partir de los datos en sí (por ejemplo, que se miden todas las variables de confusión importantes). Como resultado, no hay garantía de que un modelo con buen desempeño proporcione una estimación real y precisa del efecto. Sin embargo, se garantiza que los modelos con peor rendimiento proporcionarán resultados poco fiables, e *IBM Research Causal Inference* 360 permite identificar y eliminar dichos modelos.

El enfoque de IBM

El toolkit IBM Causal Inference 360

El paquete *IBM Causal Inference* 360, desarrollado sobre *Python*, proporciona un conjunto de métodos, bajo una API unificada, para estimar el efecto causal de una intervención en un resultado utilizando datos de observación del mundo real. Lo hace mediante la implementación de varios meta-algoritmos que permiten conectar algoritmos de aprendizaje automático complejos para proporcionar modelos de inferencia causal altamente flexibles.

Accesible en https://cif360-dev.mybluemix.net/

Este paquete proporciona:

- Una demostración interactiva que permite recorrer dos casos de uso de ejemplo y comparar los resultados de un análisis causal con uno no causal (1).
- Tutoriales con contenidos orientados a los científicos de datos. (2).
- Un conjunto de datos y eliminar modelos con mal rendimiento.

Orientación sobre la elección de métodos

IBM Causal Inference 360 incluye varios métodos de estimación, cada uno de los cuales puede utilizar varios modelos de *Machine Learning*. Esta flexibilidad plantea la cuestión de qué combinación produce el mejor modelo causal general. Debido a la naturaleza de la predicción contrafactual, no se puede evaluar los modelos contra la verdad sobre el terreno como se hace generalmente en el aprendizaje automático. Sin embargo, se pueden discutir los pros y los contras de diferentes estrategias de modelado y proporcionar herramientas para eliminar candidatos de modelo de bajo rendimiento. Para ello, seguiremos el siguiente flujo de trabajo (4):

(1) https://cif360-dev.mybluemix.net/demo
(2) https://cif360-dev.mybluemix.net/resources#tutorials
(3) https://github.com/IBM/causallib/blob/master/docs/source/index.rst
(4) descrito en https://cif360-dev.mybluemix.net/resources#guidance y https://arxiv.org/abs/1906.00442

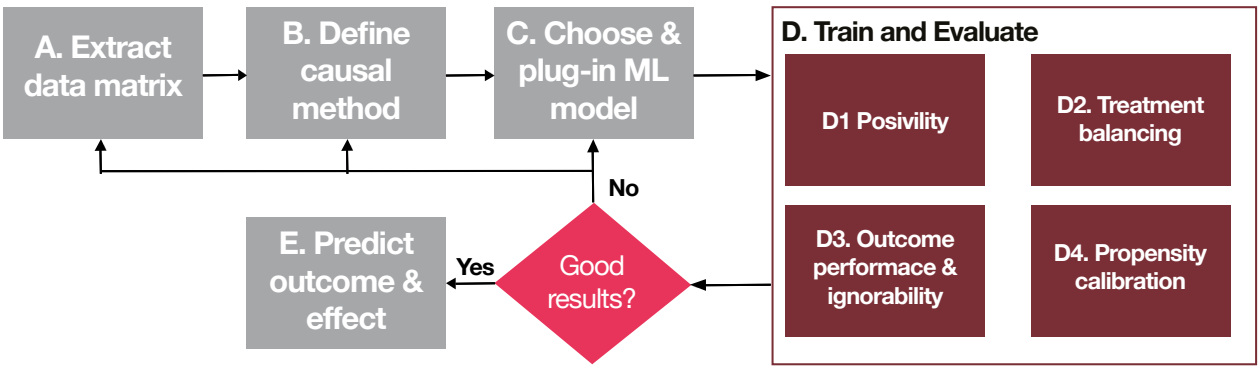


Figure 1: A schematics of the pipeline to guide model selection and cohort definition in casual inference. The pipeline involves an iterative process, in which a) the causal inference defined and a data matix is extracted; b) the causal methods in chosen; c) the underlying machine learning models are chosen; d) the model performance is evaluated. If the models perform well then causal inference prediction can be drawn to estimate outcome and effect. Otherwise, the process need to be reiterated following some modifications in steps a-c.

Fuente: IBM

Principio de Robustez

En esta sección vamos a desarrollar:

- Qué es la robustez para IBM.
- *IBM Adversarial Robustness* 360, *opensource* para gestionar el principio ético de Robustez.
  - Qué es una muestra adversaria, el elemento clave para poder realizar un ataque a un modelo inteligente que ponga en jaque su robustez.
  - Cómo gestionar ataques y defensas según el momento en el que se ataca o prueba el modelo.

Qué es la robustez para IBM

La robustez de un modelo de inteligencia artificial es la capacidad que tiene dicho sistema para:

- Asegurar sus predicciones, evitando equivocaciones que tengan un impacto en los procesos de negocio soportados por dicho modelo.
- Defenderse frente a atacantes que quieren modificar el *output* del modelo en su propio beneficio o provocar un mal funcionamiento del mismo.

El hecho de que un modelo inteligente contemple esta característica asegura su buen funcionamiento en casos de uso como los ejemplos mostrados a continuación:

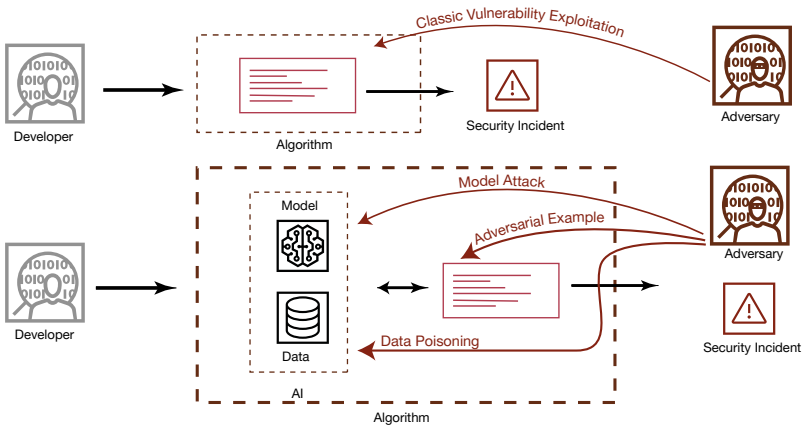
- **Reconocimiento facial**, evitando errores en casos de uso soportados por dicho modelo (control de accesos, seguridad y vigilancia, etc.).
- Funcionamiento de **vehículos autónomos o maquinaria industrial**, evitando accidentes.
- **Sistemas de seguridad**, eliminando fallos en el sistema que puedan ser aprovechados por *malware*.
- Análisis de **tendencias de opinión y reputación en RRSS**, impidiendo que las personas que las utilizan y los sistemas inteligentes que realizan dicho análisis puedan ser engañados.
- Evitar que un sistema de gestión de **fondos financieros** se equivoque en sus recomendaciones de inversión, o que un sistema de **detección de fraude** pueda ser sorteado.
- Identificar modificaciones de imágenes y descripciones tergiversadas en la **cobertura de seguros**.

La robustez en un modelo inteligente se consigue entrenando a dicho modelo ante los potenciales ataques y estableciendo mecanismos que le permitan identificarlos y gestionarlos cuando se produzcan.

La forma de atacar a un modelo inteligentes es diferente a la utilizada con un sistema de software tradicional. Un sistema de software tradicional está basado en reglas y el foco del ataque puede ser denegar su funcionamiento o robar información del mismo. Sin embargo, un modelo inteligente está diseñado para aprender y realizar predicciones a partir de dicho aprendizaje, por lo que un ataque puede tener dos consecuencias adicionales a las de un sistema de software tradicional:

- Tergiversar su aprendizaje, provocando predicciones o decisiones incorrectas.
- Engañar al modelo y forzar la obtención de un resultado (una predicción o incluso una decisión) que no debiera ser la realizada en condiciones normales.

Ataque clásico vs ataque a un modelo de AI



Fuente: IBM

IBM Adversarial Robustness 360, *opensource* para gestionar el principio ético de Robustez

***IBM Adversarial Robustness Toolbox* (ART)** es una biblioteca *opensource* desarrollada en *Python* para la robustez del aprendizaje automático.

ART proporciona herramientas que permiten a los desarrolladores e investigadores evaluar, defender, certificar y verificar modelos y aplicaciones de *Machine Learning* contra las amenazas adversas de evasión, envenenamiento, extracción e inferencia que veremos posteriormente.

ART es compatible con los *frameworks* de aprendizaje automático más utilizados (TensorFlow, Keras, PyTorch, MXNet, scikit-learn, XGBoost, LightGBM, CatBoost, GPy, etc.), todos los tipos de datos (imágenes, tablas, audio, video, etc.) y aprendizaje automático en tareas de clasificación, detección de objetos, generación, certificación, etc.

Accesible en https://art360.mybluemix.net/?_ga=2.19480378.325754436.1642231065-1367751274.1637864447

Esta es parte de la información y recursos relevantes que contiene:

- Una librería *OpenSource* publicada en GitHub (1)
- Acceso al API con toda su funcionalidad (2).
- Vídeos, *papers*, glosario de términos.
- Tutoriales de casos de uso.
- Varios demos que permiten explorar la robustez de modelos de clasificación de imágenes.

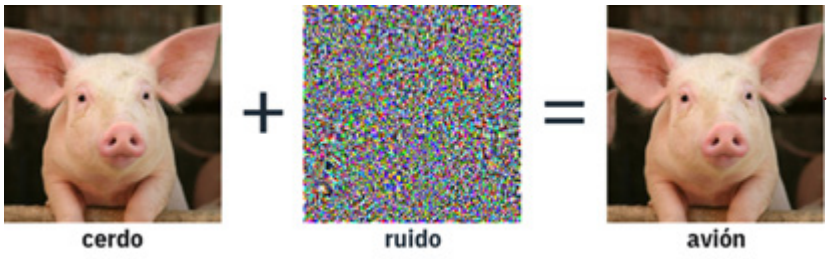
(1) https://github.com/Trusted-AI/adversarial-robustness-toolbox
(2) https://adversarial-robustness-toolbox.readthedocs.io/en/latest/

Qué es una muestra adversaria

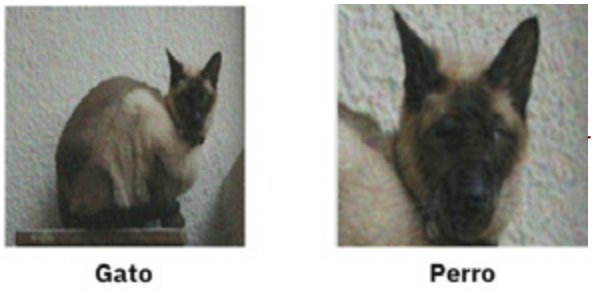
Una muestra adversaria es el elemento clave para poder realizar un ataque a un modelo inteligente que ponga en jaque su robustez.

Los ejemplos de modelos inteligentes de *computer* visión son muy didácticos para entender qué es una muestra adversaria. En dichos sistemas se puede manipular el *input* de diferentes maneras para obtener un resultado que no debiera ser el obtenido:

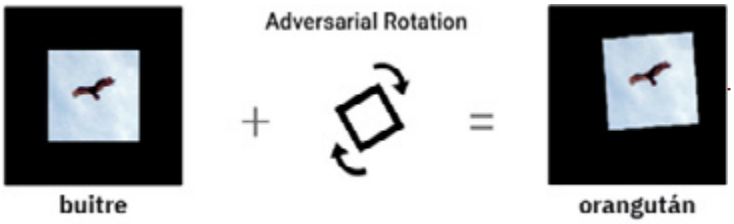
Generando "ruido pixelar", no detectable por el ojo humano



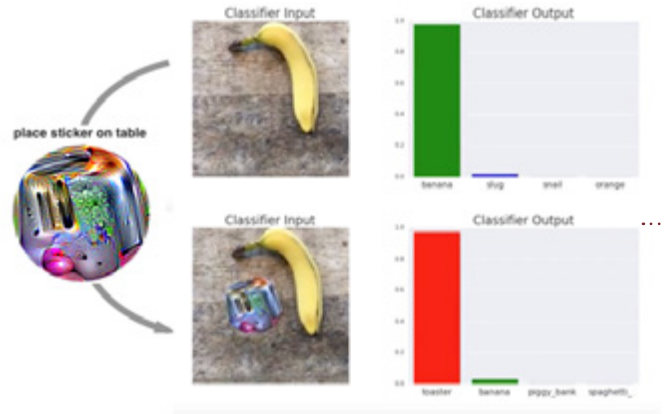
Focalizando una parte de la imagen



Rotando la imagen



Agregando un elemento adicional al elemento objeto de la evaluación



El primero de los ejemplos, el del ruido pixelar, requiere una elaborada manipulación técnica. Sin embargo, como reflejan los otros tres ejemplos, no es necesaria una manipulación especializada para engañar o equivocar a un sistema inteligente en su aprendizaje y en sus predicciones o toma de decisiones.

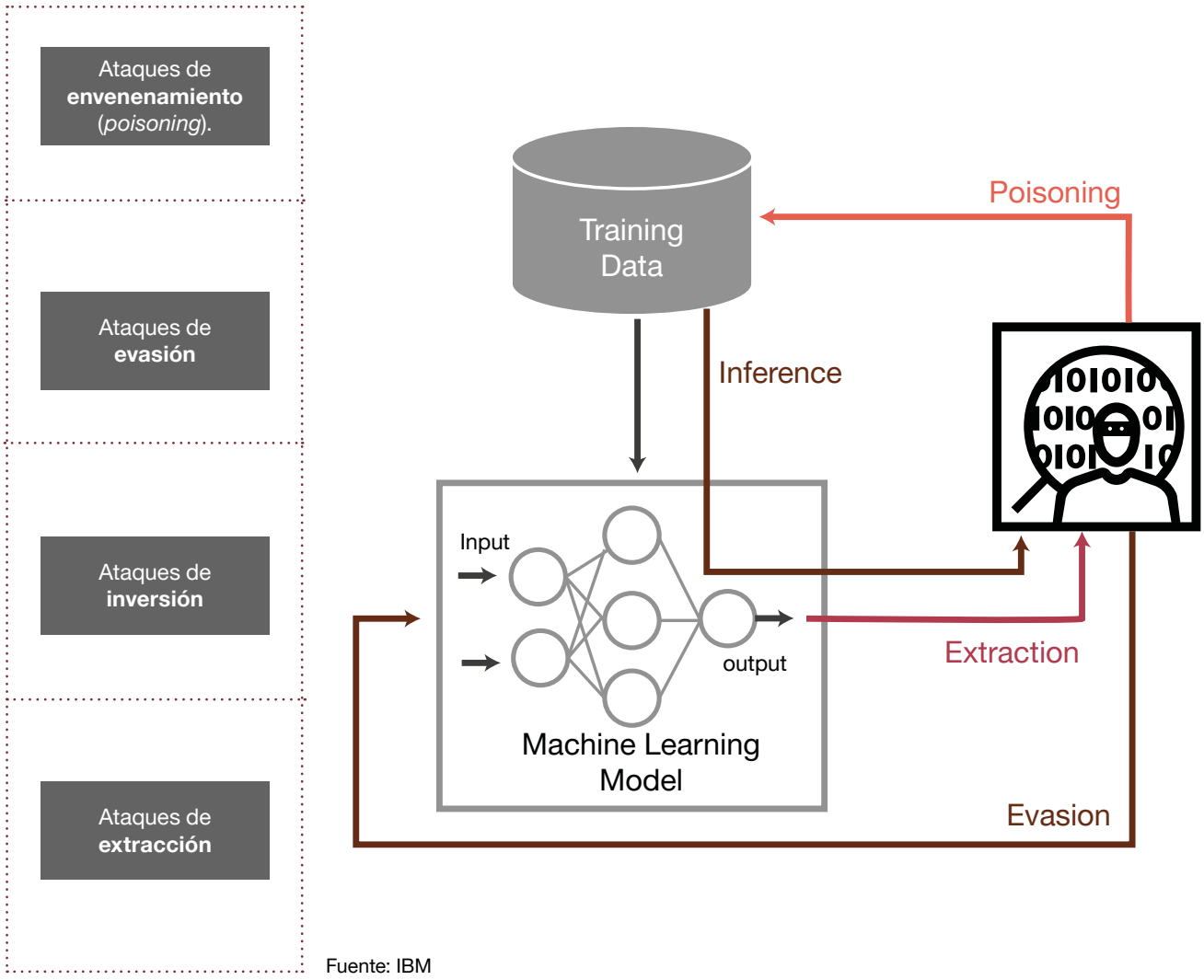
Estos comportamientos no robustos de un sistema inteligente se pueden producir a partir de un *input* introducido de manera intencionada. O simplemente por un proceso de aprendizaje no lo suficientemente profundo antes de ponerlo en producción. Se puede producir en la fase de entrenamiento del modelo, o cuando el sistema esté en producción. Por tanto, es importante poner el foco en estas posibles excepciones que puedan dar como resultado una predicción incorrecta, o incluso pervertir el proceso de aprendizaje del modelo hasta derivar en un mal funcionamiento.

Cómo gestionar ataques y defensas según el momento en el que se ataca o prueba el modelo

Primero, es necesario entender la diferencia entre ataque de caja blanca y de caja negra (no confundir con modelos inteligentes de caja blanca y caja negra):

- En un **ataque de caja blanca**, el atacante conoce interioridades del modelo (datos que se han utilizado para ser entrenado, variables evaluadas, etc.). Son los ataques más peligrosos dada la información manejada por el atacante.
- En un **ataque de caja negra**, el atacante únicamente dispone de las interfaces que el sistema proporciona, lo que le lleva a probar al sistema para conocer su razonamiento o incluso los datos con los cuales ha sido entrenado.

Existen **cuatro tipos de ataque/defensa** en función del momento en el que se prueba el modelo, si es cuando está en desarrollo o cuando está en producción.



Fuente: IBM

Ataques de **envenenamiento** (*poisoning*).	• Se produce en la **fase de entrenamiento**, actuando sobre los datos. • **Busca conseguir** predicción errónea, ya sea afectando de forma global al modelo o afectando sólo a determinados casos. • **Se puede hacer**: - Modificando algunas etiquetas de los datos de entrenamiento, provocando confusión u malfuncionamiento en el modelo. - Alterando algunas características para que, independientemente del *input*, el resultado sea el deseado por el atacante, pudiendo generar puertas traseras en el sistema inteligente. **Para detectar** este tipo de situaciones, dentro de su *toolkit* ART, IBM proporciona el **algoritmo *NeuralCleanse***.
Ataques de **evasión**	• Se produce en la **inferencia**, sin que el atacante tenga acceso directo a los datos, únicamente se comunica con el modelo a través de su interfaz. • **Busca conseguir** predicción errónea, ya sea afectando de forma global al modelo o afectando sólo a determinados casos. • **Se puede hacer** creando *inputs* que, siendo a priori muy parecidos a otros, enmascaren información que tenga intención de equivocar al modelo, como los ejemplos de los animales vistos anteriormente. **Para evitar** este tipo de situaciones, dentro de su *toolkit* ART, IBM proporciona el **algoritmo FGSM**. Dicho método genera muestras adversarias con el objetivo de que, cuando le lleguen esos *inputs* al modelo, éste sepa diferenciarlos.
Ataques de **inversión**	• Se produce en la **inferencia**, sin que el atacante tenga acceso directo a los datos, únicamente se comunica con el modelo a través de su interfaz. • **El objetivo** es obtener conocimiento de los datos del modelo. • Existen **tres tipos** de ataque de inversión en función del objetivo del atacante: - **MIA** (*Membership Inference Attack*). Busca conocer si una determinada muestra fue utilizada para entrenar al modelo. - **PIA** (*Property Inference Attack*). Busca conocer los datos utilizados por el modelo en su evaluación para realizar las predicciones. - **Reconstrucción**. Es el más ambicioso, ya que pretende reconstruir el conjunto de datos de entrada. Aplica en modelos muy preciso. • Estos ataques son especialmente relevantes si el modelo trata con datos confidenciales.
Ataques de **extracción**	• Se produce en la **inferencia**, sin que el atacante tenga acceso directo a los datos, únicamente se comunica con el modelo a través de su interfaz. • **El objetivo** es obtener conocimiento del modelo, entender su razonamiento. • El atacante proporciona *inputs* al modelo y utiliza el *output* para obtener una muestra con la que poder entrenar un modelo similar. Para evitar este tipo de situaciones, el *toolkit* IBM ART, permite generar defensas modificando la salida del modelo haciendo redondeos o añadiendo ruido estadístico, por ejemplo, para dificultar así al atacante el conocimiento del razonamiento del modelo.

Principio de Transparencia

En esta sección vamos a desarrollar:

- Qué es la transparencia para IBM.
- IBM *Uncertainty Quantification* 360, para gestionar el principio ético de Transparencia.
  - Cómo estimar la incertidumbre en las predicciones de un modelo de ML.
  - Cómo evaluar la calidad de esas incertidumbres y, si es necesario, mejorarla.
  - Cómo comunicar esas incertidumbres a las personas que hacen uso del modelo.
- La *FactSheet*, herramienta de gobernanza para una IA transparente y confiable.

Qué es la Transparencia para IBM

Existen incertidumbres inherentes en las predicciones de los modelos de aprendizaje automático (*Machine Learning*). Saber cuán incierta puede ser la predicción influye en cómo las personas actúan a partir de la misma. Por ejemplo, es posible que un modelo de pronóstico del tiempo prediga que mañana no lloverá con un 60 % de confianza. En ese caso, es posible que aún quieras llevar un paraguas cuando vayas a trabajar.

Para los usuarios finales de un sistema de IA, es **necesario proporcionar transparencia** sobre sus predicciones, ayudándolos a evaluar si quieren confiar (y aceptar) una predicción en particular, si deben recopilar más información o recurrir a un juicio alternativo. **Es imperativo presentar la información de UQ en aplicaciones de alto riesgo, como atención médica, finanzas y seguridad**, para evitar una dependencia excesiva de la IA y facilitar una mejor toma de decisiones.

La transparencia también **es útil para los desarrolladores de modelos como una herramienta para mejorar su modelo**. Estimar la incertidumbre del modelo y la incertidumbre de los datos por separado es fundamental.

La incertidumbre es un área importante de la investigación de ML, que ha producido una gran cantidad de algoritmos, métricas y formas de cuantificar la incertidumbre (UQ) para comunicar la incertidumbre a los usuarios finales. Los investigadores de IBM *Research* han estado trabajando activamente en estos temas para permitir una transparencia crítica en los modelos de ML y generar confianza en la IA.

IBM Uncertainty Quantification 360, opensource para gestionar el principio ético de Transparencia

Para su gestión automatizada, IBM facilita ***Uncertainty Quantification* 360 (UQ360)**, un amplio conjunto de herramientas de código abierto basado en *Python* para proporcionar a los profesionales y desarrolladores de la ciencia de datos acceso a los algoritmos más avanzados. Estas herramientas permiten agilizar el proceso de estimación y evaluación, mejorando y comunicando la incertidumbre de los modelos de aprendizaje automático como prácticas comunes para la transparencia de la IA.

Este paquete de IBM incluye:

- Tutoriales y cuadernos que ofrecen una introducción más profunda y orientada a los científicos de datos (1).
- Una demostración interactiva que permite explorar más a fondo estos conceptos y las capacidades que ofrece UQ360 recorriendo un caso de uso en el que se realizan diferentes tareas relacionadas con UQ (2).
- El API abierto de dicho paquete (3), disponible en *GitHub* (4)

Capacidades de IBM UQ360

Este paquete permite producir información transparente de alta calidad durante el desarrollo del modelo inteligente con los siguientes objetivos:

- Cómo estimar la incertidumbre en las predicciones de un modelo de ML.
- Evaluar la calidad de esas incertidumbres y, si es necesario, mejorarla.
- Comunicar esas incertidumbres de forma eficaz a las personas que hacen uso de la información UQ.

(1) https://uq360.mybluemix.net/resources/tutorials
(2) https://uq360.mybluemix.net/demo/0
(3) https://uq360.readthedocs.io/en/latest/
(4) https://github.com/IBM/UQ360

Cómo estimar la incertidumbre en las predicciones de un modelo de ML

El aprendizaje automático supervisado normalmente implica aprender un mapeo funcional entre entradas (características) y salidas (predicciones/recomendaciones/respuestas) a partir de un conjunto de ejemplos de entrenamiento que comprenden pares de entrada y salida.

La función aprendida se usa luego para predecir salidas para nuevas instancias o entradas que no se vieron durante el entrenamiento. Estos resultados pueden ser valores reales en el caso de un modelo de regresión o etiquetas de clase en el caso de un modelo de clasificación. En este proceso, **la incertidumbre puede surgir de múltiples fuentes:**

- Los **datos** disponibles pueden ser inherentemente ruidosos; por ejemplo, dos ejemplos con los mismos perfiles de características pueden tener resultados diferentes. Esto a menudo se denomina incertidumbre aleatoria. En adelante, nos referiremos a ella simplemente como incertidumbre de los datos. La incertidumbre de los datos se refiere a la variabilidad inherente a las instancias de datos y los objetivos.
- **La función de mapeo del modelo puede ser ambigua**: dado un conjunto de datos de entrenamiento, puede haber diferentes funciones que los expliquen. Esta incertidumbre sobre el modelo se denomina incertidumbre epistémica, y nos referiremos a ella como incertidumbre del modelo. Múltiples modelos (cada modelo se caracteriza por un conjunto de parámetros) pueden ser consistentes con los datos observados. La falta de conocimiento sobre un único modelo apropiado da lugar a la incertidumbre del modelo.

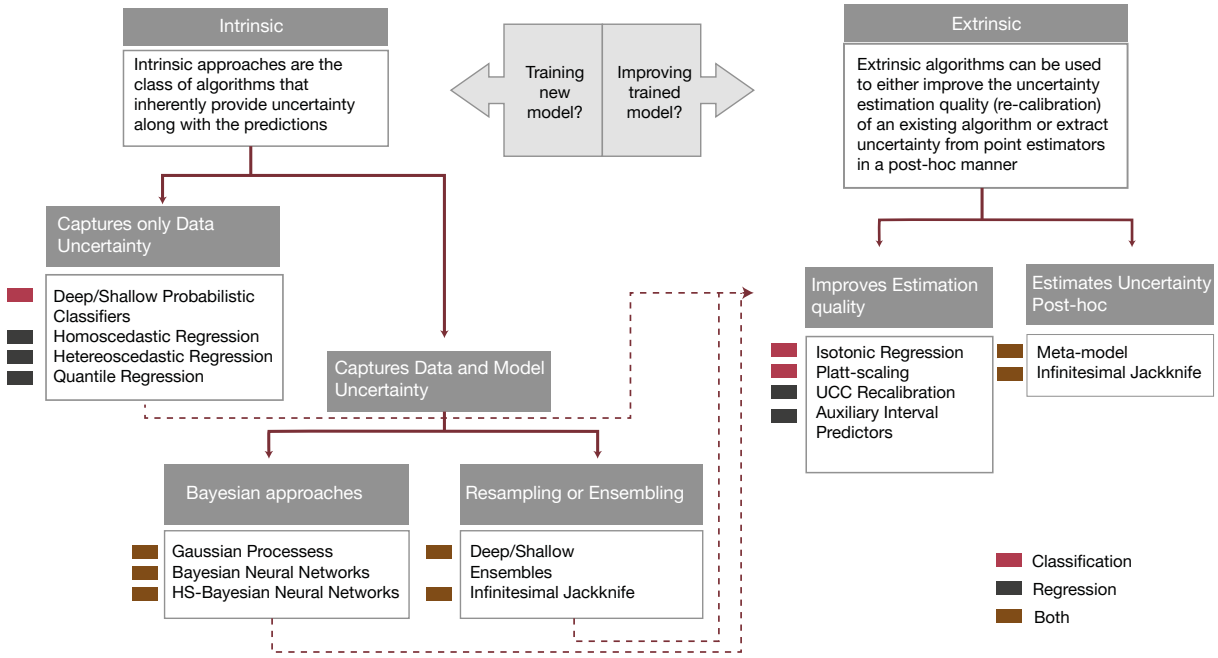
La incertidumbre en torno a las predicciones de un modelo de aprendizaje automático, **incertidumbre predictiva, es el agregado de las incertidumbres del modelo y de los datos**.

El kit de herramientas UQ360 proporciona algoritmos para estimar estos diferentes tipos de incertidumbres. Según el tipo de modelo y la fase de desarrollo del mismo, se deben aplicar diferentes algoritmos UQ. UQ360 actualmente proporciona 11 algoritmos UQ y una guía para elegir los algoritmos UQ para ayudarle a encontrar el apropiado para su caso de uso.

El enfoque de IBM

Cómo elegir algoritmos UQ

Los algoritmos de cuantificación de la incertidumbre (UQ) se pueden clasificar en términos generales como intrínsecos o extrínsecos, según cómo se obtengan las incertidumbres de los modelos de IA. El siguiente diagrama muestra la taxonomía de los algoritmos UQ, la mayoría de ellos incluyen UQ360. (1)



Fuente: IBM

¿Métodos UQ intrínsecos (entrenando un nuevo modelo) o extrínsecos (trabajando con un modelo entrenado)?

En primer lugar, la elección del algoritmo UQ depende de si se tiene la intención de entrenar un nuevo modelo o si ya se tiene un modelo entrenado. En el primer caso, se pueden usar métodos intrínsecos para entrenar un modelo que proporcione estimaciones de incertidumbre. Si ya entrenó un modelo, se pueden usar métodos extrínsecos para mejorar la calidad de las estimaciones de incertidumbre existentes de su modelo o generar estimaciones de incertidumbre post-hoc si su modelo no las proporciona.

Si es intrínseco, ¿le importa la incertidumbre del modelo?

Como comentábamos anteriormente, las dos fuentes principales de incertidumbre predictiva son incertidumbre de los datos e incertidumbre del modelo. Se pueden usar algunos algoritmos UQ intrínsecos para entrenar modelos ML que capturan ambos con las predicciones. Sin embargo, pueden ser computacionalmente costosos y, a veces, no pueden generar estimaciones de incertidumbre de alta calidad (más adelante profundizaremos sobre las métricas de evaluación). En estos casos, se puede optar por un algoritmo UQ intrínseco que captura solo la incertidumbre de los datos. Si es necesario, la salida de cualquier algoritmo UQ intrínseco se puede mejorar aún más utilizando métodos de recalibración (igualmente, profundizaremos sobre esto más adelante).

(1) https://uq360.mybluemix.net/resources/guidance

Métodos intrínsecos que capturan la incertidumbre tanto en datos como en el modelo

El conjunto de herramientas incluye dos clases de algoritmos para entrenar un modelo que captura tanto los datos como la incertidumbre del modelo: enfoques bayesianos y enfoques de conjunto/remuestreo. En general, los enfoques bayesianos son computacionalmente más costosos de entrenar, pero tienen una base teórica más sólida.

Los enfoques bayesianos incluidos en UQ360 son redes neuronales bayesianas (BNN) (1) y procesos gaussianos (2). Los BNN, en lugar de encontrar un conjunto único de parámetros de modelo óptimos, generan una distribución posterior sobre los parámetros del modelo dada una distribución previa y los datos de entrenamiento. Sin embargo, bajo una especificación incorrecta del modelo, los BNN pueden generar estimaciones de incertidumbre de baja calidad. UQ360 también incluye HS-BNN (3) que pueden generar incertidumbres de mejor calidad, especialmente cuando se trabaja con conjuntos de datos relativamente pequeños.

Los conjuntos de modelos permiten enfoques atractivos para estimar la incertidumbre del modelo, incluso si los modelos individuales del conjunto no capturan las incertidumbres. La variabilidad en las predicciones entre los miembros del conjunto puede verse como la incertidumbre predictiva. Si bien son prometedores, los métodos estándar para crear conjuntos requieren ajustes repetidos del modelo: un ajuste para cada miembro, un proceso prohibitivamente costoso para grandes conjuntos de datos. UQ360 incluye el método *jackknife* infinitesimal (4), que permite la construcción de conjuntos a partir de un solo ajuste de modelo.

Métodos intrínsecos que capturan solo la incertidumbre de los datos

Para la clasificación, los modelos de clasificación probabilística estándar pueden capturar la incertidumbre de los datos. Para la regresión, los modelos de regresión homoscedástica (5), regresión heteroscedástica (6) y de regresión cuantil (7) pueden capturar la incertidumbre de los datos. Sin embargo, difieren en sus suposiciones subyacentes del ruido del modelo para la distribución predictiva. La regresión homoscedástica asume que el ruido es constante en todas las características, mientras que la heteroscedástica permite que el ruido varíe con diferentes características. La regresión de cuantiles es un método no paramétrico que predice directamente los cuantiles de incertidumbre de los resultados predichos sin asumir una distribución de ruido.

Si es extrínseco, ¿el modelo entrenado proporciona UQ o no?

Si se está trabajando con un modelo entrenado, primero debe determinar si el modelo puede proporcionar estimaciones de incertidumbre directamente. Por ejemplo, los modelos de clasificación probabilística de uso común, como las redes neuronales, pueden capturar la incertidumbre de los datos, mientras que las SVM, los árboles de decisión y los *k-nearest neighbors* generalmente no pueden capturar directamente las incertidumbres en sus predicciones. Si el modelo proporciona estimaciones de incertidumbre, se puede evaluar su calidad utilizando las métricas UQ que veremos más adelante. Si no está satisfecho con la calidad, UQ360 proporciona una clase de algoritmos extrínsecos para mejorar estas estimaciones. Si su modelo no proporciona estimaciones de incertidumbre directamente, UQ360 incluye una clase de enfoques *post-hoc* para generar incertidumbre en estas estimaciones.

Métodos extrínsecos para mejorar la calidad UQ

UQ360 proporciona un conjunto de algoritmos para mejorar la calidad, específicamente la calibración de las estimaciones de incertidumbre existentes. En resumen, las estimaciones de incertidumbre mal calibradas significan que la distribución de resultados observada probada con un conjunto de datos no se alinea con las estimaciones de incertidumbre dadas, y la recalibración implica corregir las estimaciones para que coincidan con la distribución observada. Para las tareas de clasificación, puede utilizar la regresión isotónica y el escalado de Platt (8). Para tareas de regresión, se pueden emplear predictores de intervalos auxiliares (9) y recalibración UCC (10). Los métodos de recalibración también se pueden utilizar para mejorar las estimaciones de incertidumbre de los algoritmos UQ intrínsecos.

(1) https://uq360.readthedocs.io/en/latest/intrinsic.html#bayesian-neural-network-regression
(2) https://uq360.readthedocs.io/en/latest/intrinsic.html#homoscedastic-gaussian-process-regression
(3) https://uq360.readthedocs.io/en/latest/intrinsic.html#bayesian-neural-network-regression
(4) https://uq360.readthedocs.io/en/latest/extrinsic.html#infinitesimal-jackknife
(5) https://uq360.readthedocs.io/en/latest/intrinsic.html#heteroscedastic-regression
(6) https://uq360.readthedocs.io/en/latest/intrinsic.html#heteroscedastic-regression
(7) https://uq360.readthedocs.io/en/latest/intrinsic.html#quantile-regression
(8) https://uq360.readthedocs.io/en/latest/extrinsic.html#classification-calibration
(9) https://uq360.readthedocs.io/en/latest/extrinsic.html#auxiliary-interval-predictor
(10) https://uq360.readthedocs.io/en/latest/extrinsic.html#ucc-recalibration

Métodos extrínsecos para generar UQ post-hoc

Para los modelos existentes que no pueden generar estimaciones de incertidumbre directamente, UQ360 proporciona metamodelos (MM) (1) que pueden obtener estas estimaciones de manera post-hoc. En el caso de la regresión, un MM aumenta el modelo base para obtener un intervalo de predicción, mientras que, en el caso de la clasificación, un MM devuelve un valor escalar que indica la confianza en la predicción del modelo base. El algoritmo IJ comentado anteriormente también se puede usar para generar UQ post-hoc al aproximar el efecto de las perturbaciones de datos de entrenamiento en las predicciones del modelo.

Cómo evaluar la calidad de esas incertidumbres y, si es necesario, mejorarla

La calidad de las incertidumbres generadas por un algoritmo UQ también debe evaluarse. Las incertidumbres de baja calidad no deben ser confiables ni comunicadas a los usuarios. Primero deben ser mejoradas. Para ello, UQ360 proporciona un conjunto de métricas (2) para modelos de clasificación y regresión que permiten medir la calidad de las incertidumbres producidas por diferentes algoritmos y el conjunto de técnicas para mejorar la calidad de las incertidumbres estimadas que se han descrito anteriormente.

(1) https://uq360.readthedocs.io/en/latest/extrinsic.html#blackbox-metamodel-classification

(2) https://uq360.readthedocs.io/en/latest/metrics.html

Cómo comunicar esas incertidumbres a las personas que hacen uso del modelo

Por último, la mejor manera de comunicar las estimaciones de incertidumbre depende del tipo de modelo, la forma en que se utilizará la información de UQ y los destinatarios de la información. Para un modelo de clasificación como el de pronóstico del tiempo comentado anteriormente, la incertidumbre suele ser una puntuación que a menudo se denomina confianza. Para un modelo de regresión, la incertidumbre podría comunicarse de varias maneras, incluido un rango en el que el resultado predicho posiblemente puede caer, a menudo denominado intervalo de predicción, o mediante visualizaciones. UQ360 proporciona una guía para comunicar la incertidumbre para ayudar a elegir una forma adecuada de presentar la cuantificación de la incertidumbre.

Qué considerar al elegir métodos de comunicación

Comunicar UQ significa presentar las estimaciones de incertidumbre a las partes interesadas, suponiendo que haya elegido el algoritmo UQ correcto para generar el tipo correcto de estimaciones de incertidumbre (los vistos anteriormente). Este es un paso crucial porque incluso las personas pueden malinterpretar las estimaciones de incertidumbre de alta calidad si tienen dificultades o sesgos al interpretar los números o las estadísticas. A continuación presentamos algunas consideraciones clave para comunicar UQ y métodos de ejemplo. En la práctica, es necesario realizar pruebas con sus usuarios objetivo para asegurar que el método de comunicación de UQ elegido se entienda correctamente.

Comencemos con algunas preguntas clave que deberían guiar la elección de los métodos de comunicación de UQ. Puede navegar a las pestañas de arriba para ver métodos de ejemplo para comunicar UQ de modelos de regresión y clasificación.

¿Cuál es la forma de la UQ?

El primer paso es identificar la forma del UQ que se comunicará, es decir, si se trata de una puntuación de confianza única o de un rango o distribución de los posibles resultados previstos. En los actuales modelos de ML, la primera es la forma en que suelen aparecer las estimaciones de incertidumbre de un modelo de clasificación, y la segunda es la forma en que aparecen las estimaciones de incertidumbre de un modelo de regresión.

Por ejemplo, para un modelo de regresión, el UQ de diferentes fuentes, ya sea la incertidumbre de los datos, la incertidumbre del modelo o la incertidumbre predictiva general, se pueden comunicar como rangos de posibles resultados, pero es posible que los usuarios los perciban o actúen en consecuencia diferentemente.

¿Comunicando UQ de una sola instancia o un grupo de instancias?

La siguiente pregunta es si se desea presentar el UQ de una sola instancia o de un grupo de instancias. Los usuarios finales a menudo están interesados en la estimación de la incertidumbre de una predicción en particular para evaluar y actuar sobre la predicción. En esta guía nos enfocamos en diferentes métodos de comunicación UQ en el caso de una sola instancia.

A veces, un usuario o un científico de datos puede estar interesado en investigar el UQ de un grupo de instancias. Por ejemplo, grupos de diferentes valores de características, o qué tipo de instancias obtienen estimaciones de incertidumbre alta o baja del modelo. Esto a menudo se puede lograr trazando visualmente las estimaciones de incertidumbre de las instancias con respecto a diferentes valores de características, o presentando las estimaciones de incertidumbre agregadas por los grupos interesados (p. ej., estimaciones de incertidumbre media).

¿Cómo de preciso debe ser el UQ?

UQ se puede mostrar con diferentes niveles de precisión. Por ejemplo, describir el UQ de un modelo de clasificación utilizando categorías predefinidas, como confianza baja/media/alta, es menos preciso que mostrar valores numéricos de confianza. Proporcionar un rango numérico para UQ de un modelo de regresión es menos preciso que mostrar la distribución de probabilidad de los valores posibles.

En general, los métodos de comunicación de alta precisión pueden ayudar a los expertos de ML a comprender la escala total de la incertidumbre del modelo. Mientras que los métodos de baja precisión a menudo son preferidos por los juristas. Una vez más, es necesario diseñar y probar los métodos de comunicación con las partes interesadas objetivo para identificar el nivel de precisión adecuado.

¿Qué medio utilizar?

Otra cuestión que podría correlacionarse con la elección del nivel de precisión es mediante qué medio comunicar UQ, ya sea a través de expresiones verbales, numéricas o visuales. A menudo, las expresiones verbales combinan bien con métodos de comunicación de baja precisión para permitir que las partes interesadas consuman fácilmente la información de UQ. La visualización es a menudo un buen enfoque para comunicar información UQ probabilística de alta precisión. A veces, la elección del medio está predeterminada por la interfaz o el flujo de trabajo del usuario, y debe diseñar el contenido en consecuencia.

¿Cómo comunicarlo en modelos de regresión?

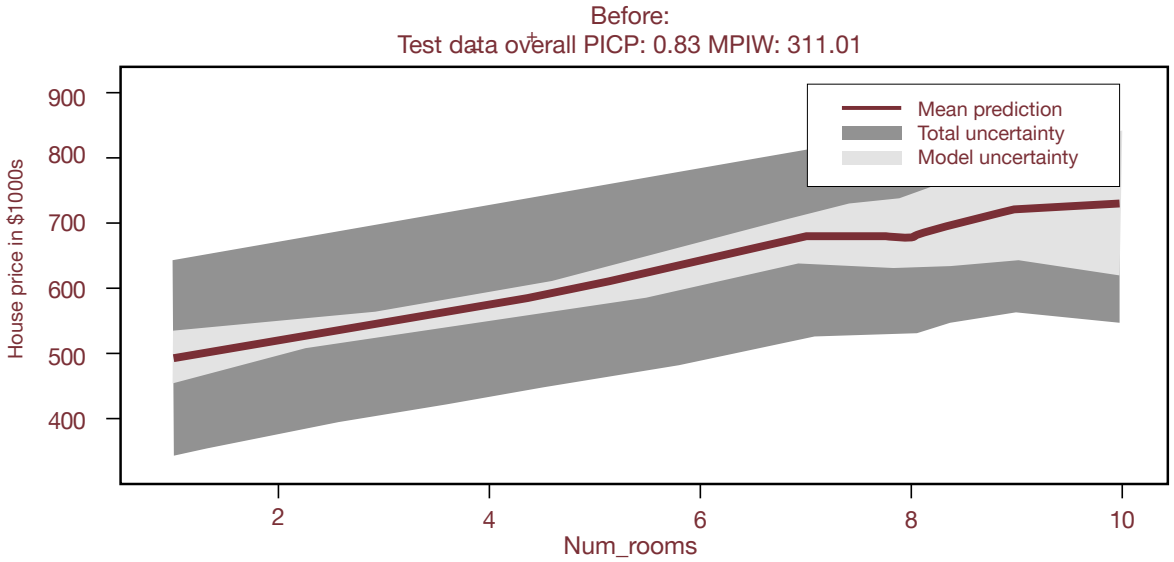
Para modelos de regresión, la biblioteca de *Python* UQ360 proporciona funciones para generar rangos numéricos, intervalos visuales, gráficos de densidad y gráficos de puntos de cuantiles.

En los modelos de regresión, a menudo las personas optan por comunicar el rango en el que el posible resultado de la predicción puede caer con una probabilidad preespecificada como nivel de confianza (p. ej., 95 %). Es posible utilizar intervalo de predicción o una barra de error para comunicar visualmente este rango. Tiene los beneficios de ser fácil de entender y no involucrar probabilidades y, por lo tanto, a menudo se prefiere cuando se comunica UQ a audiencia con conocimientos numéricos relativamente bajos.

Según el caso de uso, no siempre es deseable presentar un diagrama visual. También puede comunicar el rango del intervalo de predicción en lenguaje natural, como en la imagen mostrada.

La desventaja de un rango de intervalo es que las partes interesadas no pueden ver los detalles de distribución, por lo que podrían malinterpretar que los posibles resultados se distribuyen por igual en el rango de intervalo, lo que a menudo no es el caso. Además, no todos están familiarizados con el concepto de intervalo de predicción o entienden que debe leerse con un nivel de confianza específico, por lo que pueden ser necesarias explicaciones adicionales.

Otro beneficio de un rango de intervalo visual es que es lo suficientemente simple como para combinarlo con otra codificación visual para comunicar el UQ para un grupo de predicciones. Por ejemplo, es común mostrar una banda de incertidumbre en un gráfico de líneas para un grupo de predicciones, como se muestra a continuación

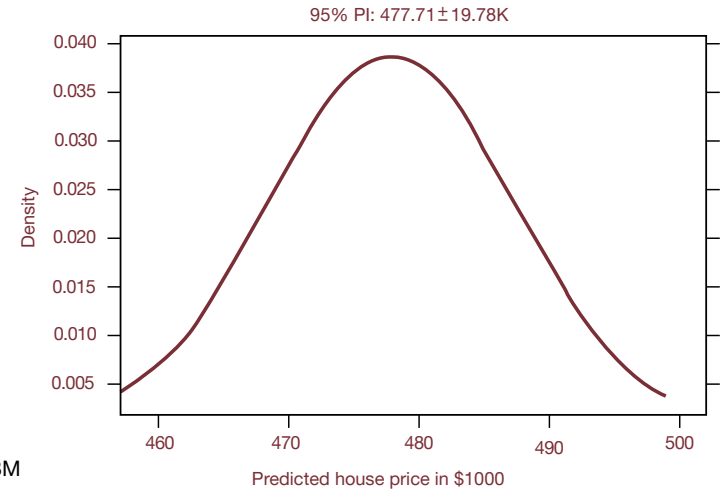


Fuente: IBM

En ocasiones, es posible que desee mostrar información detallada de UQ sobre cómo se distribuyen los posibles resultados previstos. Esto ayudaría a comprender que los posibles resultados no se distribuyen por igual en todo el rango y, por lo tanto, evaluar mejor la incertidumbre.

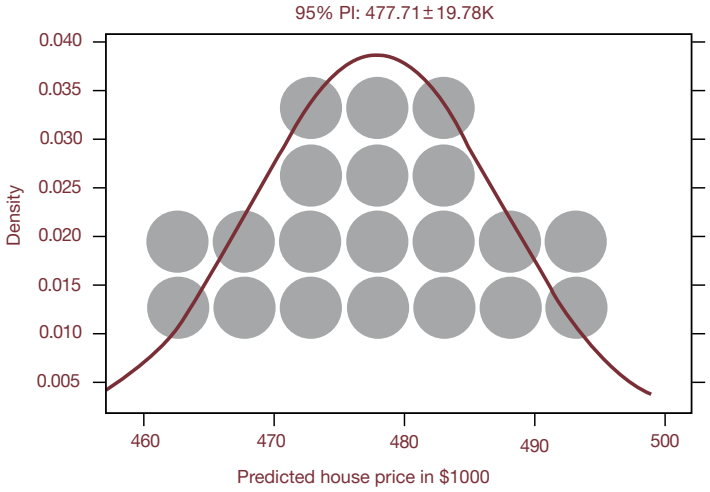
Existen algunos métodos para visualizar una distribución de probabilidad, incluidos el gráfico de densidad, el gráfico de violín y el gráfico de gradiente. Se pueden obtener de bibliotecas de visualización estándar.

Vale la pena señalar que las estimaciones de incertidumbre no siempre siguen una distribución gaussiana simétrica centrada alrededor del valor predicho. En ese caso, es aún más importante mostrar los detalles de distribución para alertar de que las estimaciones de incertidumbre están sesgadas o tienen múltiples picos.



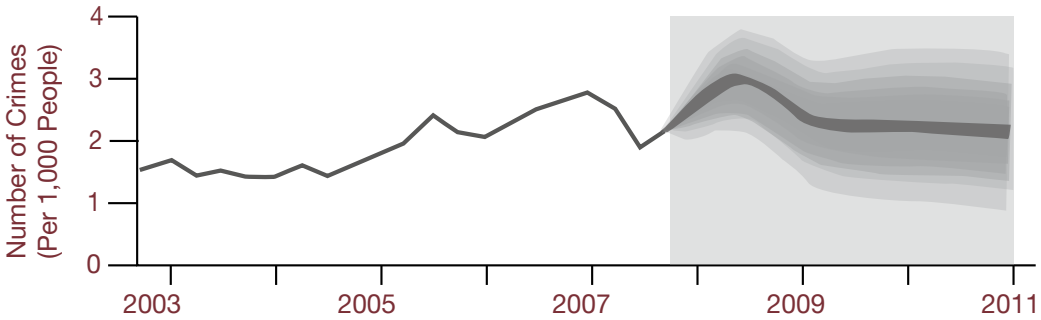
Fuente: IBM

Una variación útil de las gráficas de densidad, desarrollada por una investigación reciente de interacción humano-computadora, son las gráficas de puntos de cuantiles. Un inconveniente crítico de un gráfico de densidad tradicional es que las personas pueden tener problemas para evaluar correctamente la probabilidad relativa de diferentes valores posibles, especialmente para aquellos que no están capacitados para leer la visualización de datos. Un gráfico de puntos de cuantiles mitiga este problema mediante el uso de puntos apilados para representar visualmente las frecuencias aproximadas de diferentes valores. Por ejemplo, en el siguiente gráfico, es fácil ver que 12 de 20 veces los precios caen por debajo de 480K.



Fuente: IBM

Al comunicar la incertidumbre de un grupo de datos, es posible superponer estas distribuciones visuales con otra codificación visual. Por ejemplo, el gráfico de abanico es un enfoque común para codificar gráficos de gradiente en un gráfico de líneas para un grupo de predicciones.



Fuente: IBM

¿Como comunicarlo en modelos de clasificación?

Para modelos de clasificación, la biblioteca *Python* UQ360 proporciona funciones para generar evaluaciones de confianza para algoritmos UQ relevantes.

En el caso de un modelo de clasificación, la estimación de la incertidumbre de una predicción suele ser una sola puntuación. Existen algunas opciones para comunicar la puntuación:

Categorías ordenadas

A menudo, para los legos, una puntuación de confianza numérica es difícilmente procesable. Uno puede tener problemas para juzgar, digamos, si un nivel de confianza del 85% es lo suficientemente bueno para confiar en la predicción del modelo. La respuesta depende en gran medida del dominio de la tarea y de lo que está en juego. El uso de una categorización ordenada predefinida, como confianza alta, moderada, baja y muy baja, podría guiar de manera eficiente a las personas para hacer ese juicio.

La principal desventaja de usar categorías ordenadas es la falta de precisión. Además, el mapeo entre un puntaje de confianza a una categoría, o los criterios de umbral, debe diseñarse cuidadosamente en función de la tarea y las partes interesadas objetivo, como lo que está en juego en sus acciones de seguimiento al aceptar o rechazar una predicción del modelo, y su riesgo-tolerancia. Por ejemplo, una puntuación de confianza del 90 % podría verse como una confianza alta para un recomendador de películas, pero podría no serlo para la IA de imágenes médicas para el diagnóstico de enfermedades.

Puntuación de confianza

A veces es mejor presentar el puntaje de confianza real, si las partes interesadas objetivo están interesadas en los valores precisos o en comparar las estimaciones de incertidumbre de diferentes predicciones.

Con cierta pérdida de precisión, los legos a menudo son mejores para interpretar una frecuencia verbal (p. ej., el modelo cree que 9 de cada 10 posibilidades de que la predicción sea...) que un dato numérico.

Si es importante enfatizar la existencia de incertidumbre, también puede usar un gráfico circular o una matriz de iconos para comunicar la naturaleza probabilística de la predicción, como se muestra a continuación.

La *FactSheet*, una herramienta de gobernanza para fomentar confianza en una IA transparente y confiable

Los modelos y servicios de IA se utilizan en un número creciente de áreas de alto riesgo, como la evaluación de riesgos financieros, el diagnóstico médico y la planificación del tratamiento, las decisiones de contratación y promoción, la determinación de selección de los servicios sociales, la vigilancia predictiva y las recomendaciones jurídicas, por ejemplo.

El objetivo del proyecto *FactSheet* (1) es fomentar la confianza en la IA aumentando la transparencia y habilitando gobernanza. Una mayor transparencia proporciona información para que los consumidores de IA comprendan mejor cómo un modelo de IA o un servicio fue creado.

Esto permite que el consumidor del modelo determine si es apropiado para sus necesidades. La gobernanza de la IA permite a una empresa especificar y aplicar políticas que describen cómo se debe construir e implementar un modelo o servicio de IA. Esto puede evitar situaciones no deseadas, como el entrenamiento de un modelo con conjuntos de datos no aprobados, modelos con sesgos o modelos con variaciones de rendimiento inesperadas.

El aumento de la transparencia de la IA y la mejora de la gobernanza de la IA son dos *inputs* importantes para el proyecto *FactSheet*. A continuación describimos las ideas clave que son comunes a ambos *inputs*.

¿Qué es una FactSheet?

Una *FacSheet* es una recopilación de información relevante (hechos) sobre la creación y el despliegue de un modelo o servicio de IA. Los hechos pueden variar desde información sobre el propósito y la importancia del modelo, características medidas del conjunto de datos, modelo o servicio, o acciones realizadas durante el proceso de creación e implementación del modelo o servicio. Dichos modelos son creados por varios roles a lo largo del ciclo de vida del modelo de *Machine Learning*, como el dueño de un negocio, un científico de datos, el validador del modelo, y el operador del mismo.

Cada uno de dichos roles en el ciclo de vida aporta datos sobre cómo se creó e implementó el modelo. Por ejemplo, el propietario de la empresa puede proporcionar el uso previsto para el modelo. El científico de datos puede describir diversas actividades de recopilación y manipulación de datos. Un evaluador de modelos puede describir medidas de prueba clave y un operador de modelos puede proporcionar métricas de rendimiento clave.

Una *FactSheet* se basa en el estándar de declaración de conformidad de un proveedor -SDoC- (2), que se utiliza en muchas industrias diferentes de USA para mostrar que un producto cumple con un estándar o reglamento técnico. Sin embargo, las hojas informativas se pueden representar en muchos formatos diferentes, no solo en documentos impresos.

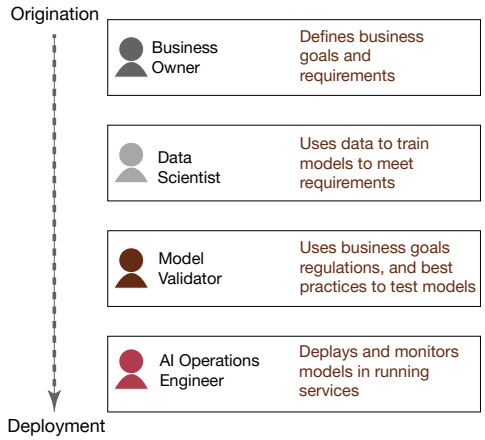
(1) https://aifs360.mybluemix.net/

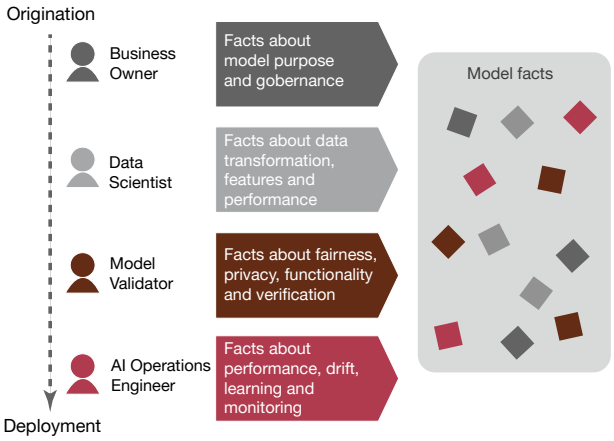
El enfoque de IBM

Un modelo basado en plantillas, adaptable para cada caso de uso

IBM considera que una *FacSheet* estándar única para todos los casos de uso no es factible porque el contexto, el dominio de la industria y el consumidor objetivo varían en cada caso. Por eso proporciona una plantilla adaptable para cada caso de uso (1), donde los de mayor riesgo en la aplicación de la IA requerirían de mayores adaptaciones. Además, las *FacSheets* varían según el propósito de la comunicación y la audiencia objetivo. Por ello IBM proporciona:

- **Un modelo de gobernanza** para el ciclo de vida de las soluciones inteligentes (2), que permite identificar roles y actividades en cada fase de dicho ciclo de vida.
- **Una metodología que permite crear una *FacSheet*** a partir de una plantilla genérica (3), teniendo en cuenta el modelo de gobernanza previamente definido.

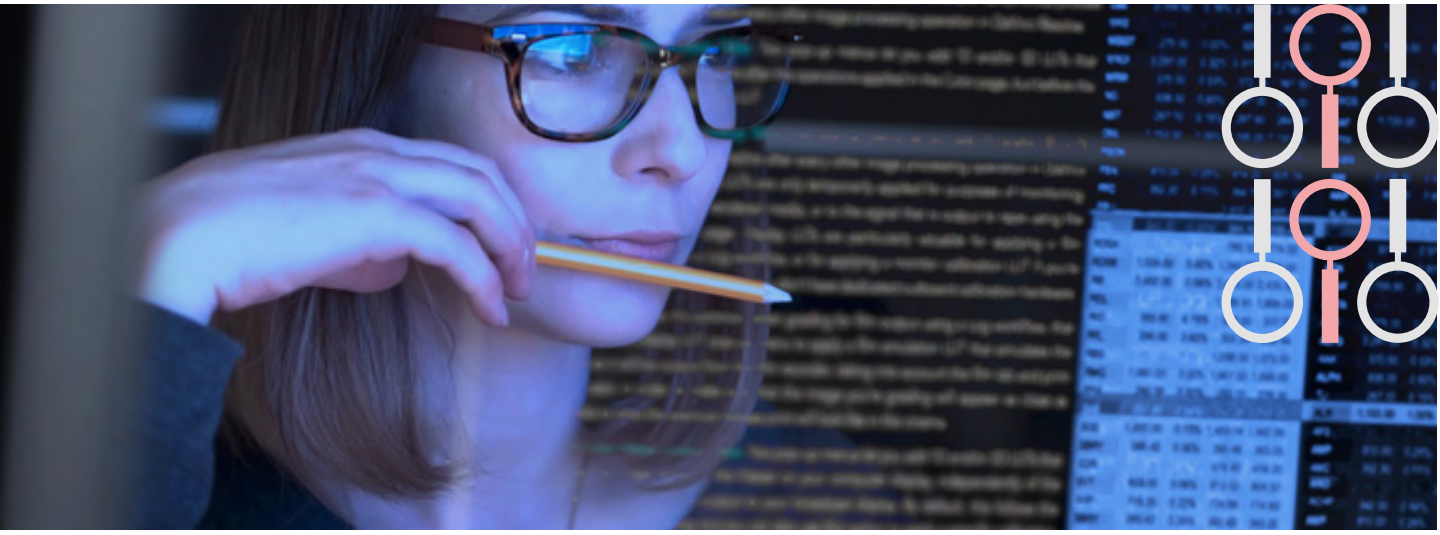




Fuente: IBM

- **Varios ejemplos** diseñados para casos de uso y audiencias específicas (4).

(1) https://arxiv.org/pdf/1911.08293.pdf
(2) https://aifs360.mybluemix.net/governance
(3) https://aifs360.mybluemix.net/methodology
(4) https://aifs360.mybluemix.net/examples



Principio de Privacidad

En esta sección vamos a desarrollar:

- Qué es la privacidad para IBM.
- *IBM AI Privacy* 360, opensource para gestionar el principio ético de Privacidad.
  - Gestión de datos cifrados
  - Privacidad diferencial.
  - Anonimización.
  - Minimización de datos.
  - Evaluación del riesgo de privacidad.
  - Privacidad en el aprendizaje federado

Qué es la privacidad para IBM

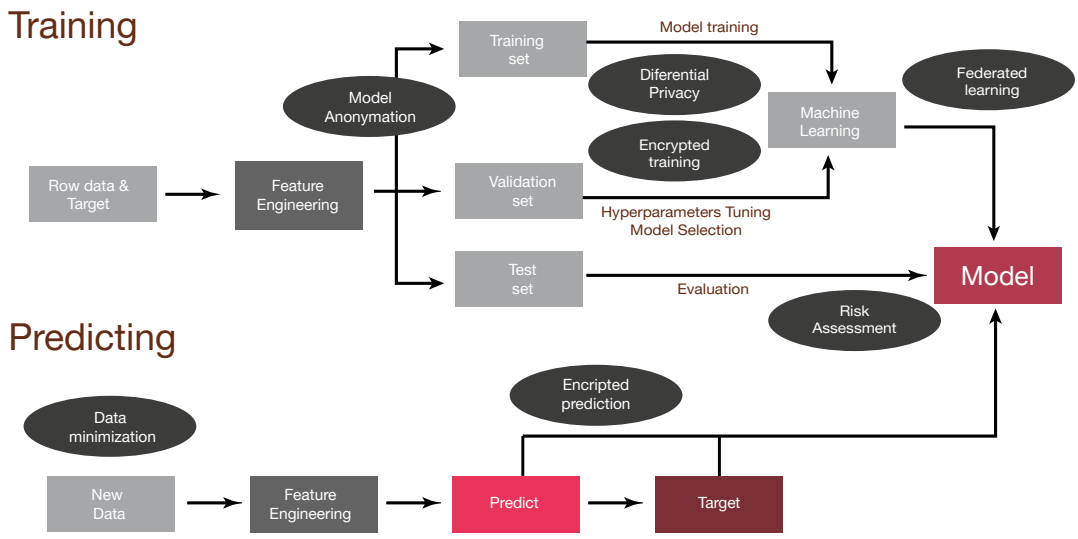
Muchas regulaciones de privacidad, incluida la GDPR, exigen que las organizaciones cumplan con ciertos principios de privacidad al procesar información personal. Esto es relevante para el *Machine Learning* ya que un tercero malintencionado con acceso a un modelo de aprendizaje automático entrenado, incluso sin acceso a los datos de entrenamiento en sí, aún puede revelar información personal confidencial sobre las personas cuyos datos se usaron para entrenar el modelo. Por lo tanto, es crucial poder reconocer y proteger los modelos de IA que pueden contener información personal.

IBM *AI Privacy* 360, *opensource* para gestionar el principio ético de Privacidad

IBM cuenta el *toolkit IBM AI Privacy* 360 para su gestión automatizada, accesible en https://aip360.mybluemix.net/

Dicho *toolkit* cuenta con herramientas para abordar este reto compensando las necesidades de privacidad, precisión y rendimiento de los modelos resultantes, proporcionando las siguientes capacidades:

- Gestión de datos cifrados.
- Privacidad diferencial.
- Anonimización.
- Minimización de datos.
- Evaluación del riesgo de privacidad.
- Privacidad en el aprendizaje federado.



Fuente: IBM

El enfoque de IBM

A continuación describimos cada una de las capacidades, indicando los recursos técnicos proporcionados por IBM para cada una de ellas.

inteligencia artificial sobre datos cifrados

Si bien el cifrado permite proteger los datos tanto durante el tránsito como durante el almacenamiento, los datos generalmente deben descifrarse mientras se accede a ellos para operaciones informáticas y comerciales críticas. El cifrado totalmente homomórfico (FHE) es una forma de cifrado más avanzada, que está diseñada para cerrar esta brecha, al permitir que los datos permanezcan cifrados incluso durante el cálculo. Las matemáticas detrás de FHE están diseñadas para que los cálculos se puedan realizar en datos cifrados (texto cifrado), sin que el servicio detrás de él necesite "ver" esos datos para proporcionar resultados precisos. Por lo tanto, al usar FHE, podemos implementar diferentes soluciones de análisis e inteligencia artificial sobre datos encriptados.

Documentación de la API disponible en GitHub (1).

Código fuente en Python y C++ disponible en GitHub (2).

Privacidad diferencial

Desde su concepción en 2006, la privacidad diferencial se ha convertido en el estándar de facto en la privacidad de datos, debido a sus sólidas garantías matemáticas, aplicabilidad generalizada y abundante literatura. A lo largo de los años, los investigadores han estudiado la privacidad diferencial y su aplicabilidad a un campo de temas cada vez más amplio.

Este método permite ejecutar consultas sobre datos sensibles mientras se preserva la privacidad de los individuos en los datos con sus sólidas garantías matemáticas. La privacidad diferencial se basa en el ruido aleatorio para proteger la privacidad de las personas al mismo tiempo que preserva la precisión en las estadísticas agregadas y tiene aplicaciones en ML y análisis de datos en general.

La biblioteca de privacidad diferencial de IBM es una biblioteca de código abierto de propósito general para investigar, experimentar y desarrollar aplicaciones de privacidad diferencial en el lenguaje de programación Python. La biblioteca incluye una serie de mecanismos, los componentes básicos de la privacidad diferencial, junto con una serie de aplicaciones para el aprendizaje automático y otras tareas de análisis de datos. Se ha priorizado la simplicidad y la accesibilidad en el desarrollo de la biblioteca, haciéndola adecuada para una amplia audiencia de usuarios, desde aquellos que usan la biblioteca para sus primeras investigaciones sobre privacidad de datos, hasta los expertos en privacidad que buscan contribuir con sus propios modelos y mecanismos para que otros los usen. Recursos disponibles:

- Documentación de la API (3).
- Código fuente en Python y C++ disponible en GitHub (4).
- Leading Paper acerca de privacidad diferencial (5).
- Notas adicionales (6).

(1) https://github.com/IBM/fhe-toolkit-linux/blob/master/GettingStarted.md
(2) https://github.com/IBM/fhe-toolkit-linux
(3) https://diffprivlib.readthedocs.io/en/latest/
(4) https://github.com/IBM/differential-privacy-library
(5) https://arxiv.org/pdf/1907.02444.pdf
(6) https://github.com/IBM/differential-privacy-library/tree/main/notebooks

Anonimización

Las organizaciones a menudo tienen la necesidad de entrenar modelos ML en datos personales, pero de una manera que preserve el anonimato de las personas cuyos datos se usaron durante el entrenamiento. El aprendizaje sobre datos anonimizados suele dar como resultado una degradación significativa de la precisión. Otros métodos, como los que se basan en la privacidad diferencial, tienden a ser mucho más complejos y consumen muchos recursos, lo que requiere reemplazar los algoritmos de entrenamiento existentes por otros nuevos y, a menudo, requiere el uso de varias herramientas o implementaciones diferentes.

IBM propone una herramienta basada en *Phyton* (1 y 2) que permite anonimizar los datos de entrenamiento de una manera que se adapta a un modelo específico, lo que da como resultado modelos anonimizados con una precisión mucho mayor que cuando se aplican algoritmos de anonimización tradicionales que no tienen en cuenta el uso objetivo de los datos.

En un artículo (3), IBM demuestra que este método logra resultados similares en su capacidad para prevenir ataques de inferencia de membresía como enfoques alternativos basados en la privacidad diferencial. También demostramos que nuestro método es capaz de defenderse contra otras clases de ataques, como la inferencia de atributos.

Esto significa que la anonimización guiada por modelos puede, en algunos casos, ser un sustituto legítimo de dichos métodos, al tiempo que evita algunos de sus inconvenientes inherentes, como la complejidad, la sobrecarga de rendimiento y la adaptación a tipos de modelos específicos. A diferencia de los métodos que se basan en agregar ruido durante el entrenamiento, nuestro enfoque no se basa en realizar modificaciones en el propio algoritmo de entrenamiento y puede funcionar incluso con modelos de "caja negra" en los que el propietario de los datos no tiene control sobre el proceso de entrenamiento. Como tal, se puede aplicar en una amplia variedad de casos de uso.

Recursos adicionales disponibles:

- Defensa contra la inferencia de membresía usando anonimización ML (4).
- Defensa contra la inferencia de atributos mediante la anonimización de ML (5).

(1) https://github.com/IBM/ai-privacy-toolkit

(2) https://ai-privacy-toolkit.readthedocs.io/en/latest/

(3) https://arxiv.org/abs/2007.13086

(4) https://github.com/IBM/ai-privacy-toolkit/blob/main/notebooks/membership_inference_anonymization_adult.ipynb

(5) https://github.com/IBM/ai-privacy-toolkit/blob/main/notebooks/attribute_inference_anonymization_nursery.ipynb

Minimización de datos

El Reglamento General Europeo de Protección de Datos (RGPD) dicta que “los datos personales deberán ser: adecuados, pertinentes y limitados a lo necesario en relación con los fines para los que son tratados”. Este principio, conocido como minimización de datos, requiere que las organizaciones y los gobiernos recopilen solo los datos necesarios para lograr el propósito en cuestión. Se espera que las organizaciones demuestren que los datos que recopilan son absolutamente necesarios, mostrando medidas concretas que se tomaron para minimizar la cantidad de datos utilizados para cumplir un propósito determinado. De lo contrario, corren el riesgo de violar las normas de privacidad, incurrir en grandes multas y enfrentarse a posibles demandas.

Los algoritmos avanzados de aprendizaje automático, como las redes neuronales, tienden a consumir grandes cantidades de datos para hacer una predicción o clasificación. Además, estos modelos de “caja negra” hacen que sea difícil derivar exactamente qué datos influyeron en la decisión. Por lo tanto, es cada vez más difícil mostrar adherencia al principio de minimización de datos.

IBM propone una herramienta cuyo código fuente en *Phyton* está disponible en *GitHub* (1) para la minimización de datos que puede reducir la cantidad y la granularidad de los datos de entrada utilizados para realizar predicciones mediante modelos de aprendizaje automático para realizar la clasificación o la predicción, ya sea mediante la eliminación (supresión) o la generalización.

Actualmente admite la minimización de los datos recién recopilados para el análisis (es decir, datos de tiempo de ejecución), no los datos utilizados para entrenar el modelo, aunque consideran extenderlo en el futuro. El tipo de minimización de datos que realizamos implica la reducción del número y/o granularidad de las características recopiladas para el análisis. Las funciones pueden suprimirse (eliminarse) por completo o generalizarse. La generalización significa reemplazar un valor con un valor menos específico, pero semánticamente consistente. Por ejemplo, en lugar de una edad exacta, representada por el dominio de los números enteros entre 0 y 120, una edad generalizada puede consistir en rangos de 10 años.

Este método no requiere volver a entrenar el modelo y ni siquiera asume la disponibilidad de los datos de entrenamiento originales. Por lo tanto, proporciona una solución simple y práctica para abordar la minimización de datos en los sistemas existentes.

(1) https://github.com/IBM/ai-privacy-toolkit

Evaluación del riesgo de privacidad

Estudios recientes muestran que un tercero malintencionado con acceso a un modelo de aprendizaje automático entrenado, incluso sin acceso a los datos de entrenamiento en sí, aún puede revelar información personal confidencial sobre las personas cuyos datos se usaron para entrenar el modelo. Por ejemplo, puede ser posible revelar si los datos de una persona son o no parte del conjunto de entrenamiento del modelo (inferencia de membresía), o incluso inferir atributos confidenciales sobre ellos, como su salario (inferencia de atributo).

Por lo tanto, es crucial poder evaluar los modelos de IA de riesgo de privacidad que pueden contener información personal antes de que se implementen, dando tiempo para aplicar estrategias de mitigación adecuadas. Tales evaluaciones también son importantes para permitir comparar y elegir entre diferentes modelos de ML en función no solo de la precisión sino también del riesgo de privacidad, y así tomar una decisión informada sobre qué modelo es el más adecuado para un caso de uso determinado.

La evaluación del riesgo de privacidad puede basarse en tres tipos de información:

- Resultados empíricos de aplicar ataques de inferencia al modelo.
- Computar puntajes de métricas de fuga de membresía/privacidad.
- Analizando los factores de riesgo que se han encontrado asociados con un mayor riesgo de privacidad.

El objetivo es eventualmente poder calcular un puntaje general de riesgo de privacidad basado en estas diferentes dimensiones y componentes básicos.

Un ejemplo para la evaluación del riesgo de privacidad puedes ser la evaluación de privacidad de un modelo de evaluación de hipotecas, MDA (1).

Recordamos que los tipos de ataque son los mencionados en el principio ético de Robustez tratado por IBM en este mismo documento.

Materiales adicionales:

- Montaje de un ataque de inferencia de membresía (2).
- Montaje de un ataque de inferencia de atributos (3).
- Montaje de un ataque de reconstrucción de base de datos (4).
- Montaje de un ataque de inversión de modelo (5).

(1) https://aifs360.mybluemix.net/examples/hmda_privacy?_ga=2.11842614.325754436.1642231065-1367751274.1637864447

(2) https://github.com/Trusted-AI/adversarial-robustness-toolbox/blob/main/notebooks/attack_membership_inference.ipynb

(3) https://github.com/Trusted-AI/adversarial-robustness-toolbox/blob/main/notebooks/attack_attribute_inference.ipynb

(4) https://github.com/Trusted-AI/adversarial-robustness-toolbox/blob/main/notebooks/attack_database_reconstruction.ipynb

(5) https://github.com/Trusted-AI/adversarial-robustness-toolbox/blob/main/notebooks/model_inversion_attacks_mnist.ipynb

El enfoque de IBM

Privacidad en el aprendizaje federado

El aprendizaje federado (FL) es un enfoque del aprendizaje automático en el que los datos de entrenamiento no se gestionan de forma centralizada. Los datos son retenidos por las partes que participan en el proceso de FL y no se comparten con ninguna otra entidad

Así pueden trabajar juntos para entrenar un modelo de forma colaborativa sin compartir datos de entrenamiento, sino intercambiando y fusionando los parámetros de modelos entrenados localmente (*Main paper* de IBM en 1).

Esto hace que FL (Código *Python* y API en 2, 3 y 4) sea una solución cada vez más popular para la tarea de aprendizaje automático en la que reunir datos en un repositorio de datos centralizado es problemático, ya sea por razones de privacidad, reglamentarias, de confidencialidad o prácticas. Este enfoque funciona para proteger los datos de los consumidores en teléfonos inteligentes, así como en centros de datos en diferentes países y todo lo demás. En este enlace puedes seguir una sencilla demostración (5).

Para una mayor privacidad, este enfoque se puede combinar con otras técnicas, como la privacidad diferencial, el cifrado homomórfico y la computación multipartita segura.

Sin embargo, los modelos entrenados de forma federada siguen siendo vulnerables a los ataques de inferencia que reconstruyen los datos de entrenamiento durante el proceso de aprendizaje o en el modelo final. Las combinaciones de técnicas de privacidad, como el ruido diferencialmente privado, el cifrado homomórfico y el cómputo seguro de múltiples partes, mejoran la privacidad del aprendizaje federado más allá de simplemente no compartir datos.

(1) IBM Federated Learning: An Enterprise Framework.White Paper V0.1

https://arxiv.org/pdf/2007.10987.pdf

(2) Código Python. Aprendizaje federado de IBM (edición comunitaria)

https://github.com/IBM/federated-learning-lib

(3) Código Python. IBM CloudPak for Data (versión preliminar técnica)

https://dataplatform.cloud.ibm.com/docs/content/wsj/analyze-data/fed-lea.html?_ga=2.44651302.325754436.1642231065-1367751274.1637864447

(4) Documentación API

http://ibmfl-api-docs.mybluemix.net/index.html?_ga=2.44651302.325754436.1642231065-1367751274.1637864447

(5) https://mnist-ffl-demo.mybluemix.net/?_ga=2.49967268.325754436.1642231065-1367751274.1637864447

Las referencias descritas a continuación plantean diferentes enfoques prácticos sobre aprendizaje federado:

- HybridAlpha: An Efficient Approach for Privacy-Preserving Federated Learning.
  https://arxiv.org/abs/1912.05897
- Un enfoque híbrido para el aprendizaje federado que preserva la privacidad.
  https://arxiv.org/abs/1812.03224
- Un enfoque sintáctico para el aprendizaje federado que preserva la privacidad.
  http://ecai2020.eu/papers/1591_paper.pdf
- Secure Model Fusion para el aprendizaje distribuido mediante el cifrado homomórfico parcial.
  https://link.springer.com/chapter/10.1007/978-3-030-17277-0_9
- Análisis del aprendizaje federado a través de una lente antagónica.
  https://arxiv.org/abs/1811.12470
- Hacia el aprendizaje gráfico federado para la detección colaborativa de delitos financieros.
  https://arxiv.org/abs/1909.12946
- Aprendizaje federado con privacidad diferencial para datos de salud confidenciales.
  https://arxiv.org/abs/1910.02578
- Predicción de reacciones adversas a medicamentos en datos de salud distribuidos mediante el aprendizaje federado.
  https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7153050/
- ¿Maldición o redención? Cómo la heterogeneidad de los datos afecta la solidez del aprendizaje federado.
  https://arxiv.org/abs/2102.00655
- Compartiendo modelos o conjuntos básicos: un estudio basado en el ataque de inferencia de membresía.
  https://arxiv.org/abs/2007.02977
- FedV: Aprendizaje federado que preserva la privacidad sobre datos particionados verticalmente.
  https://arxiv.org/pdf/2103.03918.pdf
- Aprendizaje automático federado responsable en el gobierno: perspectivas de ingeniería y gestión.
  https://link.springer.com/chapter/10.1007/978-3-030-82824-0_10

Resumen en un vistazo del Toolkit de IBM para la gestión de los principios éticos

	Governance	Tecnologías	
Principio	Tools	Open Source	Productos
Fairness		AI Fairness 360 Casual inference 360*	Watson openscale
Explainability		Casual inference 360* AI Explainability 360	
Robustness		Adversial Robustness 360	
Transparency	AI FactSheets 360	Uncertainty Quantification 360	
Privacy		AI Privacy 360	

* IBM Causal Inference 360, permite gestionar la inferencia Causal, un concepto relacionado con la Equidad y la Explicabilidad, aunque el fin último de este toolkit no sea gestionarlas.

3. Framework GuIA

Cómo aterrizar cada principio ético

Caso de éxito de Telefónica

Caso de éxito de Telefónica

Telefónica, con una historia casi centenaria, es una de las multinacionales emblemáticas que han acompañado al mercado español e internacional a lo largo de los años. Desde sus orígenes en 1920, los servicios de telecomunicaciones y las infraestructuras en las que se basan han evolucionado hasta convertirse en los ejes básicos de una innovación, sostenida por los datos. La compañía se ha adaptado proactivamente a la secuencia de tecnologías disruptivas, desde la adopción temprana de *big data*, hasta reconocidas iniciativas de datos e inteligencia artificial (IA) como la 4ª Plataforma, y el rol pionero de la compañía en términos de gobernanza de datos e inteligencia artificial.

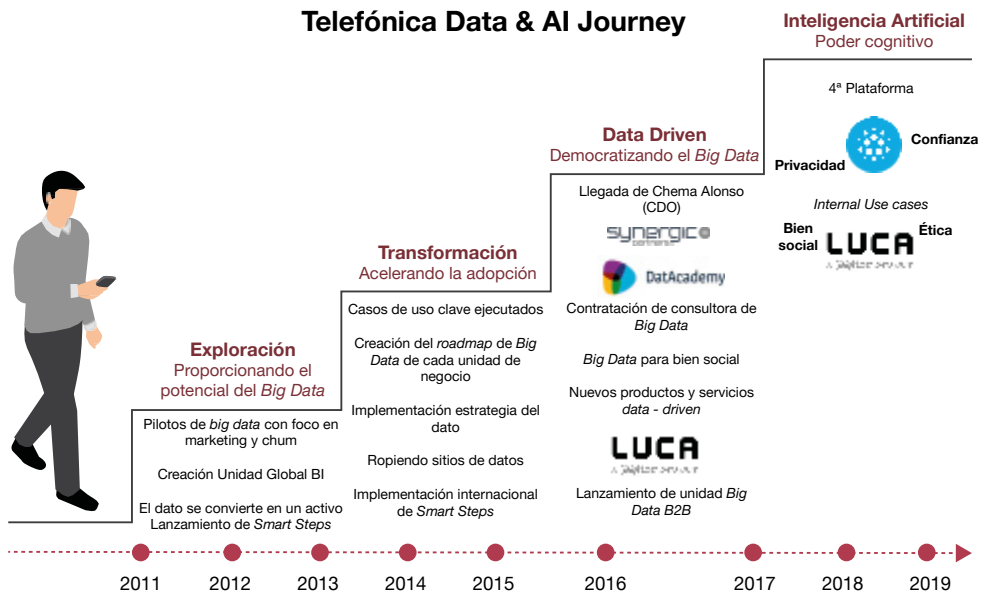
Este caso de éxito pretende contextualizar el rol de Telefónica como empresa de innovación y pionera en materia de inteligencia artificial, con un foco especial en su enfoque propio de IA responsable y la contribución al estado del arte de dicha materia, que es para la mayor parte de empresas una tarea pendiente, tanto a nivel de industria nacional e internacional, como en comparación con otros actores relevantes de la industria de telecomunicaciones.

Contexto: Orígenes e IA responsable en Telefónica

La historia de Telefónica como empresa tecnológica , que va más allá de las actividades puras de telecomunicaciones, no puede ser entendida sin una estrategia de datos, como los generados por las redes, aquellos relacionados a los usuarios de los servicios, y el conocimiento generado, a modo de *insights*. Todos ellos han sido fuente y motor del resto de actividades de la marca a lo largo de los últimos años.

La relación entre Telefónica y los datos (tradicionalmente conocidos como *big data*) no es nueva. El periodo entre 2011 y 2015 supuso el paso inicial entre la empresa que hasta entonces era considerada un actor relevante de la industria *telco* y la transformación total para dar lugar a una empresa de innovación en datos e inteligencia artificial. La experiencia interna obtenida con las iniciativas de *big data* dio lugar en 2016 a la creación de una unidad de negocio para proyectos con clientes (LUCA), integrada ahora en Telefónica Tech, y la creación de un nuevo rol ejecutivo: el *Chief Data Officer* con Chema Alonso a la cabeza, la que en su día fue una decisión pionera del actual Presidente de la empresa, José María Álvarez-Pallete.

Trayectoria de Telefónica entre 2011 y 2019



Fuente: Telefónica

Esa nueva etapa dio lugar a la Telefónica actual, la misma que atrae la atención en el *Mobile World Congress* (MWC), año tras año, con presentaciones vanguardistas, y con la innovación como ADN. Esta innovación le permite aprovechar todas las oportunidades de progreso y de negocio que ofrece el ecosistema digital actual, caracterizado por su dinamismo y rapidez, y con la que ha cambiado para siempre la manera de entender los datos y la relación entre la *telco* y sus clientes, siendo la 4ª Plataforma su exponente máximo y todos los servicios digitales que se crean sobre la misma usando sus capacidades.

La 4ª Plataforma, es la plataforma tecnológica en el corazón de la digitalización de Telefónica que permite disponer de una visión de extremo a extremo de cliente. Para conseguirlo, se apoya sobre distintas capacidades de gestión de identidad, de "apificación " (APIs) y de abstracción de datos que implementa bajo la máxima de privacidad-por-diseño. Con este ecosistema de capacidades que ofrece, se pueden desarrollar aplicaciones, productos y servicios digitales más rápido, con menos fricción y además habilita la creación de productos que incorporan inteligencia artificial para enriquecer la experiencia del cliente en Telefónica. Algunos de estos productos son: Aura, el asistente virtual con inteligencia artificial de Telefónica; y la aplicación Mi Movistar. Ambos ofrecen una nueva forma de relación con los clientes para gestionar su experiencia digital con la compañía. Soluciones a las que hay que sumar otras como las *Living Apps*, aplicaciones integradas en Movistar Plus+, que ofrecen una nueva experiencia de consumo desde la televisión o soluciones de mercado con terceros a través de Telefónica *Tech*, líder en transformación digital que integra, entre otros, el negocio de *Big Data* e IoT del Grupo Telefónica.

La visión de Telefónica: principios de inteligencia artificial

Telefónica hace girar todas sus acciones sobre los Objetivos de Desarrollo Sostenible (ODS), tal y como indica en sus Principios de negocio responsable y en la Política de derechos humanos. Esto le permite asegurar la sostenibilidad del negocio además de poder aportar, de forma significativa, a un proyecto determinante para el futuro de las personas, favoreciendo el progreso social y económico a través de la digitalización a la par que generar confianza para asegurar una transición digital centrada en las personas.

La premisa principal es que la tecnología debe contribuir a crear una sociedad más inclusiva y ofrecer mejores oportunidades para todos, sin dejar a nadie atrás y la inteligencia artificial puede contribuir a estos objetivos. Con el fin de guiar a la empresa y sus empleados en su aplicación de la inteligencia artificial y el *big data* en todas las líneas de negocio, Telefónica publicó en 2018 sus «Principios de IA»:

Principios de la inteligencia artificial según Telefónica

Justa	Transparente y explicable	Centrada en las personas	Privacidad y seguridad desde el diseño	Trabajo con terceros y socios
Nos aseguraremos de que nuestras aplicaciones de IA produzcan resultados justos, sin conducir a impactos injustos o discriminatorios. Reduciremos al mínimo la probabilidad de reforzar sesgos y discriminaciones injustas.	Seremos explícitos y transparentes en el uso de datos personales, así como en su finalidad. Garantizaremos la comprensión de las decisiones que tome nuestro sistema de IA. Seremos transparentes cuando las personas interactúen directamente con uns sistema de IA.	Nuestra IA deberá estar al servicio de la sociedad y generar beneficios tangibles para las personas. La implementación de la IA en nuestros productos y servicios no debe, en ningún caso, provocar un impacto negativo en los derechos humanos o en el logro de los Objetivos de Desarrollo Sostenible de la ONU.	Los aspectos de seguridad serán una parte inherente en el ciclo de vida de los sistema de IA para garantizar el derecho de las personas a la privacidad y a sus datos personales.	Nos comprometemos a que los proveedores y terceros con los que trabajamos cumplán con nuestros principios de IA.

Fuente: Telefónica

Caso de éxito de Telefónica

El motivo de la creación de dichos principios fue justamente señalar la importancia de la ética para la inteligencia artificial, la cual, según Telefónica, no puede estar completa si no es acompañada de dicha visión. Cada uno de los cinco principios tiene un razonamiento específico:

	Principios	Razonamiento
1	Justicia	Las aplicaciones de la tecnología de IA deben dar resultados justos, sin impactos discriminatorios en relación con la raza, el origen étnico, la religión, el género, la orientación sexual, la discapacidad o cualquier otra condición personal.
2	Transparencia y explicabilidad	El objetivo es que los usuarios sepan que están interactuando con un sistema de IA, qué datos suyos se usan y para qué. Telefónica se asegurará de comprender la lógica que hay detrás de las decisiones del sistema.
3	Centrada en las personas	La IA debe estar al servicio de la sociedad y generar beneficios tangibles para las personas, cuyos derechos humanos no pueden verse vulnerados. Además, la compañía se ha propuesto ayudar a cumplir los Objetivos de Desarrollo Sostenible (ODS) con la IA.
4	Privacidad y seguridad desde el diseño	Las Políticas de Privacidad y Seguridad de la Compañía cobran en estos Principios especial relevancia para preservar los datos tanto personales como anónimos y agregados.
5	Trabajando con socios y terceros	Telefónica se compromete a verificar la lógica y los datos utilizados por los proveedores.

Para más información sobre los Principios de IA Responsable en Telefonica:

- Octubre 2018 - Telefónica *AI Principles*

https://www.telefonica.com/wp-content/uploads/sites/7/2021/11/principios-ai-eng-2018.pdf

- Infografía sobre los Principios de IA Responsable en Telefónica

https://www.telefonica.com/es/wp-content/uploads/sites/4/2021/03/infografia-ia-esp.pdf

Estos principios de IA ética se enmarcan dentro de los principios de Diseño Responsable de Telefónica:

Principios de diseño responsable

Ética aplicada al diseño

Principios de responsabilidad con el cliente

- Simplicidad
- Transparencia
- Integridad

Principios éticos aplicados a inteligencia artificial y gestión de datos

- Justa
- Transparente y explicable
- Con las personas como prioridad
- Privacidad y seguridad desde el diseño
- Socios y terceras partes

Sostenibilidad aplicada al diseño

Diseño considerando su impacto en los DD.HH

- Accesibilidad de la aplicación web
- Tecnologías digitales

Impacto en el medioambiente

- Ecodiseño
- Residuos
- Consumo energético

Fuente: Telefónica

La presentación de la alianza estratégica entre Telefónica y Microsoft para impulsar la inteligencia artificial y crear nuevos servicios, con los máximos responsables de ambas empresas presentes en el MWC'19, ilustró un paso más en la etapa de madurez IA de Telefónica, caracterizada como *Cognitive Intelligence* o inteligencia cognitiva por la empresa. Desde entonces, la adopción de la inteligencia artificial ha continuado como una tendencia natural, y ha habido aportaciones muy significativas en el ámbito de la ética y la IA responsable.

En los dos últimos años, hemos visto un auge general de los conceptos de IA responsable, los cuales partieron de una realidad teórica que ha ido evolucionando hacia prácticas más aterrizadas. Telefónica no ha sido una excepción, y ha trabajado en la definición de nuevas prácticas de gobernanza y nuevos roles relacionados, los cuales vamos a explorar en la sección 3.

Factor diferenciador: Gobernanza IA interna

La gobernanza de inteligencia artificial y el enfoque propio de Telefónica son motivos suficientes para este caso de éxito que nos ocupa. Durante los últimos tres años, la empresa ha desarrollado un sistema de IA responsable con tres niveles, en los que cada persona tiene la capacidad para analizar posibles riesgos potenciales.

Los puntos principales de este sistema de gobernanza de inteligencia artificial Responsable son:

- Los principios IA como punto de partida para cualquier implementación que incluya inteligencia artificial y/o datos masivos.
- La introducción del modelo de gobernanza como parte clave de los procesos oficiales de desarrollo de productos y servicios.
- La creación de un cuestionario autoevaluación con el objetivo de que los responsables de producto, junto a los equipos de desarrollo, lo completen durante la fase de diseño de los productos y servicios que utilizan inteligencia artificial y a posteriori en sucesivas fases durante todo el ciclo de vida del producto.
- La definición de un proceso de escalado entre niveles para garantizar el análisis detallado.
- La figura de los *RAI Champions*, como parte del proceso de gobierno de la IA Ética, los cuales están cerca tanto de las unidades de negocio como de los equipos de desarrolladores.

Rol del *Responsible AI (RAI) Champion*

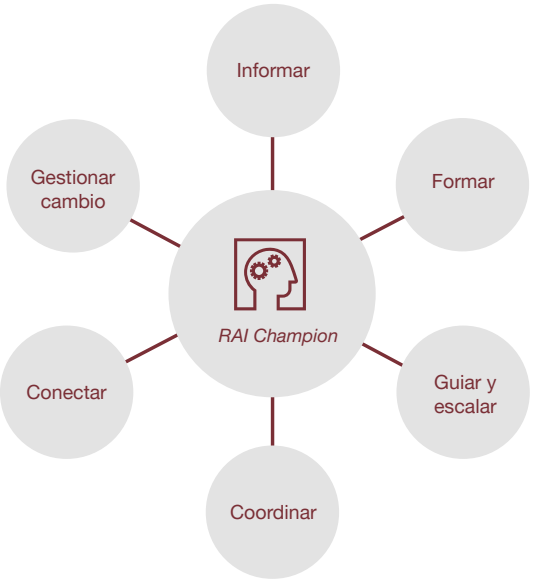
La figura del *Champion* de IA responsable forma parte de la nueva granularidad de roles relacionados a las iniciativas de inteligencia artificial, más allá de los casos típicos de científicos e ingenieros de datos. El *RAI Champion* es más un rol que un puesto especifico dentro de una organización, por lo general ocupado por profesionales con perfiles muy diversos, desde científicos de datos, responsables de privacidad, hasta la responsabilidad social de la empresa, con mayor o menor conocimiento técnico, pero que traen un punto de vista diferente a la hora de analizar el impacto ético de un proyecto de inteligencia artificial.

Este rol lo utilizan también algunas empresas que están muy avanzadas en la implementación de la IA responsable, pero Telefónica ha adaptado dicho perfil a los procesos y necesidades internas. La premisa principal es que se necesita un *RAI Champion* siempre que la unidad de negocio use o planee utilizar inteligencia artificial o sistemas de datos masivos en sus soluciones. Si el nivel de adopción es intensivo, puede haber más de un *Champion*.

Las funciones de un RAI Champion incluyen:

- Informar sobre la importancia de aplicar los principios de inteligencia artificial.
- Formar a sus áreas de influencia sobre cómo aplicar dichos principios de inteligencia artificial y el modelo de gobernanza.
- Guiar y escalar, y ser el punto de referencia para dudas y cuestiones éticas, así como escalar al grupo de expertos cuando sea necesario.
- Coordinar con diferentes áreas como Privacidad, Seguridad, *Chief Data Office*, Responsabilidad Social Corporativa, Legal, etc.
- Conectar y fomentar la construcción de una comunidad de expertos en inteligencia artificial IA ética.
- Gestionar el cambio, para que las consideraciones éticas sean parte del *business as usual* de los equipos.

Funciones del *Responsible AI Champion*



Fuente: Telefónica

Sistema de escalado en tres niveles

Tal y como hemos comentado anteriormente, el escalado forma parte del alcance de actividades del *RAI Champion*. Dicho escalado no se corresponde a la jerarquía tradicional de empresa, sino que Telefónica ha creado una estructura dedicada a la IA ética y responsable. No hay muchas empresas que hayan alcanzado dicho nivel de madurez interna en el mundo, y en el caso de Telefónica incluye no sólo los *RAI Champions* de cada área (definidos como el nivel 1 del modelo), sino un equipo *core RAI* (nivel 2), que coordina a todos los *RAI Champions* y otros actores como CDO (*Chief Data Office*), RSC (Responsabilidad Social Corporativa) o DPO (*Data Privacy Office*), y la Oficina de Negocio Responsable (nivel 3). Concretamente, la relación entre niveles y los factores de escalado son:

- Nivel 1 a 2: Si hay dudas en torno a las preguntas o resultados del cuestionario (detalles en la sección 4), y dichas dudas no pueden ser resueltas por el *RAI Champion* o su comunidad de expertos, entonces el caso se escala al nivel del equipo *core RAI* para consulta.
- Nivel 2 a 3: En el caso extraordinario de que el equipo *core RAI* y su red ampliada de expertos multidisciplinares no puedan encontrar una solución, se procederá al nivel final de escalado, en el que la Oficina de Negocio Responsable podrá intervenir y solucionar el caso.

Dicho sistema promueve la responsabilidad individual, y una habilidad de escalado bajo demanda, en vez de un sistema tradicional de aprobación *top-down*. Se parte de la base de que todo el mundo quiere implementar una IA ética y responsable, y para ello se ponen a disposición una serie de herramientas explicadas en la sección 4.

Para más información sobre el modelo de gobierno:

- Junio 2021 - *Telefónica´s Approach to the Responsible Use of AI*

https://www.telefonica.com/es/wp-content/uploads/sites/4/2021/06/ia-responsible-governance.pdf

Implementación: De los principios éticos a la acción

Punto clave 1. Formación de empleados como herramienta de conocimiento y concienciación

Siendo la inteligencia artificial y la ética en inteligencia artificial campos de conocimiento relativamente nuevos para muchas personas, Telefónica ha dado el peso necesario a la formación interna para democratizar el acceso a dicho conocimiento, y para que cada uno de los empleados entienda la importancia absoluta de la ética en la inteligencia artificial. Hoy en día, este programa de formación ha formado a más de 3.500 empleados de la compañía, y el numero sigue creciendo.

Los aspectos cubiertos en dicha formación incluyen los fundamentos de la inteligencia artificial y los datos masivos, los principios de IA Responsable de Telefónica y su implementación en función del tipo de proyecto, el sistema de gobernanza y los roles en el día a día, las herramientas disponibles para implementar una IA ética y responsable, y por supuesto ejemplos prácticos para aterrizar los conceptos aprendidos.

Este programa, a la altura de cualquier iniciativa de formación de calidad en IA ética y responsable, permite desarrollar el conocimiento y un entendimiento común de los principios, el argot, y el potencial realista de la inteligencia artificial y los datos masivos.

Formación interna de IA y ética

Fuente: Telefónica

Punto clave 2. El cuestionario de evaluación como punto de partida para el análisis ético por proyecto

Después de la formación, los empleados tienen la información y el conocimiento necesario para aplicar los principios en sus proyectos. Con el fin de facilitar el análisis, Telefónica ha creado un cuestionario dinámico que debe ser integrado en el diseño de productos y servicios que utilizan inteligencia artificial.

Este cuestionario está disponible a través de un acceso en línea y contiene una serie de preguntas relevantes sobre cada uno de los principios IA de Telefónica. Dichas preguntas aportan una reflexión en función de las respuestas introducidas y recomendaciones para garantizar la ética y la responsabilidad entre las fases de diseño e implementación. Este cuestionario es completado por los responsables del producto o servicio a analizar.

Entre otras, las cuestiones se centran en:

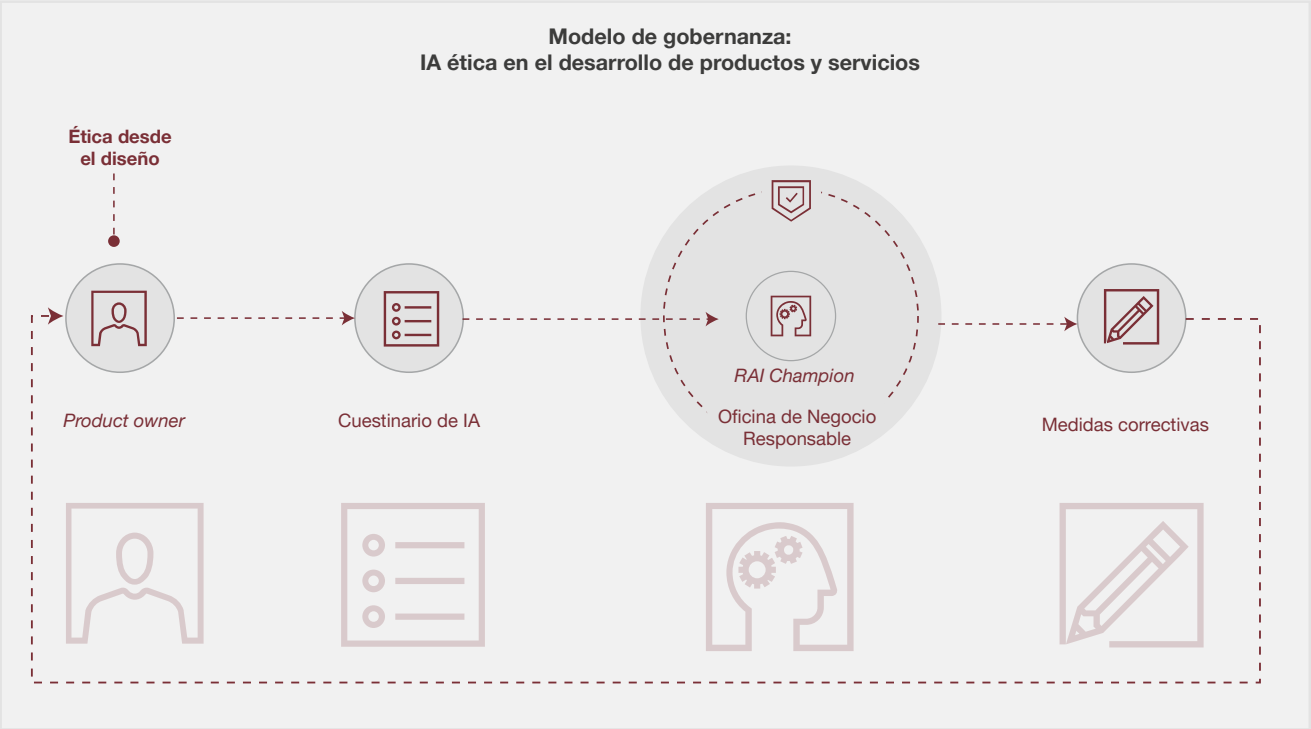
- Los sets de datos y sus variables, con foco especial en aquellas que puedan ser sensibles (por ejemplo, el género o la nacionalidad de una persona, o la presencia de grupos minoritarios).
- La correlación entre variables, como parte del ejercicio de preparación de los datos que permite reducir datos innecesarios que puedan introducir sesgos en los modelos.
- La explicabilidad de los sistemas y el posible impacto en la vida de las personas.
- Las cuestiones de privacidad de datos y el cumplimiento de la regulación.
- La relación con sistemas de terceras partes, y su alineamiento con los principios IA de Telefónica.

Por último, y tal como hemos visto en la sección 3, el *RAI Champion* correspondiente solventa dudas, propone medidas, e incluso escala el caso si es necesario.

Punto clave 3. Inclusión de la gobernanza como parte del proceso oficial de desarrollo de productos

El sistema de gobernanza IA responsable pretende privilegiar un enfoque: *Responsible AI first* en el que todo tipo de *stakeholders* (desde los más técnicos, hasta las unidades de negocio y los niveles ejecutivos) formen parte del análisis y la reflexión para garantizar una implementación ética y responsable de la IA en los productos y servicios. Este sistema es una realidad del día a día en Telefónica, y alinea a toda la empresa en torno a los procesos y principios previamente definidos. Telefónica utiliza los procesos ya existentes de desarrollo de productos y servicios, para integrar como una fase más del desarrollo, esta evaluación de los principios de IA Ética siempre que el producto o servicio incorpore inteligencia artificial o datos masivos.

Modelo de gobernanza de IA ética



Fuente: Telefónica

Este sistema, compartido y adoptado internamente, en conjunto con el resto de las herramientas y acciones formativas previamente explicadas, permite tener un enfoque estandarizado y validado que garantice la implementación eficaz del sistema en el día a día y que tenga en consideración la ética en todas las fases del desarrollo del producto, partiendo del diseño. Esta madurez de pensamiento y organización ha permitido a Telefónica y sus empleados desarrollar una experiencia muy valiosa, con aprendizajes interesantes que van a ser presentados en la sección 5.

Aprendizajes y recomendaciones

Si consideramos el camino de Telefónica para convertirse en una empresa de datos e inteligencia artificial desde 2011 (sección 1), así como el nivel de innovación y soluciones presentadas a lo largo de los años, es justo pensar que Telefónica ha desarrollado un conocimiento y una opinión formada en torno a las mejores prácticas para facilitar una adopción ética y responsable de la inteligencia artificial. Pocos ejemplos nacionales pueden ilustrar mejor este tipo de recorrido y sus aprendizajes.

Aprendizajes clave de Telefónica

Desde las áreas de *Chief Data Office* y *Chief Responsability Office* nos comparten sus reflexiones finales:

- Los principios de IA son un excelente punto de partida para unificar el enfoque de innovación de inteligencia artificial responsable y ética en cualquier empresa. Cualquier compañía que comience su andadura en el mundo de los datos e inteligencia artificial debe adoptar unos principios, se recomienda crear un set propio de principios, en base a los objetivos de la empresa, contexto de negocio, valores y visión de la compañía, y el uso o aplicación específico de la inteligencia artificial.
- Definir y adoptar un modelo de gobernanza de IA responsable permite aterrizar los conceptos, y facilita la implementación en el día a día de la empresa. Introducir dicha gobernanza en los procesos oficiales de desarrollo de productos es clave, y una metodología de éxito probada durante los últimos años por Telefónica.
- Formar a los empleados es clave para garantizar la comprensión de los conceptos y del modelo. Se deben privilegiar formaciones híbridas que cubran aspectos técnicos y de negocio, y que sean accesibles a cualquier empleado de la empresa, más allá de su rol, como *adopters* potenciales de la inteligencia artificial y los datos masivos en sus productos y servicios.
- Crear la figura de los *RAI Champions* y sus comunidades de conocimiento permiten tener un punto de referencia para la adopción correcta de una inteligencia artificial responsable y ética. Dichos *Champions* son embajadores de la visión y los principios de inteligencia artificial, y suelen adoptar ese rol como parte de sus responsabilidades, sin que ello suponga una dedicación exclusiva. Las ganas de continuar aprendiendo y ayudando a los equipos de desarrollo y de negocio son la motivación clave para este tipo de profesionales.

Beneficios de la IA ética y responsable

Desde Telefónica conocen la importancia que tienen tanto sus grupos de interés (clientes, proveedores, empleados y la sociedad en su conjunto), como sus accionistas e inversores. Estos son cada vez más conscientes de la importancia del desarrollo de una tecnología ética y responsable, así como de los crecientes requerimientos y exigencias a tener en cuenta.

Por todo lo anterior, Telefónica fue una de las primeras empresas en el mundo en definir un marco de principios éticos de inteligencia artificial, que permitiera, además de hacer las cosas bien, adelantarse a los posibles riesgos de cumplimiento, así como amenazas, que pudieran surgir, reforzando así especialmente sus sistemas de control y prevención de riesgos.

Compañías como Telefónica que cuenten con un sólido sistema de control de sus riesgos en materia de IA, tendrán un especial atractivo e interés para accionistas e inversores que desde hace años recalcan la importancia de los aspectos ESG (*Environmental*, *Social and Governance*) en la estrategia de las compañías.

Caso de éxito de Telefónica

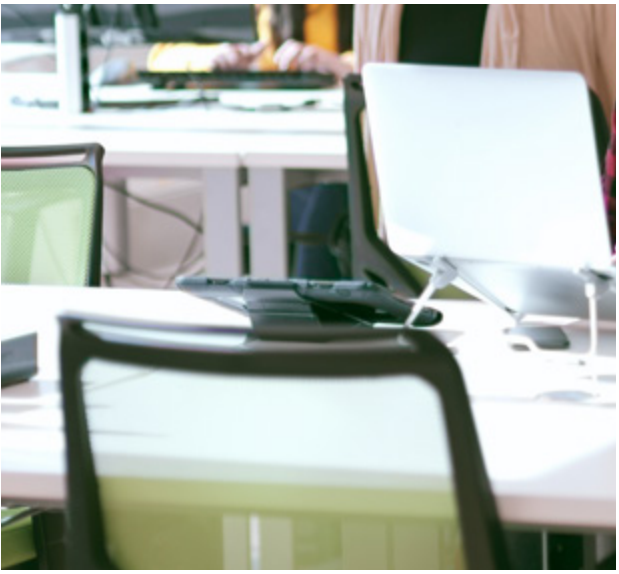
La tecnología es muy poderosa por lo que es importante marcarle el camino. En Telefónica aplicamos la ética a la inteligencia artificial con el fin de aprovechar todo su potencial sin consecuencias negativas. Para ello contamos con la metodología Diseño Responsable, que nos permite formar a nuestros empleados para evitar discriminaciones en los algoritmos, respetar al máximo la privacidad y procurar que nuestra IA sea justa, explicable y beneficiosa para las personas.

Elena Valderrábano
Global Chief Sustainability (ESG) *Officer*, GCSO

En Telefónica, nuestros principios sitúan a las personas en el centro y garantizan el respeto de los derechos humanos en cualquier entorno y proceso en el que se use la inteligencia artificial. Así, estamos trabajando para implementar, sobre la base ya existente de Privacidad por Diseño, una capa adicional que denominamos Ética por Diseño. En ella, implementamos técnicamente los valores y principios éticos durante el ciclo de vida de nuestros productos de inteligencia artificial, además de definir las técnicas organizativas necesarias para su consecución.

Francisco Montalvo
Chief Data Officer, CDO

4



Próximos pasos: Adaptación sectorial y Formación

El objetivo que recorre toda la iniciativa GuIA como un mantra es **aterrizar**. Ese objetivo lo hemos perseguido a lo largo del presente documento creando el *FrameWork* GuIA. Un marco global que integra de manera trazable tres marcos: el ético, el jurídico y el tecnológico.

De esta manera pretendemos ayudar a las empresas a adaptar de manera pragmática la inteligencia artificial ética y normativa en su día a día. **Aterrizar** los conceptos éticos mediante recomendaciones, *guidelines*, tecnologías y modelos de gobierno que ayuden a las empresas a saber **cómo hacerlo**. En definitiva, un **aterrizaje** desde el principio y su normativa asociada hasta, en muchos casos, el algoritmo necesario para implementarlo, quedándonos a las puertas del *coding*.

Pero hay dos pilares estratégicos más de GuIA que persiguen este objetivo de aterrizaje: **la adaptación sectorial y la formación**.

Adaptación Sectorial, aportando además un valor de negocio

Este documento GuIA es una base sólida y válida para todos los sectores empresariales. Sin embargo, cada sector necesita de sus necesidades particulares. Por ejemplo:

- Un sector regulado necesita de un mayor soporte en el marco jurídico.
- Los principios éticos no son los mismos en un sector industrial donde prevalece la componente “máquina” física gestionada mediante IA, que en el sector *retail* donde debiera prevalecer la sensibilidad de los consumidores, o en el sector salud donde el uso de la IA puede llegar a salvar vidas.

Además, incluso cada caso de uso dentro de un sector tiene sus propias necesidades éticas, normativas, tecnológicas, de gobierno, etc., etc.

Esta GuIA ha sido diseñada de manera modular para poder enlazarla directamente con los sectores y sus casos de uso. Un ejercicio que ya hemos comenzado en la Fase 2 de GuIA, donde vamos a adaptar esta GuIA a 10 sectores empresariales de la mano de más de 50 empresas. Cerraremos así el círculo y **aterrizaremos también** los conceptos de IA ética desde un punto de vista de negocio.

Y en este ejercicio, el objetivo de **aterrizar** lo conseguiremos gracias a los **caso de uso**. Desde GuIA proporcionaremos a cada sector casos de uso que puedan ser soportados sobre IA, y los complementaremos con la experiencia y las necesidades de las empresas.

Co-crearemos un conjunto de casos de negocio donde la IA aporte un **valor real de negocio**, para después adaptar el *FrameWork* GuIA a cada caso de uso.

En esta Fase sectorial vamos a contar con:

- **Empresas grandes y pequeñas**, porque nuestra economía se sustenta en grandes compañías, en pymes, y también nuevas empresas que están creciendo para aportar valor a la economía española.
- **Empresas sectoriales** con amplio recorrido y conocimiento del negocio, que además comienzan a adoptar la inteligencia artificial y son sabedoras del valor añadido que la IA puede suponer.
- **Empresas tecnológicas verticales** a un sector en particular, que tienen un enorme *Know-How* acerca de cómo automatizar procesos concretos de negocio mediante inteligencia artificial dentro de dicho sector.

Inicio de los trabajos sectoriales

Ya hemos comenzado este trabajo en el marco de **INDESIA**, el consorcio nacional para la adopción de la inteligencia artificial en el **sector industrial**, liderado por Repsol, Gestamp, Navantia, Técnicas Reunidas, Telefónica, Microsoft, Airbus y Ferrovial, donde OdiseIA co-lidera el grupo de trabajo de inteligencia artificial ética junto a Telefónica y Microsoft.

Por otra parte, durante los próximos 18 meses abordaremos 10 sectores empresariales. Estamos comenzando con los sectores de Seguros, Salud y Publicidad. A estos tres sectores les seguirán otros siete más que consensuaremos con todos los actores involucrados en GuIA. Hemos elegido estos tres primeros sectores por varios motivos entre los que destacan:

- Seguros. Para contar con un sector regulado.
- Publicidad. Por la sensibilidad de la sociedad respecto a como se utilizan sus datos.
- Salud. Por el impacto social positivo que supone en la sociedad, puesto aún más de manifiesto durante la pandemia por covid-19.

Además, estos sectores están identificados en la ENIA (Estrategia Nacional de Inteligencia Artificial) publicada por la SEDIA (Secretaría de Estado de Digitalización e Inteligencia Artificial) como los sectores en los que más impacto puede tener la IA.

En esta fase ya han confirmado su presencia empresas como **Mapfre, Generali, Mutualidad de la abogacía, Verti, Bdeo, t2ó, Inspide, Quiron Salud y Savia**. Un conjunto de empresas de distinto tamaño y naturaleza como indicábamos anteriormente. Y que, al igual que ha pasado con las empresas Tecnológicas que han formado parte de esta primera fase cuyo resultado recoge el presente documento, tienen en muchos casos negocios solapados pero todas tienen un mismo interés común alineado con la misión de OdiseIA:

> Trabajar activamente por un uso ético y responsable de una inteligencia artificial que tenga en su centro el bienestar del ser humano

Cualquier empresa interesada en participar en estos sectores, o en alguno de los *site* restantes que publicaremos en el mes de Marzo, puede dirigirse a guia.contacto@odiseia.org

Formación

El otro pilar estratégico de GuIA que permite nuestro objetivo global de aterrizaje es la Formación. La inteligencia artificial es una tecnología emergente, cuyo nivel de adopción está comenzando a desarrollarse, y cuyo crecimiento exponencial en los próximos años es ya una realidad tangible. Las empresas que no recorran ese camino corren el riesgo de perder competitividad en un mercado que cada vez es más rápido y no espera a los rezagados.

La inteligencia artificial ética es una disciplina aún más joven. Por tanto, para conseguir que estos conceptos fundamentales sean asimilados y **"aterricen"** en la mente de las personas de manera clara es necesario algo más que la elaboración de un informe como el presente. La IA ética no es un adendum opcional a la IA, y por tanto las empresas deben tomar consciencia de ello y **capacitar** a sus equipos relacionados con la IA.

La propia Secretaría de Estado de Digitalización e Inteligencia Artificial (SEDIA) es consciente de ello, y desde Abril a Julio de 2021 invitó a OdiseIA a impartir un programa formativo a más de 100 empleados públicos en IA en general e IA ética en particular. Un programa que tuvo una valoración media de 8,8 (sobre 10) por parte de los asistentes.

Por estos motivos en GuIA hemos llegado a acuerdos con escuelas de negocio y unversidades para formar a ejecutivos y universitarios, a los profesionales de hoy y del futuro.

Cualquier empresa y entidad académica (universidad, escuela de negocios, etc.) interesada en incorporar estos contenidos de alguna manera a sus planes formativos ya existentes o en nuevos programas generados ad-hoc puede contactar con guia.contacto@odiseia.org

5

Accede a nuestra comunidad GuIA

OdiseIA es una plataforma de colaboración abierta a todos los actores del ecosistema de la IA (empresas, administraciones públicas, centros de investigación, escuelas de negocio, universidades y profesionales independientes) que se identifican con nuestra misión.

> Trabajar activamente por un uso ético y responsable de una inteligencia artificial que tenga en su centro el bienestar del ser humano

La iniciativa GuIA es por tanto también una plataforma de colaboración abierta a cualquier de los mencionados actores.

Además del aterrizaje que tanto hemos comentado anteriormente, el otro objetivo de GuIA indicado en el prólogo del presente documento es crear un ecosistema donde cualquier entidad y profesional se pueda integrar para compartir y conocer las mejores prácticas aterrizadas en inteligencia artificial atendiendo a principios éticos y su normativa asociada.

El proceso de investigación que ha dado lugar a este informe y que dará lugar a los diez informes sectoriales que hemos detallado anteriormente, es realizado con un grupo reducido de unas cinco empresas por motivos operativos. Estas empresas son consensuadas entre todos los actores que forman la iniciativa. De hecho, los sectores de seguros, salud y publicidad están ya configurados bajo este criterio.

En el mes de Marzo haremos públicos los 7 sectores restantes. Cualquier empresa interesada en participar en dichos podrá dirigirse al buzón guia.contacto@odiseia.org, y nos pondremos en contacto para evaluar entre todos la mejor forma de colaborar.

Participar directamente en el proceso de investigación que dará lugar a estas GuIAs sectoriales no es la única forma de participar en GuIA. Ya hemos habilitado un canal de Linkedin donde:

- Todas las entidades y profesionales interesados puedan acceder a los documentos GuIA.
- Podrán compartir sus reflexiones sobre dichos contenidos. Así los mantendremos vivos y actualizados, y las reflexiones podrán ser incorporadas en futuras releases de las GuIAs.
- Estar al tanto de todas las novedades y eventos relacionados con GuIA.

LinkedIn

https://www.linkedin.com/showcase/odiseia-canalguia

Además, te animamos a seguir a OdiseIA en nuestras Redes Sociales:

https://www.linkedin.com/company/odiseia

https://twitter.com/odise_ia

https://www.youtube.com/c/OdiseIAImpactoSocialyÉticodelaIA

Y en nuestra web https://www.odiseia.org/

6

Anexos

- Equipo
- Qué es OdiseIA
- Qué es PwC

6. Anexos

Equipo

Este primer informe podría haber sido realizado por un eticista, un tecnólogo y un jurista. Pero ese planteamiento no habría cumplido con uno de los objetivo de GuIA: **crear comunidad** alrededor de las buenas prácticas aterrizadas en inteligencia artificial ética y normativa. El resultado no habría tenido la misma resonancia, y no habría un enriquecedor **trabajo colaborativo** que crease dicha comunidad.

Las personas indicadas a continuación han colaborado activamente en la elaboración de contenidos que han servido para generar esta primera GuIA. Algunas lo han hecho a través de sus empresas, otras a título personal. Todas han puesto su **granito de arena**, más o menos grande en función de su disponibilidad, para poder crear este documento. Todas han tenido **la misma ilusión**, y **son la semilla de una comunidad** que deseamos crezca exponencialmente en los próximos meses.

Muchas gracias a todas ellas.

Idea y Estrategia (OdiseIA)

- Idoia Salazar
- Juan Manuel Belloto Gómez
- Richard Benjamins

Conceptualización (OdiseIA)

- Juan Manuel Belloto Gómez

Coordinación global *Delivery*

- Juan Manuel Belloto Gómez (OdiseIA)
- Armando Martínez Polo (PwC España)

Equipo OdiseIA

- Adaya Esteban
- Adrián González Sánchez
- Adrián Palma
- Anna Danés
- Florence Byrd
- Idoia Salazar
- Juan Manuel Belloto Gómez
- Juan Murillo Arias
- Lorenz Cotino Hueso
- Paul Van Branteghem
- Pedro Albarracín
- Richard Benjamins
- Wilma Arellano

Equipo PwC España

- Aleix Catalán
- Alejandro Martínez Morillo
- Álvaro Blanco Barrios
- Ana Cendón
- Anand Rao
- Armando Martínez Polo
- Artem Motsarenko
- Asier Trapote Carpintero
- Bernat Rovira
- Britany Kerber
- Daniel Polanco Palacio
- Elena Ramírez Pérez
- Fina Sastre Coll
- Gonzalo Muelas Gironella
- Ilana Golbin
- Javier Barguño
- Julia Olano Jañez
- Maria Axente
- María Cumbreras
- Marta Llamazares Carreño
- Nacho de Gregorio Noblejas
- Nacho Mayero Mariscal de Gante
- Pablo Fernández Burgueño
- Patricia Manca Díaz
- Tamer Davut

Equipo Google

- Fernanda Viégas
- Josetxo Soria Checa
- Kathy Meier-Hellstern
- Mariana Carrillo
- Pilar Manchón

Equipo Microsoft

- Alberto Pinedo Lapeña
- Gabriel López

Equipo IBM

- Benito Martín Gassol
- Enric Delgado Samper
- Itziar Leguinazabal

Equipo Telefónica

- Francisco Montalvo
- Joaquina Salado
- Violeta Alburquerque

6. Anexos

Qué es OdiseIA

El Observatorio del impacto social y ético de la inteligencia artificial (OdiseIA), es una asociación independiente y sin ánimo de lucro fundada en 2019.

Somos una plataforma de colaboración abierta a todos los actores del ecosistema de la IA (empresas, administraciones públicas, centros de investigación, escuelas de negocio, universidades y profesionales independientes) que se identifiquen con nuestra misión:

> Trabajar activamente por un uso ético y responsable de una inteligencia artificial que tenga en su centro el bienestar del ser humano

Nuevos objetivos

- Impulsar la inteligencia artificial en España, atendiendo además a los principios éticos y normativos necesarios para su adopción. Mediante **acciones concretas**, más allá de necesarias reflexiones globales de las que también participamos.
- Ser la plataforma **hispanohablante** para el intercambio de experiencias y mejorar prácticas sobre el impacto social y ético de la inteligencia artificial.
- **Concienciar** a la sociedad, las organizaciones (empresas, administración pública, instituciones, etc.) sobre el impacto social y ético de la inteligencia artificial.
- Generar y recopilar **conocimiento** a través de la investigación y análisis en temas relacionados con la inteligencia artificial y su impacto social y ético.
- **Difusión y divulgación** del conocimiento generado
- Ejercer liderazgo de pensamiento en temas relacionados con la inteligencia artificial, su impacto social y ético, a través de artículos, organización de congresos/eventos y asesoría a instituciones públicas o privadas.

Organización y actividad

OdiseIA cuenta con más de 20 áreas:

- Las que ofrecen servicios de Investigación, Divulgación, Formación y Asesoramiento.
- Las que definen los desafíos (ética, normativa, sostenibilidad, inclusividad, etc.) sobre los servicios.
- Las que aportan la visión sectorial (salud, seguros, finanzas, administración pública, etc.) sobre dichos desafíos y servicios.

Miembros individuales

OdiseIA cuenta actualmente (Febrero de 2022) con más de 200 miembros, todos ellos profesionales de reconocido prestigio, con conocimientos y experiencia en materia de inteligencia artificial, e incorporados tras un riguroso proceso de evaluación y admisión que garantiza la excelencia de todo el equipo.

Un equipo multidisciplinar formado por tecnólogos, humanistas, juristas, expertos en negocio, docentes, comunicadores, etc., entre los que se encuentran miembros de la Real Academia de Ingeniería, premios *Pulitzer*, catedráticos de universidad, directivos de grandes empresas, reconocidos tecnólogos a nivel internacional, etc.

Miembros corporativos

OdiseIA cuenta además con 11 entidades como socios corporativos de toda naturaleza y tamaño, unidas bajo el interés común que persigue nuestra misión. Con ellos desarrollamos gran parte de nuestra actividad, apoyados por nuestros miembros individuales.

Las empresas y profesionales independientes interesados en incorporarse a OdiseIA pueden hacerlo a través de contacto@odiseia.org

6. Anexos

Qué es PwC

PwC cuenta con un equipo de profesionales altamente cualificados en diferentes ámbitos de actividad ofreciendo a nuestros clientes servicios de asesoramiento legal, fiscal, transacciones, consultoría y auditoría.

Bajo la perspectiva de la Inteligencia Artificial hemos trabajado en los últimos años en diferentes iniciativas para el asesoramiento en la definición de estrategias de Inteligencia Artificial, la construcción de modelos y soluciones cognitivas, el análisis de vulnerabilidades y riesgos de seguridad… lo que nos ha aportado una amplia visión sobre las consideraciones a tener en cuenta para una adecuada utilización de la Inteligencia Artificial.

Con la experiencia adquirida hemos ido desarrollando diferentes marcos de trabajo, persiguiendo una serie de objetivos, entre otros citar:

- Enfoque práctico y aterrizado orientado a la consecución de resultados
- Disponer de una sistemática de trabajo que permita guiar el desarrollo de las iniciativas
- Una visión completa cubriendo todas las dimensiones que consideramos relevantes (estrategia, gobierno, personas, operaciones y tecnologías)
- Aportar buenas prácticas que puedan ser aprovechadas por cualquier organización

Consideramos desde PwC que un adecuado uso ético y responsable de la Inteligencia Artificial es un elemento al que prestar una especial atención, de la misma forma que las organizaciones ya tienen perfectamente interiorizadas otras cuestiones como la protección de datos, la no explotación infantil, la no discriminación por diferentes aspectos… por citar algunas significativas.

Nuestra involucración en GuIA es una oportunidad brindada por OdiseIA, que nos ha parecido desde el primer minuto fundamental para poder aportar nuestro granito de arena a que un tema tan novedoso, y con tanto impacto en los próximos años, disponga de una orientación clara, esté abierta a cualquier destinatario y ofrezca una visión amplia fruto de la colaboración de un amplio número de profesionales de PwC España pertenecientes a diferentes disciplinas (legal, fiscal, tecnología, Analytics, Seguridad, auditoría…).

Asimismo, consideramos relevante señalar la existencia de una comunidad dentro de PwC a nivel internacional muy activa formada por casi 500 personas, a través de la cual se ponen común todas las iniciativas en torno a lo que denominamos “IA responsable”. Dicho grupo es liderado por grandes expertos en esta materia que colaboran con diferentes administraciones públicas en la definición de las normas y regulaciones específicas.

El propósito de PwC es generar confianza en la sociedad y resolver problemas importantes. Somos una red de firmas presente en 156 países con más de 295.000 profesionales comprometidos en ofrecer servicios de calidad en auditoría, asesoramiento fiscal y legal, consultoría y transacciones. Cuéntanos qué te preocupa y descubre cómo podemos ayudarte en www.pwc.es