

LabS Inteligencia Artificial Responsable e Inclusiva

Triple E, Ética = Explicabilidad + Equidad



Contenidos

Resumen ejecutivo

03

01 Sobre el LabS IA

04

02 Contexto actual – marco europeo

07

03 Explicabilidad y Equidad: Introducción

09

04 Explicabilidad: Comprender el proceso de toma de decisiones de los algoritmos

4.1 La necesidad de modelos explicables

4.2 La explicabilidad es un concepto importante en todos los niveles de la organización

4.3 Relación entre la complejidad del modelo y el grado de explicabilidad

11

05 Equidad: Garantizar resultados justos para todas las empresas

5.1 Conocer los tipos de sesgos en los modelos de IA

18

06 ¿Cómo aplicar la explicabilidad y la equidad en proyectos de IA?

23

Conclusiones

25

Sobre Fundación SERES & NTT DATA

27

Resumen ejecutivo

La inteligencia artificial (IA) es una tecnología que tiene el potencial de transformar muchos aspectos de nuestras vidas. Sin embargo, también existe el riesgo de que la IA se utilice de forma discriminatoria o sesgada o que aparezcan errores que la hagan funcionar incorrectamente. Para evitar este riesgo, es importante que los sistemas de IA sean comprensibles y justos y que las empresas y organizaciones adopten los principios de una IA ética en sus operaciones.

En el LabS IA de 2023, la Fundación SERES y NTT DATA hemos logrado profundizar en los principios de explicabilidad y equidad, su fuerte conexión con el plano ético, y cómo estos afectan de manera directa a las empresas.

La explicabilidad es la capacidad de interpretar cómo funciona un algoritmo. Nos permite confiar en el algoritmo, y adoptar un proceso de mejora continua. También se ha demostrado cómo el grado de explicabilidad está relacionado con la complejidad del algoritmo.

La equidad es la capacidad de un algoritmo de ser justo e imparcial, garante de la inclusión y la diversidad de los individuos y grupos sociales. Se ha ahondado en el concepto de sesgo, y demostrado cómo estos pueden producirse en todas las fases del ciclo de vida de los modelos de IA.

Marcos de referencia como el *CDO Journey*, elaborado por NTT DATA, permiten a las empresas desarrollar e implementar sistemas de IA éticos que sean beneficiosos para todas las partes interesadas, cumpliendo con las regulaciones existentes y preparados para las regulaciones futuras.

En conclusión, los principios de explicabilidad y equidad son fundamentales para el desarrollo y uso ético de la IA.

“ Las empresas deben incorporar estos principios en su cultura y procesos desde el inicio de sus proyectos de IA para poder asegurar un uso responsable de esta tecnología.

1. Sobre el Labs IA

Desde el año 2020, NTT DATA y la Fundación SERES colaboran en la creación de diferentes laboratorios temáticos sobre inteligencia artificial (IA), que tienen como objetivo ayudar a las empresas y organizaciones a afrontar los retos actuales de la IA.

La primera edición del LabS IA celebrada en 2020 buscaba **impulsar el papel de la empresa en el desarrollo de una IA responsable e inclusiva que evite daños a la sociedad**, tanto a través de la definición de objetivos estratégicos como a través de programas de formación. Para ello, en este primer laboratorio identificamos un **decálogo** con 12 enunciados sobre cómo ayudar a las empresas a impulsar el paradigma de IA ética desde diferentes perspectivas. Tomando como elemento principal de la sesión los siete principios éticos introducidos por la Unión Europea para construir una IA fiable, desarrollamos un informe final denominado Decálogo común para una IA responsable e inclusiva¹.



¹Decálogo Labs IA Responsable e Inclusiva



Ilustración 1:
Doce enunciados del Decálogo LabS IA

Después de este ejercicio, en el año 2021 celebramos la segunda edición del LabS IA que tenía por objetivo el **diseño de servicios de IA centrados en las personas**. Es decir, bajo un modelo enfocado en las necesidades de las personas, tanto clientes, como usuarios y colaboradores. Sobre este concepto, introducimos nuestra metodología de AI Service Design a través de un workshop para comprender el diseño de servicios de IA de principio a fin donde la persona cobraba un papel protagonista. Al igual que en el primer LabS, este workshop también culminó con un entregable.²



² NTT DATA (2022) LabS IA Diseño de Servicios de IA para las personas



Ilustración 2: Metodología para el diseño de un servicio de IA

De esta manera, nos basamos en la experiencia y el conocimiento adquirido en los laboratorios anteriores para crear esta nueva edición titulada LabS IA “Triple E: Ética = Explicabilidad + Equidad”. El laboratorio de esta edición 2022/2023 ahonda en los principios de **explicabilidad y equidad**. Para ello, tomamos como referencia uno de los enunciados del Decálogo identificado en la primera edición relacionado con la diversidad y el fomento de la inclusión de los individuos en las comunidades. Es decir, con el principio de equidad.

Este principio tiene por objetivo reducir el impacto negativo de la tecnología y las vulnerabilidades que la IA puede introducir, por ejemplo, a través de los sesgos. Por otro lado, recuperamos del segundo LabS los principios de explicabilidad y transparencia, es decir, aquellos principios que establecen una IA confiable. **Esta tercera edición del LabS tiene por objetivo desgranar los principios de explicabilidad y equidad, y hacer el presente estudio de las implicaciones éticas en el desarrollo, despliegue y ciclo de vida de la IA.**



La inteligencia artificial es parte de la vida cotidiana, pero su gestión responsable y regulación global son esenciales, según han explicado líderes como Brad Smith y Sam Altman durante el Foro Económico Mundial de Davos de 2024.

Garantizar la preparación de la fuerza laboral y mantener un debate ético continuo, como enfatizó Nick Clegg, son aspectos cruciales para evitar daños no intencionales y avanzar hacia un desarrollo tecnológico ético y beneficioso para la sociedad en el que las organizaciones construyan una relación de confianza con las personas basada en la transparencia y la explicabilidad de la tecnología.

“ El objetivo de LabS es desarrollar marcos de actuación que consideren la perspectiva social para abordar los problemas éticos asociados con la IA desde su concepción hasta el diseño de productos.

2. Contexto actual – marco europeo

Los principios de explicabilidad y equidad no son dos conceptos aislados: parten del impulso a una IA ética a la que está apostando la Comisión Europea a través de diferentes publicaciones y propuestas de regulación que se remontan a 2018.

Estas propuestas tienen el objetivo de proporcionar directrices a empresas y organizaciones y sentar las bases sobre los usos de la IA.

Entre ellas destaca el **Plan coordinado de Inteligencia Artificial** publicado en 2018³, un esfuerzo entre los Estados miembros y la Comisión Europea para alinearse en materias de estrategia, políticas, regulación, inversión, etc, donde la IA sea el motor de la economía europea.

Seguidamente, en el año 2019, se publicaron las **Directrices éticas para una IA confiable**⁴, que es el primer documento que define el camino hacia una IA ética a través de la fiabilidad de la tecnología. Es decir, que para que la IA sea confiable, debe ser robusta, segura y promover los derechos de las personas. En este mismo documento se detallan los cuatro principios fundamentales de la IA confiable: prevención de daños, autonomía humana, equidad y explicabilidad.

Por último, el **AI Act**⁵ (2021) de la Comisión Europea es una de las propuestas regulatorias más importantes a nivel europeo, ya que establece las bases sobre las obligaciones tanto de los proveedores de servicios como de los usuarios, atendiendo al nivel de riesgo que puede suponer el uso de la inteligencia artificial. Las categorías son:

- Riesgo inaceptable, que prohíbe el uso de IA para armas autónomas o sistemas de vigilancia masiva, por ejemplo.
- Riesgo alto, englobando sistemas de IA utilizados para el acceso a empleo, educación o servicios públicos.
- Riesgo limitado, como chatbots o categorización biométrica tienen establecidas obligaciones de transparencia.
- Riesgo mínimo, incluyendo sistemas que no tienen obligaciones ni limitaciones de ningún tipo.



³ [Comisión Europea \(2018, 7 de diciembre\) Plan coordinado de Inteligencia Artificial](#)

⁴ [Comisión Europea \(2019\) Directrices éticas para una IA confiable](#)

⁵ [Comisión Europea \(2021, 21 de abril\) Artificial Intelligence Act](#)

Además, durante 2023, el Parlamento Europeo incorporó nuevas cláusulas en el AI Act que afectan a los proveedores de servicios de IA generativa, que estarán obligados a cumplir con requisitos adicionales de transparencia (evaluación y mitigación de riesgos, requisitos de diseño, información y medioambiente y registro en la base de datos de la UE). En cualquier caso, el texto se encuentra aún en fase de negociación y se espera su aprobación formal en el primer trimestre de 2024 y su entrada en vigor en 2026.

De esta manera, a través de espacios de cocreación como el LabS IA y en especial de esta nueva edición, NTT DATA y Fundación SERES se unen con el propósito de ayudar a las empresas y organizaciones a prepararse para todas las implicaciones y ajustes necesarios para alinearse con esta nueva regulación, y que van a afectar de manera estructural a todas las empresas que desarrollen o utilicen IA.

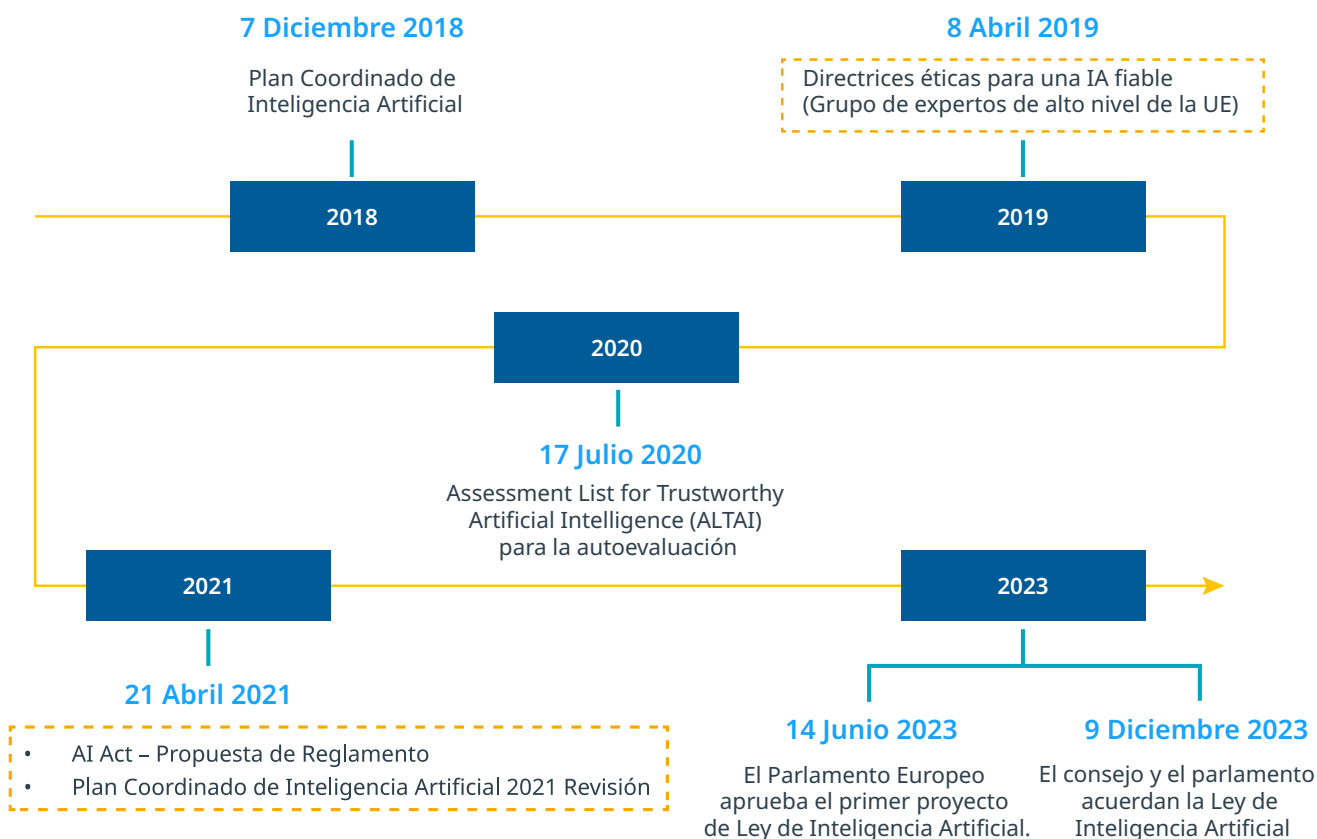


Ilustración 3:
Cronología con los hitos más importantes en materia de regulación de la IA en la Unión Europea

3. Explicabilidad y Equidad: Introducción

La explicabilidad es el principio que se ocupa de la interpretabilidad de los modelos de IA. Dentro del campo de la IA también es conocido como explainable AI (xAI) y a menudo es asociado a la transparencia. Este principio persigue entender y explicar las decisiones de los modelos de IA y conseguir que estos sean seguros, imparciales, explicables y respetuosos con la privacidad.

Por otro lado, podemos definir la **equidad** como el ámbito de investigación que se ocupa de la promoción de la igualdad y la mitigación de sesgos en modelos de IA. El objetivo de este principio es desplegar medidas para abordar los sesgos, errores e inexactitudes, mejorar la calidad de los datos y empoderar a las personas.

Gartner (2022) ⁶ analiza la IA centrada en las personas, y la define como una innovación imprescindible en el área y técnicas de IA en los próximos años. Esta IA centrada en las personas se basa en la inclusión del valor empresarial y social, la transparencia, la reducción de sesgos, la justicia, la explicabilidad, la seguridad y el cumplimiento reglamentario, entre otros. Según el mismo estudio, **la adopción generalizada de la IA responsable se alcanzará dentro de cinco y diez años, pero tendrá un efecto profundamente transformador en las empresas**, de manera que es importante integrar la ética de manera estructural en las estrategias de IA para promover su adaptación.

Para explorar el nivel de conocimiento sobre estos temas, durante el LabS IA preguntamos a los asistentes si estaban familiarizados con los conceptos de explicabilidad y equidad. El **48%** respondió que no estaba familiarizado con estos conceptos, el **20%** respondió que sí conocía estos principios pero que aún no los aplicaba. Por último, el **32%** respondió que sí estaba familiarizado con los principios de explicabilidad y equidad y los aplicaba en su organización.



⁶ [Garner \(2022\) Novedades del Hype Cycle de Gartner para la inteligencia artificial](#)



“ Aunque muchas compañías han comenzado a adoptar medidas básicas para comprender las respuestas de los modelos de IA, aprovechar al máximo los beneficios sobre la explicabilidad requiere de una estrategia integral en todos los niveles de la compañía.

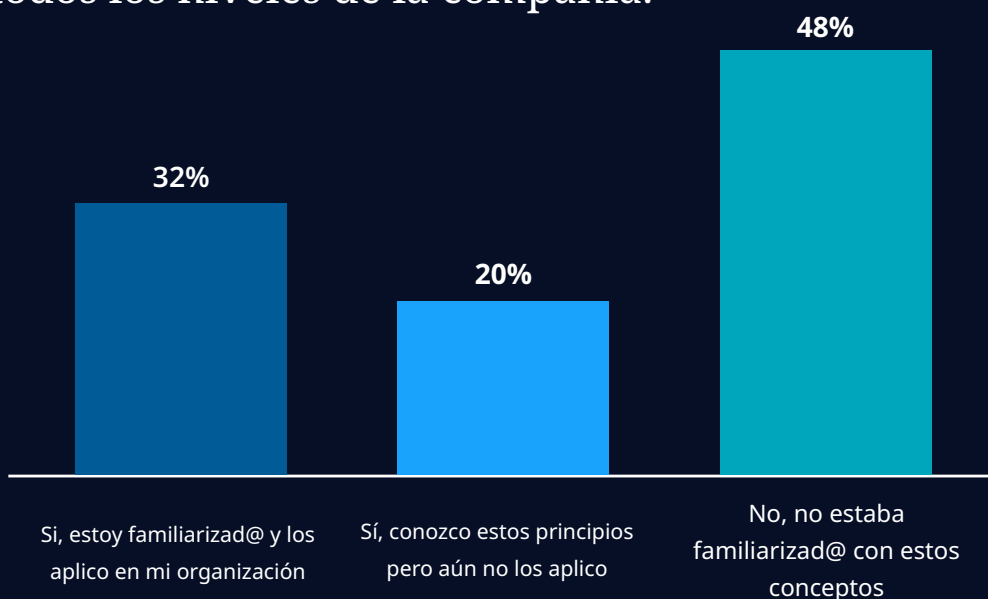


Ilustración 4:

Resultados a la pregunta lanzada en la edición del IA referente a la familiarización con los conceptos de equidad y explicabilidad

4. Explicabilidad: Comprender el proceso de toma de decisiones de los algoritmos

4.1 La necesidad de modelos explicables

Poder interpretar un modelo y explicar sus decisiones nos permite crear un proceso de mejora continua. Sin embargo, debido a la complejidad de los modelos de IA, en muchas ocasiones es imposible para un ser humano entender y comprender las decisiones de dichos modelos. Esto es conocido como el efecto de caja negra (black box en inglés). McKinsey ⁷ revela en su artículo **“Why businesses need explainable AI and how to deliver it”** que las empresas cada vez dependen más de los sistemas de IA para tomar decisiones que pueden afectar significativamente los derechos individuales, la seguridad humana y las operaciones críticas del negocio. Pero ¿cómo derivan estos modelos sus conclusiones? ¿Qué datos utilizan? ¿Podemos confiar en los resultados? Abordar estas preguntas es la esencia de la “explicabilidad” y hacerlo bien se está convirtiendo en algo esencial.

En el siguiente diagrama (en la página 12) podemos ver de qué manera puede impactar en un proyecto de IA la inclusión o no de interfaces y herramientas que favorezcan la explicabilidad y mitiguen el efecto de caja negra.

De acuerdo con el planteamiento de un modelo no explicable de IA, para la tarea que tenemos que realizar, utilizamos una serie de datos de entrenamientos que se procesan a través de *machine learning* y generamos una función. Este proceso nos acaba ofreciendo una decisión o recomendación en forma de conclusión. Sin embargo, sin las medidas apropiadas, es muy probable que no seamos capaces de explicar por qué el algoritmo ha llegado a ese resultado, qué se podría considerar una respuesta correcta, qué se podría considerar una respuesta incorrecta, cómo corregir un error y lo más importante, si se puede confiar en ese modelo.



⁷ Grennan, L., Kremer, A., Singla, A., & Zipparo, P. (2022, 29 de septiembre) [Why businesses need explainable AI—and how to deliver it.](#) McKinsey

En cambio, en un modelo de IA explicable donde incorporamos una interfaz de explicabilidad podemos ver de qué manera se establece una relación de feedback entre la tarea y los resultados de la función algorítmica, de manera que podemos seguir el recorrido mediante el cual el algoritmo nos da una respuesta y no otra, podemos ser capaces de corregir los errores de manera más precisa y, en definitiva, podemos confiar en el modelo.

Además, las organizaciones que establecen confianza digital entre los consumidores a través de prácticas como hacer que la IA sea explicable tienen más probabilidades de ver crecer su ingreso anual y EBIT a tasas del 10% o más.

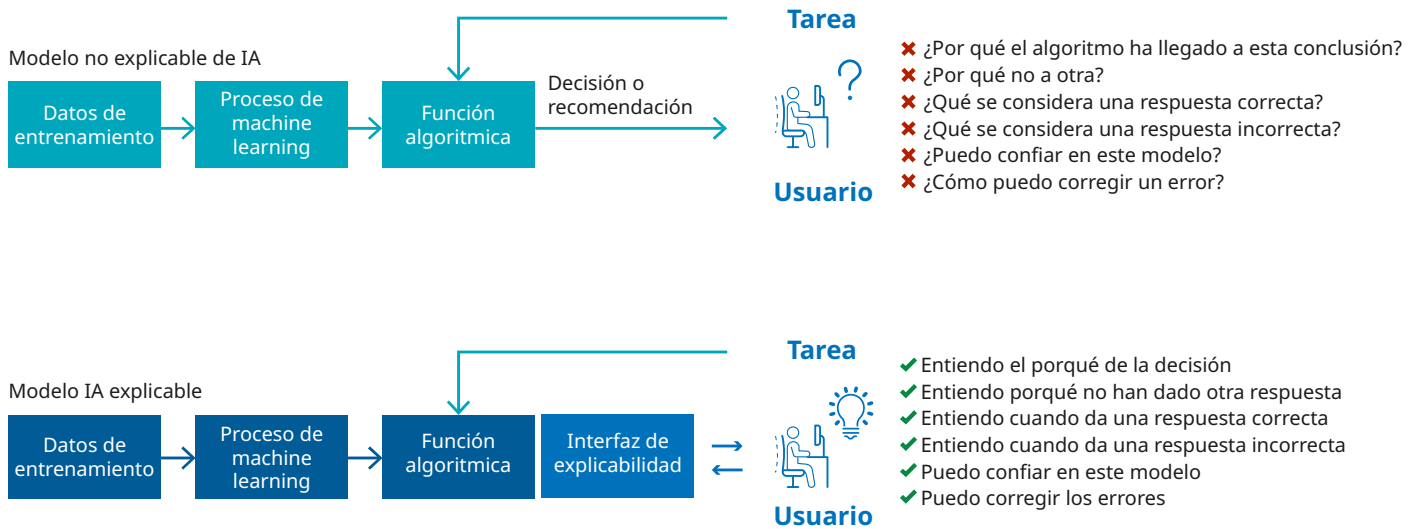


Ilustración 5:

Comparación de un modelo no explicable de IA frente a uno que sí lo es

“ Según McKinsey, las empresas que obtienen los mayores beneficios económicos debido al uso de la IA son aquellas que dedican al menos el 20% del EBIT⁸ a su uso. Estas empresas tienen más probabilidades de seguir buenas prácticas para hacer que la IA sea más explicativa.

⁸ Earnings before interest and taxes (Beneficio antes de intereses e impuestos): indicador de la rentabilidad de una empresa calculado como los ingresos menos los gastos sin incluir impuestos e intereses.

4.2 La explicabilidad es un concepto importante en todos los niveles de la organización

La IA ética es un paradigma que afecta a los proyectos de IA de manera transversal, no solamente en los procesos técnicos, todos los equipos involucrados en el ciclo de vida pueden beneficiarse de la aplicación de medidas que favorezcan la explicabilidad.

En **perfiles ejecutivos**, la explicabilidad puede ayudarnos a mejorar la propuesta de valor, facilitar el cumplimiento de la legislación (la futura regulación de AI Act exige a los proveedores de servicios de IA que establezcan normas de transparencia para los sistemas de IA destinados a interactuar con personas físicas). Asimismo, la explicabilidad puede ayudarnos a aumentar la confianza en los clientes, así como atraer nuevos clientes y, en definitiva, mejorar la imagen de la empresa.





Pero no solo afecta a niveles ejecutivos. Es importante incorporar medidas de explicabilidad en nuestros **equipos técnicos**, ya que va a favorecer la producción de nuevos modelos algorítmicos, la mejora del conocimiento de los algoritmos, la transparencia y el incremento de la performance. También vamos a poder comprender mejor el impacto de nuestros proyectos y aumentar la seguridad al crear informes y análisis, y detectar los sesgos que se puedan producir de manera más eficiente.

De cara a los **clientes**, encontramos beneficios como el aumento de la confianza y de la fidelidad. La explicabilidad nos puede ayudar a llegar a nuevos clientes, aumentar las ventas y, en definitiva, mejorar la calidad del servicio.



4.3 Relación entre la complejidad del modelo y el grado de explicabilidad

El Instituto Nacional de Estándares y Tecnología (NIST) en su publicación “**Four Principles of Explainable Artificial Intelligence**” (2020) ⁹ presenta cuatro principios fundamentales para sistemas de IA explicables. Estos principios son:

-  **Explicabilidad:** obliga a los sistemas de inteligencia artificial a proporcionar evidencia, apoyo o razonamiento para cada resultado.
-  **Significado:** los sistemas deben proporcionar explicaciones comprensibles para cada usuario individual.
-  **Precisión de la explicación:** la explicación debe reflejar correctamente el proceso del sistema para generar el resultado.
-  **Límites del conocimiento:** el sistema solo debe operar bajo las condiciones para las que fue diseñado y cuando alcanza suficiente confianza en su resultado.



Ahora bien, es importante señalar que la explicabilidad no es un factor binario, sino que es un espectro, y el grado de explicabilidad está íntimamente relacionado con la precisión del modelo de IA.

Tal y como venimos diciendo a lo largo de este informe, la explicabilidad es la capacidad de comprender y explicar cómo el modelo de IA toma decisiones o realiza predicciones a partir de los datos de entrada. Por otro lado, la precisión (a veces viene referida como accuracy) de un modelo de IA es la capacidad que tienen los modelos para predecir con exactitud. De esta manera, la complejidad del modelo se relaciona con el grado de explicabilidad y el grado de precisión del algoritmo.

Inciendo en la relación entre precisión y explicabilidad, hay una balanza entre ambos factores. Veámoslo en el siguiente gráfico en donde se pueden ver algunos de los algoritmos más comunes utilizados en sistemas de IA:

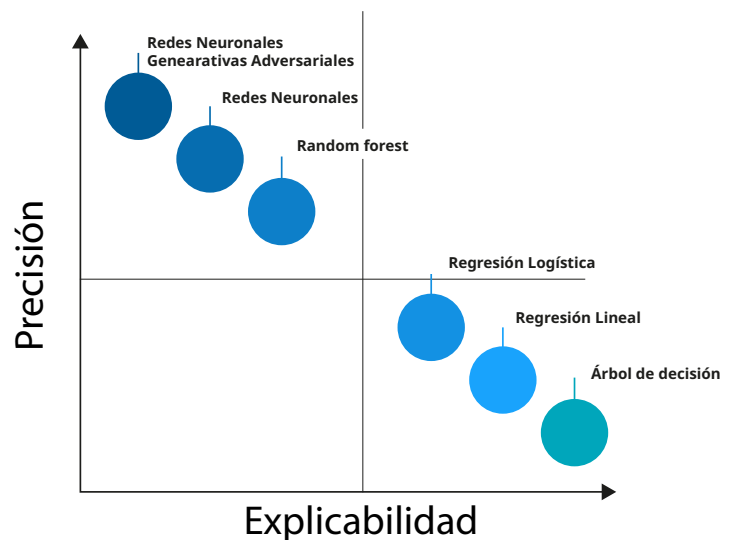


Ilustración 6:

Gráfico que presenta la precisión frente a la explicabilidad del modelo incluyendo distintos algoritmos reales de IA

⁹ [Phillips, P. J., Hahn, C. A., Fontana, P. C., Broniatowski, D. A., & Przybocki, M. A. \(2020\). Four principles of explainable artificial intelligence. Gaithersburg, Maryland, 18](#)



En el anterior gráfico podemos encontrar algunas de las principales tipologías de algoritmos de IA (redes neuronales generativas adversariales, redes neuronales, regresión lineal, etc.), repartidos en dos ejes: precisión y explicabilidad. Lo que podemos observar es que los algoritmos más explicables (como los árboles de decisión) a menudo son menos precisos que algoritmos más complejos como, por ejemplo, los sistemas de redes neuronales, que son altamente precisos, pero, sin embargo, menos explicables. En definitiva, cuanto más complejo es el sistema de IA más preciso es, pero menos explicables son sus respuestas. Aunque es importante destacar que no se puede indicar que sea mejor utilizar un tipo de algoritmo u otro, ya que para cada proyecto necesitaremos desarrollar un modelo algorítmico acorde con la tarea a resolver.

Vamos a ver cómo se relaciona esto con una serie de ejemplos:

Supongamos que tenemos un hipotético algoritmo que nos ayuda a predecir la probabilidad de que un cliente incumpla un préstamo. En este caso, el banco determina los distintos tipos de clientes y genera un modelo de IA (como el que se ve abajo) para calcular su solvencia teniendo en cuenta su edad, si estudia o no y sus ingresos.

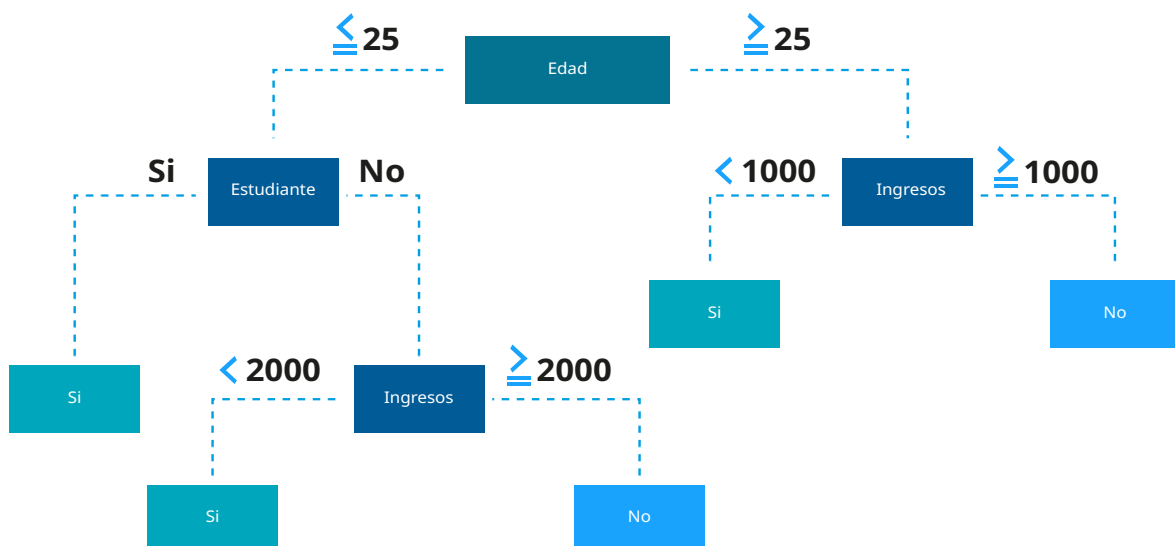


Ilustración 7: Ejemplo de funcionamiento de un árbol de decisión de forma gráfica

Vamos a ver cómo se relaciona esto con una serie de ejemplos:

Aquí, podemos determinar con facilidad si Marta (que tiene 30 años y unos ingresos de más de 3.000 euros) va a incumplir su préstamo o no. Para este ejemplo concreto, estamos hablando de un modelo de árbol de decisión en el cual el camino de toma de decisiones se puede seguir con facilidad y podemos identificar porqué nos ha dado esta respuesta.

Por otro lado, si nos vamos a otro ejemplo y nos imaginamos un hipotético sistema de diagnóstico de neumonías a través de radiografías, veremos cómo la explicabilidad cambia radicalmente. Este sistema de clasificación nos permite analizar y detectar patrones y anomalías a través de imágenes que nos puedan indicar la presencia de enfermedades pulmonares.

Analizamos las imágenes de nuestra paciente Lola, de 60 años y nos damos cuenta de que los resultados nos son precisos, generan fallos y esto supone un problema para la pronta detección. En este caso estamos ante una red neuronal. Como vemos en el gráfico, el sistema permite hacer una detección altamente precisa. Sin embargo, debido a la falta de transparencia y a lo complejo de la red, es muy difícil identificar en qué punto del ciclo de vida del sistema de IA se está produciendo el error.

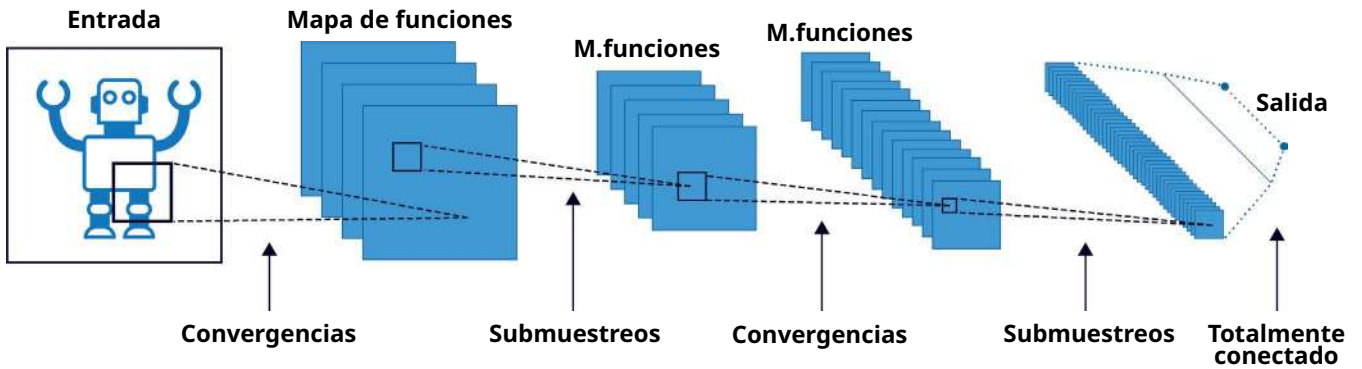


Ilustración 8:

Ejemplo gráfico de Red Neuronal Convolutiva. Se presentan sus distintas capas y la complejidad de la misma

Estos dos ejemplos ponen de manifiesto la relación entre el grado de precisión y la explicabilidad. Además, nos ayudan a comprender que los modelos de inteligencia artificial se mueven dentro del espectro de la interpretabilidad. Dentro de este espectro podemos clasificar a los modelos en:



Modelos interpretables: Un modelo interpretable puede ser entendido por un humano sin la necesidad de utilizar ninguna técnica o herramienta especial. Por ejemplo, un árbol de decisión.



Modelos explicables: Un modelo explicable es demasiado complejo para ser entendido por los humanos y requiere de técnicas adicionales para entenderlo. A este tipo de modelos los conocemos también como cajas negras (ver referencia en la página 11). En el caso mencionado, la red neuronal.

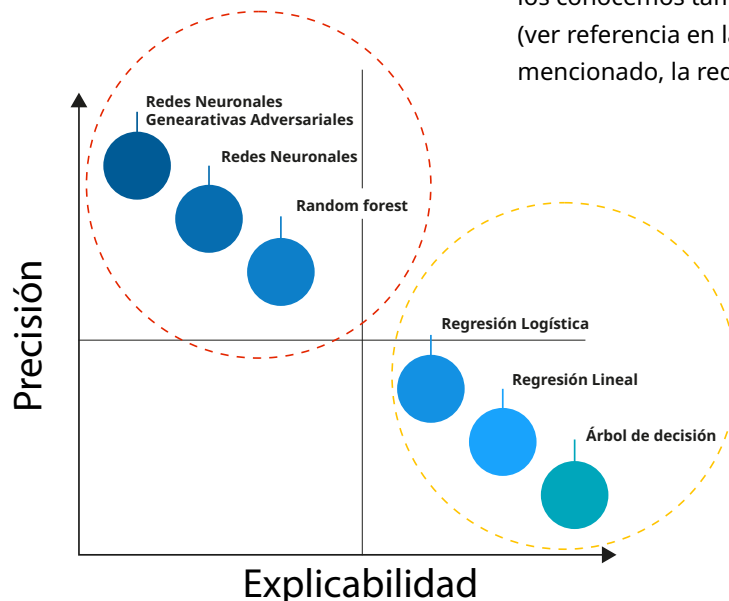


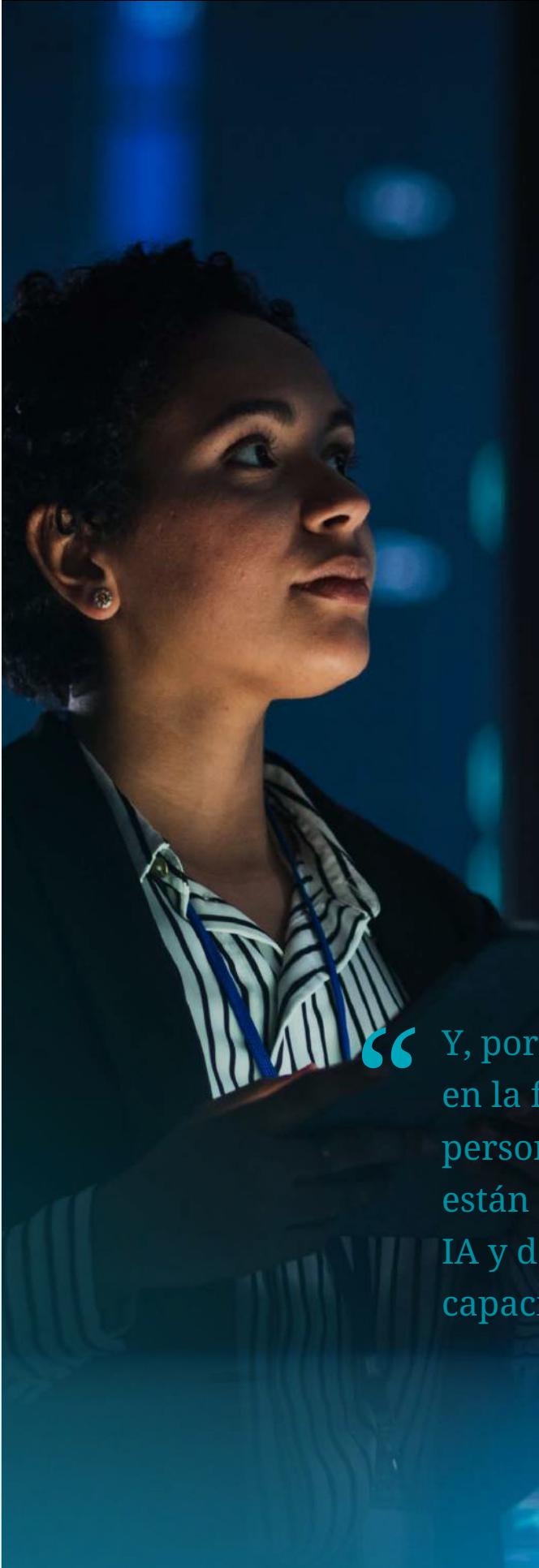
Ilustración 9:

Gráfico que presenta la precisión frente a la explicabilidad del modelo incluyendo distintos algoritmos reales de IA. En este caso, los algoritmos se agrupan en Modelos Interpretables y Modelos Explicables

Debemos destacar también que es importante crear una metodología de auditoría, así como disponer de programas de formación que permitan a nuestros equipos ser capaces de explicar las decisiones o recomendaciones del sistema.

IBM¹⁰ destaca que una IA explicable mitiga los riesgos legales, de cumplimiento, seguridad y reputación de la IA en producción. Por lo tanto, es crucial para una organización comprender plenamente los procesos de toma de decisiones de la IA con supervisión de modelos y rendición de cuentas de la IA, sin confiar ciegamente en ellos. Destacan, además, la contribución que la IA explicativa puede aportar al ayudar a las personas a comprender y explicar los algoritmos de machine learning, deep learning y las redes neuronales.

En ese sentido, los procesos de documentación son clave para posibilitar la trazabilidad, la auditabilidad y aumentar la transparencia como señala un informe de Bain & Company y el Foro Económico Mundial (2021)¹¹, destacando que la trazabilidad digital permite a las empresas cumplir con sus objetivos de sostenibilidad y lograr un conjunto más amplio de objetivos empresariales, incluyendo eficiencia, resiliencia y capacidad de respuesta.



“ Y, por último, en línea con lo establecido en la futura normativa del AI Act, las personas tienen derecho a saber cuándo están interactuando con un sistema de IA y debieran ser informadas sobre sus capacidades y limitaciones.

¹⁰ [What is explainable AI. IBM](#)

¹¹ [Bain & Company, Foro Económico Mundial \(2021, 16 de septiembre\) The traceability transformation: How transparent value chains can help companies achieve their sustainability goals](#)

5. Equidad: garantizar resultados justos para todas las personas

5.1 Conocer los tipos de sesgos en los modelos de IA

Para comprender cómo interviene el principio de equidad en los proyectos de IA, primero debemos conocer el ciclo de vida de un modelo de IA y entender qué es un sesgo.

Según la Real Academia Española (RAE), se define sesgo como un error sistemático en el que se puede incurrir cuando al hacer muestreos o ensayos se seleccionan o favorecen unas respuestas frente a otras. Los sesgos pueden influir en la forma en que percibimos y procesamos la información, y pueden afectar nuestras decisiones y juicios. Así, un algoritmo sesgado o injusto es aquel cuyas decisiones están influenciadas por determinados atributos que éticamente no deberían afectar, dando lugar a decisiones discriminatorias hacia un determinado grupo personas.

Los modelos de IA pueden encontrar sesgos a lo largo de todo su ciclo de vida puesto que estos no son agnósticos a la persona o equipo que los desarrolla ni a los datos y patrones de los que se nutren en su aprendizaje. A continuación, podemos ver el diagrama del ciclo de vida de un modelo de IA:



Ilustración 10:
Fases del ciclo de vida de un modelo de IA

Al tener una comprensión clara del proceso y las posibles fuentes de sesgo que pueden surgir en cada etapa del ciclo de vida del modelo, podemos tomar medidas para garantizar que los modelos sean justos y equitativos. Generalmente los sesgos se clasifican según la fase del ciclo en la que se generan. Estos son algunos de los ejemplos de sesgos que podemos encontrar:



1. Oportunidad de negocio:

Este paso implica la conceptualización, la planificación, y la recopilación de la información comercial. En esta fase se determinan los objetivos y requisitos del modelo de aprendizaje automático y se planifica el proceso de desarrollo. Los sesgos dentro de la definición de la propuesta de negocio que se pueden dar son:

- **Sesgo de confirmación:** Se seleccionan o se buscan aquellos datos que respalden la oportunidad de negocio o la hipótesis previa, mientras se ignoran o se descartan aquellos que la contradicen.
- **Desalineamiento de la iniciativa:** Riesgo de desalineamiento con el propósito para el que se creó el modelo. Este desalineamiento podría dar lugar a que los resultados no sean éticos.
- **Impacto en stakeholders:** No se tiene en cuenta ninguna medida para evaluar el impacto de las iniciativas en las demás partes interesadas.



2. Adquisición y descubrimiento de datos:

En este paso, se recopilan y procesan los datos que se utilizarán para entrenar el modelo. Esto puede incluir la limpieza y el procesamiento de los datos para eliminar el ruido y garantizar su calidad. Los sesgos inherentes a la adquisición de datos pueden reflejarse en:

- **Sesgo de calidad de los datos:** Utilizar bases de datos desactualizadas y poco precisas puede hacer que el algoritmo aprenda información incorrecta y, por tanto, hacer predicciones incorrectas.
- **Sesgo histórico:** Ocurre cuando el conjunto de datos utilizado para entrenar un algoritmo refleja las desigualdades históricas en la sociedad, lo que puede llevar a predicciones injustas y discriminatorias en el futuro.

- **Sesgo de muestreo:** Se produce cuando la muestra representada en los datos utilizados para entrenar un algoritmo no es representativa de la población para la que se está haciendo una predicción.
- **Multiplicidad de etiquetas dentro de la base de datos:** Problemas derivados de la gran cantidad de valores existentes en la base de datos, que afectan a la calidad y precisión y que pueden no estar controlados por el administrador de la base de datos.



3. Creación y desarrollo del modelo:

Durante esta fase se crea un modelo de entrenamiento, se transforman los datos, se perfecciona el modelo predictivo y se evalúa su rendimiento. Los datos se combinan, manipulan y formatean para identificar patrones. En ese sentido, los sesgos que se pueden observar son:

- **Sesgo de atributo:** Ocurre cuando se utilizan atributos en el modelo que son irrelevantes para el problema que se quiere resolver.
- **Sesgo del programador:** Sesgo inconsciente que introduce el programador o equipo trabajando con parámetros conocidos sin darse cuenta de a quién dejan fuera. Ocurre especialmente en equipos no diversos.
- **Sesgo de error de medición:** Ocurre cuando los datos utilizados para entrenar el modelo están sujetos a errores de medición o incertidumbre.





4. Evaluación y despliegue del modelo:

Una vez el modelo ha alcanzado el rendimiento satisfactorio, es hora de implementarlo y llevarlo a producción. Algunos de los sesgos y riesgos que podemos identificar en esta fase son:

- **Sesgo de retroalimentación:** Surge cuando el modelo recibe prejuicios o información errónea en los datos de reentrenamiento que se le proporcionan en la retroalimentación y que pueden perpetuar los sesgos del modelo.
- **Sesgo de popularidad:** El modelo de IA favorece ciertos elementos o resultados simplemente porque son populares o tienen una alta frecuencia de aparición en los datos de entrenamiento.
- **Sesgo de características:** Las características utilizadas para entrenar el modelo pueden volverse irrelevantes o insuficientes a medida que los datos evolucionan. Es decir, características que antes eran informativas pueden volverse redundantes o perder su capacidad predictiva.
- **Sesgo de deriva de los datos:** Se produce cuando los conceptos subyacentes que describen los datos cambian con el tiempo. Por ejemplo, en el análisis de sentimiento en redes sociales, las opiniones y tendencias de los usuarios pueden cambiar con el tiempo, lo que afecta la interpretación de los datos y el rendimiento del modelo.



5. Monitorización del modelo:

Finalmente, monitorizamos el desempeño del modelo en producción, lo que nos permitirá hacer ajustes y actualizaciones para mantener su rendimiento y precisión. Los sesgos y riesgos que pueden aparecer en esta fase son:

- **Problemas de interpretabilidad:** Dificultades para comprender la relación entre entradas y salidas y predecir la respuesta del modelo cuando hay cambios en las entradas.
- **Cambios en el entorno:** Malfuncionamiento del modelo frente a los objetivos previstos debido a la introducción de sesgos o factores no deseados en el proceso de entrenamiento, cambios en el entorno o en la forma en que se utiliza el modelo.
- **Deterioro del rendimiento:** Disminución en la capacidad de realizar la tarea (ya sea por la falta de actualización de datos de entrenamiento, cambios en la distribución de datos, etc.).



“ El World Economic Forum (2021)¹² discute en su artículo “Research shows AI is often biased. Here’s how to make algorithms work for all of us” cómo los distintos tipos de sesgos son transferidos con bastante frecuencia a los modelos de IA.

En el mismo, se hace referencia a la necesidad de una mayor transparencia y responsabilidad frente a esta situación y propone distintas estrategias para identificar y mitigar los riesgos de equidad y no-discriminación.

Pero ¿cómo puede afectar la aparición de sesgos en un desarrollo de IA? Uno de los casos reales más sonados de sesgos en IA es quizás el **Caso COMPAS**, un algoritmo de predicción del riesgo de reincidencia en acusados utilizado en EE.UU. El sistema estaba diseñado para captar los detalles del perfil de un acusado (la edad, el género, el nivel de educación, el nivel de ingresos, lugar de residencia, entorno social, delincuencia familiar, entre otros) y estimar las posibilidades de reincidir.

En los resultados arrojados por el sistema, vemos como las personas afroamericanas obtienen mayores condenas y sentencias más duras (**riesgo 10 frente a 3**) en comparación con las personas blancas con antecedentes similares, es decir, tienen el doble de posibilidades que las personas blancas de ser clasificadas por error como “alto riesgo”. Los algoritmos de aprendizaje automático utilizan estadísticas para encontrar patrones en los datos, en este caso, **el sistema ha sido entrenado con datos históricos que están sesgados**, ya que la población afroamericana ha sido perseguida por las fuerzas de seguridad de manera desproporcionada a lo largo de los años, especialmente aquellas comunidades de ingresos bajos. El resultado final es que el algoritmo marca falsamente a los acusados afroamericanos como futuros delincuentes, etiquetándolos erróneamente en casi el doble de la tasa de acusados blancos. Y por tanto tienen el riesgo de recibir puntuaciones más altas de reincidencia. Como resultado, el algoritmo puede amplificar y perpetuar los sesgos ya existentes y generar aún más datos sesgados que alimenten un ciclo vicioso.



Ilustración 11:

Imagen del Caso COMPAS donde se refleja el grado de reincidencia estimado por el sistema .

Fuente: ProPublica en Psychology Today

En conclusión, podemos decir que para mitigar los sesgos en nuestras organizaciones y poder orientar nuestros proyectos hacia una IA ética y equitativa, es importante garantizar que los datos utilizados tengan en cuenta la multitud de expresiones culturales y sociales para garantizar la inclusión y reflejar la diversidad de los individuos y grupos sociales. También es de vital importancia que los resultados que den los modelos estén enfocados a preservar y proteger las características culturales de los individuos. Debemos establecer medidas y procedimientos para identificar y mitigar la aparición de sesgos en el modelo a lo largo del ciclo de vida. Y tal y como señalábamos en el capítulo sobre explicabilidad, debemos formar a los equipos en materia de diversidad e inclusión.

¹² [World Economic Forum \(2021, 19 de julio\) Research shows AI is often biased. Here’s how to make algorithms work for all of us. Artificial Intelligence](#)



Por último, la inclusión de las diferentes partes interesadas es fundamental para prevenir y mitigar la inclusión de sesgos. Por un lado, la inclusión de diferentes puntos de vista permite una evaluación más completa y holística de los posibles sesgos que puedan surgir. Además, involucrar agentes de diversas culturas y contextos puede ayudar a garantizar que el modelo sea sensible y equitativo para todas las comunidades representadas en los datos.

La inclusión de las distintas partes interesadas con conocimiento en ética, leyes y regulaciones, y aspectos sociales y políticos, puede ayudar a garantizar que se tomen decisiones equitativas y se aborden las preocupaciones éticas. Incluso la participación de los usuarios finales puede jugar un papel crucial en la adopción y aceptación de los sistemas de inteligencia artificial. Si los usuarios perciben que el modelo es sesgado o injusto, es menos probable que confíen en él y lo utilicen adecuadamente. Involucrar a los usuarios en el proceso de desarrollo ayuda a comprender mejor sus necesidades y preocupaciones, lo que puede llevar a una mayor confianza y adopción.

6. ¿Cómo aplicar la explicabilidad y la equidad en proyectos de IA?

Desde NTT DATA nos preocupamos porque estos principios de equidad y explicabilidad no se queden en un plano teórico y en una declaración de buenas intenciones, sino que acompañamos a las empresas a que puedan llevar la aplicación de estos principios al día a día de las operaciones.

Para ello NTT DATA ha creado el CDO Journey, un framework que permite a los clientes tener una visión global de la práctica de Data & Intelligence (D&I) y acortar la curva de aprendizaje al identificar áreas clave para mejora. En ese sentido, los paradigmas de explicabilidad y equidad, junto con el resto de los requerimientos éticos, son la brújula que guía las acciones de cada una de las fases de negocio: Business Value, Responsible Governance, Core Tech & Next gen Operations, Ecosystem & Innovation and Culture & Change Management.

En el área de Business Value sentamos las bases para el uso responsable de la IA en las organizaciones, promoviendo una cultura de D&I responsable que transmita los principios éticos en toda la organización. Para ello guiamos a las organizaciones con una comprensión profunda de la Ética de los Datos e IA mediante el desarrollo de una **Guía de Ética de Datos e IA**. Así mismo, proporcionamos herramientas de auditoría como el **AI Audit Model** que ayuda a los clientes a entender el grado de madurez de cumplimiento de sus modelos de IA con la nueva normativa regulatoria europea del AI Act.



También es muy importante cómo se cubre el ciclo de vida de los modelos, con nuestra propuesta de Responsible Governance ayudamos a las empresas a comprender los procesos que intervienen en cada fase del algoritmo de IA y establecer normas que guíen el gobierno responsable y ético de las soluciones de D&I. Proporcionando un **modelo estandarizado de supervisión y control de sesgos y riesgos** que conlleva su puesta en producción y garantiza la escalabilidad de las soluciones de IA.

Además, desde NTT DATA apoyamos a las organizaciones en el diseño de un plan de formación que ayude a elevar el conocimiento analítico y responsable de la IA creando caminos personalizados atendiendo a cada rol y perfil. Como consecuencia, en la parte de Culture & Change Management proporcionamos programas de formación en ética de la IA a través de talleres para toda la organización, que son sesiones de concienciación sobre IA ética, y a través de talleres específicos para perfiles técnicos, que pueden ser en materia de explicabilidad en modelos de machine learning o sobre equidad de datos y algoritmos.



Ilustración 12:
Esquema del CDO Journey desarrollado por NTT DATA

“ Como resultado, los empleados avanzados con habilidades formadas en D&I contribuirán a fomentar e impregnar una cultura de D&I en todas las áreas, desarrollando la madurez, la concienciación y las habilidades analíticas en D&I de la organización.

Conclusión

Durante la sesión del LabS IA nos hemos adentrado en la IA ética, comprendiendo principios de explicabilidad y equidad y cómo afectan a los proyectos de IA.

En términos de explicabilidad resaltamos que entender cómo funciona un algoritmo nos permite confiar más en él, aumentar la calidad del servicio, mejorar la propuesta de valor y dar mejor respuesta a los requisitos regulatorios. A través de ejemplos prácticos que se realizaron en la sesión y que explicamos en el presente informe, demostramos que el grado de explicabilidad está profundamente relacionado con la complejidad del algoritmo.

En el bloque de equidad, evidenciamos la importancia de emplear algoritmos que tengan en cuenta la multitud de expresiones culturales y sociales para garantizar la inclusión y reflejar la diversidad de los individuos y grupos sociales. También analizamos cómo los sesgos se pueden producir en todas las fases del ciclo de vida de los modelos de la IA.

“ Sin embargo, el cambio cultural y los procesos de concienciación son fundamentales para priorizar el uso ético de la IA en toda la organización.

Es por esto, que los principios de explicabilidad y equidad deben estar presentes desde las fases iniciales de definición de un proyecto de IA, para lo cual, deben formar parte de la cultura de las empresas. Marcos de referencia como el CDO Journey permiten a las empresas iniciar un camino de transformación o identificar procesos de innovación en los que la aplicación de la IA ética es beneficiosa en todas las áreas de negocio además de que ayuda a preparar a las empresas para los futuros requisitos regulatorios.

Empresas participantes en el LabS IA Responsable e Inclusiva desde su lanzamiento:

BBVA

 **CaixaBank**


CUATRECASAS

El Corte Inglés

ferrovial

FUJITSU

FFP FUNDACIÓN FERNANDO POMBO

gsk

 **IESE**
Business School
University of Navarra

 **ILUNION**

MELIÀ
HOTELS & RESORTS

 **NTT DATA**

Pérez-Llorca

 **randstad**

randstad
fundación.

^BSabadell

 **Santander**

TENDAM
GLOBAL FASHION RETAIL

URÍA
MENÉNDEZ

 **vodafone**


Willis Towers Watson

Sobre Fundación SERES & NTT DATA

Fundación SERES, entidad sin ánimo de lucro, nacida hace 15 años, acompaña la transformación de las empresas e impulsa su liderazgo ante los retos sociales. Su objetivo es posicionar el valor de lo social en las organizaciones. Como movimiento pionero, con cerca de 150 compañías adheridas, que representan el 30% del PIB y el 75% del IBEX 35, aborda el compromiso social de las empresas desde un enfoque estratégico y práctico basado en la innovación.

Desde la Fundación, trabajamos junto con las compañías para abordar los principales retos corporativos en materia social, aunando propósito y estrategia. En el campo de la Inteligencia Artificial, promovemos que las organizaciones hagan una gestión responsable de la tecnología y contribuyan a un modelo de progreso inclusivo, que no deje a nadie atrás.

NTT DATA, compañía japonesa en el TOP 10 de empresas de servicios TI más grande del mundo, cuenta con más de 140.000 profesionales y opera en más de 50 países. En NTT DATA acompañamos a nuestros clientes en su desarrollo digital a través de una amplia oferta de servicios de consultoría estratégica y Advisoring, tecnologías de vanguardia, aplicaciones, infraestructura, modernización de servicios TI y BPOs. Aportamos una profunda experiencia en todos los sectores de actividad económica y un gran conocimiento de las geografías donde tenemos presencia.

Ponemos nuestro empeño en la construcción de una comunidad de personas única y abierta, liderada por unos valores compartidos, que ha ido creciendo como una gran red de talento colectivo capaz de multiplicar nuestras capacidades y nuestro conocimiento, para responder con agilidad a las necesidades cambiantes de nuestros clientes y anticiparnos con inteligencia al futuro. En el ámbito de Data & Intelligence, aceleramos la transformación de negocio de nuestros clientes a través de la innovación y de un porfolio completo de servicios.

Contacto



David Pereira Paz

Head of Data & Intelligence Europe,
NTT DATA



Cristina Aliaga Ibañez

Directora del Área de Empresas,
Fundación SERES



Beatriz Zamora
Project Manager,
Fundación SERES

Contacto



Alicia de Manuel Lozano
Expert Analyst in AI Ethics



Alberto Martinez Caballero
Tech Advisory Consultant

