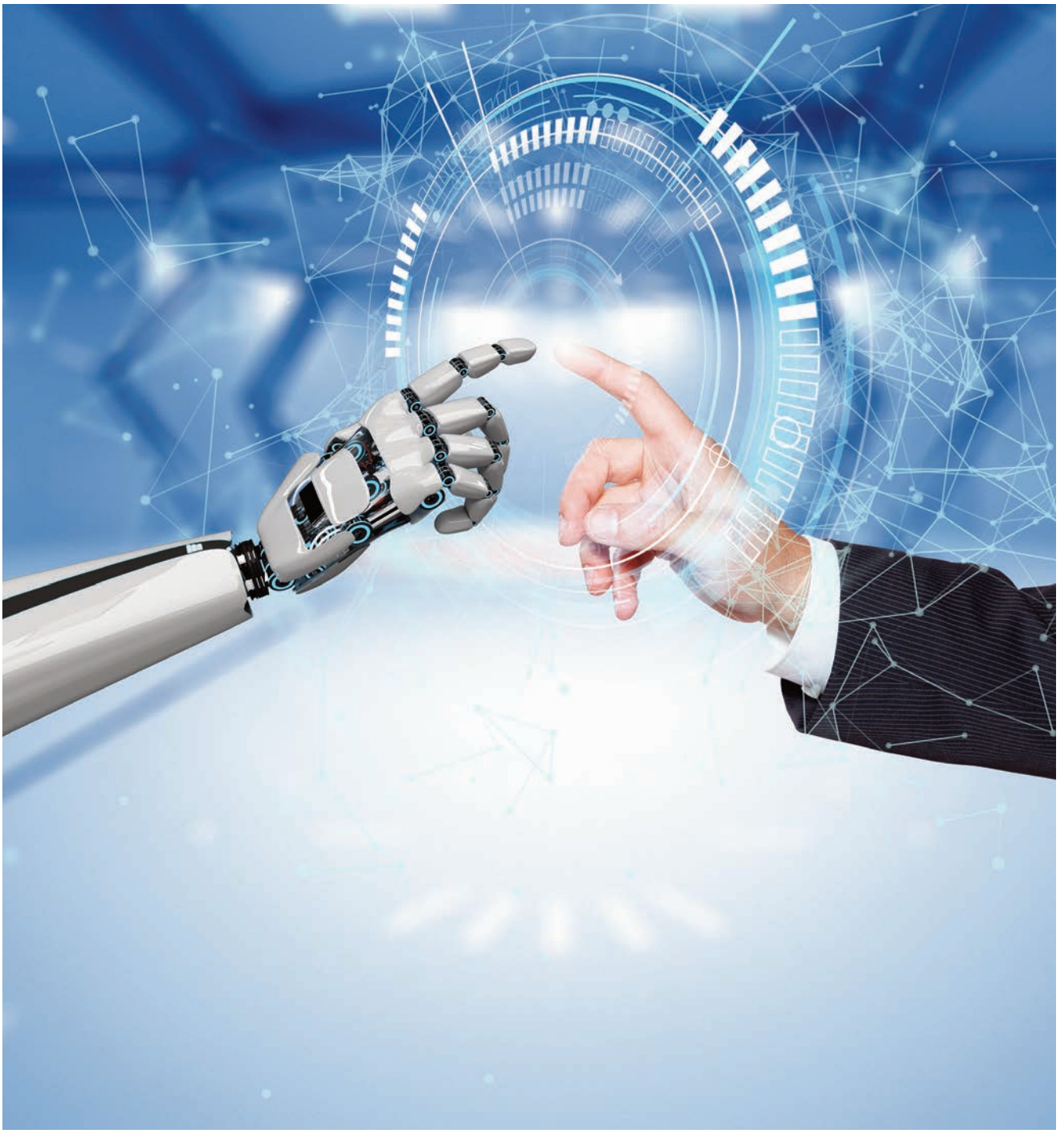
Contexto y fundamentos de la XAI

"Comprender la inteligencia artificial es un desafío que requiere una enorme capacidad intelectual; afortunadamente, contamos con la inteligencia artificial para abordarlo".

GPT-4[16]



Contexto

Una de las características más notables de la transformación digital es que está poniendo a disposición de todas las industrias una cantidad masiva de datos estructurados y no estructurados provenientes de múltiples aplicaciones; por ejemplo:

- Datos de comercio minorista procedentes de acciones de compra, transacciones y comentarios de los clientes.
- Datos financieros procedentes de fuentes bancarias, de inversión y comerciales.
- Datos de redes sociales, incluidos análisis de opiniones y análisis predictivos.
- Sensores digitales IoT (Internet de las Cosas) que miden la temperatura, la presión y otros datos del entorno.
- Datos sanitarios, como historiales médicos, diagnósticos, imágenes e información genómica.
- *Wearables*, como rastreadores de actividad, sensores de salud y relojes inteligentes.
- Sistemas de reconocimiento de voz que permiten a las máquinas entender y responder al lenguaje natural.
- Satélites y otros sensores espaciales que proporcionan información sobre el tiempo y el clima.
- Sistemas de vigilancia inteligentes que utilizan el reconocimiento facial y la detección de objetos.
- Sensores de vehículos autónomos como cámaras, lidar, radar y sensores ultrasónicos.

La disponibilidad de estos datos, junto con la presencia de enormes capacidades de almacenamiento y procesamiento computacional a coste reducido, ha impulsado un mayor apetito por la modelización avanzada, que se manifiesta en el uso de una amplia gama de técnicas de aprendizaje automático y en el desarrollo de la inteligencia artificial (AI) en prácticamente todos los sectores y ámbitos[17].

Aunque hay consenso sobre el hecho de que los modelos de AI proporcionan en general un mayor poder predictivo que los modelos tradicionales[18], también introducen una mayor complejidad y puede resultar difícil interpretarlos y explicar sus resultados.

Esto genera riesgos vinculados con el uso de estos modelos, como la falta de comprensión del modelo, la presencia de sesgos inadvertidos o la dificultad para determinar si el modelo está sobreentrenado (global o localmente), lo que puede dar lugar a una escasa capacidad de generalización y a potenciales errores en las decisiones basadas en él, y como consecuencia, derivar en una falta de confianza en el modelo.

Todo ello lleva a la pregunta de si es posible comprender lo suficientemente bien los resultados que arrojan los algoritmos de AI, especialmente cuando tienen impacto en decisiones críticas, como el diagnóstico médico, la conducción autónoma o la detección del fraude, entre otras muchas.

[16] GPT-4, Generative Pre-Trained Transformer, red neuronal profunda diseñada por la Fundación OpenAI para realizar tareas de procesamiento del lenguaje natural (NLP). En este caso, se le pidió "Inventa 10 citas ingeniosas sobre la inteligencia artificial y cómo de difícil y necesario es ser capaz de interpretar y explicar los modelos de AI". La cita presentada fue la tercera.

[17] Aunque hay diferencias, dada la falta de consenso sobre su definición, en este documento se emplearán de forma indistinta los términos "aprendizaje automático", "machine learning (ML)", "inteligencia artificial (AI)" y "modelización avanzada". Asimismo, se utilizará la abreviatura "AI" para "inteligencia artificial", por consistencia con las siglas "XAI" (que habitualmente no se traducen), incluso en las citas de publicaciones en español.

[18] LeCun, Y. et al (2015). Investigador en Facebook AI Research y la Universidad de Nueva York.

Figura 2. Número de publicaciones científicas por año sobre Explainable Artificial Intelligence (XAI).

Definición

La disciplina de XAI es relativamente nueva y, por tanto, no hay todavía una doctrina asentada que estandarice su terminología. Pese a algunos esfuerzos notables para definir los términos[19], la aproximación a la XAI es o bien heterogénea (según la fuente académica consultada), o bien intuitiva (más frecuente en la práctica industrial).

En todo caso, para la mayor parte de usos en la práctica puede ser suficiente definir XAI del siguiente modo[20]:

> La inteligencia artificial explicable (XAI) es el conjunto de procesos y métodos que permiten a los usuarios humanos comprender y confiar en los resultados y productos creados por algoritmos de aprendizaje automático. La XAI se utiliza para describir un modelo de AI, su impacto previsto y sus posibles sesgos. Ayuda a caracterizar la precisión del modelo, la imparcialidad, la transparencia y los resultados en la toma de decisiones basada en AI. La XAI es crucial para que una organización genere confianza a la hora de poner en producción modelos de AI. La explicabilidad de la AI también ayuda a una organización a adoptar un enfoque responsable del desarrollo de la AI.

Relevancia de la XAI

Un aspecto en el que hay consenso entre académicos y profesionales de la industria es en la relevancia creciente de la XAI como disciplina complementaria a la AI.

Las herramientas de análisis de publicaciones científicas identifican más de 77.000 artículos sobre XAI entre 2014 y 2022, y en tendencia exponencialmente creciente, con más de 20.000 artículos solo en 2022 (Fig. 2)[21].

Más allá del interés académico, la atención que recibe la XAI se explica por su capacidad para dar solución a una serie de inquietudes de la industria en el uso de la AI (Fig. 3); entre ellas:

- **Requerimientos regulatorios:** la obligación de cumplir con la regulación emergente sobre el uso de AI.
- **Falta de confianza:** la necesidad de generar confianza sobre el modelo de AI y los resultados que arroja en los usuarios, los validadores y auditores, y en última instancia el público en general.
- **Potencial mal uso:** la conveniencia de evitar el mal uso de los modelos debido a la falta de comprensión sobre su funcionamiento, lo que puede conllevar costes e incluso sanciones.
- **Impacto reputacional:** la prevención de impactos reputacionales sobre la compañía debidos a sesgos, decisiones discriminatorias, uso inapropiado o simplemente predicciones erróneas del modelo.
- **Impactos sociales o humanos:** la prevención de impactos sociales o humanos en usos críticos como la AI para el diagnóstico de enfermedades médicas, sentencias judiciales, identificación biométrica, polígrafos, etc.
- **Otros:** la mitigación de otros riesgos que emanan de la falta de comprensión sobre el modelo, como ciberseguridad, protección de datos, fraude, riesgo de modelo, etc.

Pese a todo lo anterior, hay casos en los que los modelos de AI no necesitan ser particularmente interpretables, porque los usos no están regulados, porque no tienen impactos potenciales relevantes o simplemente porque no es necesario interpretarlos, como los sistemas de recomendación automática de cine y música, o los algoritmos que juegan al ajedrez, por ejemplo.

[19] Marcinkevics et al. (2020). Departamento de Computer Science, ETH Zürich.
[20] IBM (2022).
[21] Dimensions (2022).

Figura 3. Inquietudes de la industria que la XAI contribuye a solucionar.

Regulación

La XAI, por tanto, se está posicionando como una disciplina de relevancia creciente; y esto está llevando a reguladores y supervisores de distintas jurisdicciones a establecer reglamentos y directrices para el uso apropiado de la AI, incluyendo los aspectos de interpretabilidad de los modelos.

En este contexto, posiblemente las referencias regulatorias más relevantes a la fecha de redacción de este documento son las siguientes:

1. GDPR (Parlamento Europeo)

En Europa, el Reglamento General de Protección de Datos, que entró en vigor en 2018, establece el "derecho a una explicación" de los ciudadanos, según el cual[22]:

> El interesado debe tener **derecho a no ser objeto de una decisión**, que puede incluir una medida, que evalúe aspectos personales relativos a él, y **que se base únicamente en el tratamiento automatizado** y produzca efectos jurídicos en él o le afecte significativamente de modo similar, como la denegación automática de una solicitud de crédito en línea o los servicios de contratación en red en los que no medie intervención humana alguna. [...]
>
> En cualquier caso, dicho tratamiento debe estar sujeto a las garantías apropiadas, entre las que se deben incluir la información específica al interesado y el derecho a obtener intervención humana, a expresar su punto de vista, a **recibir una explicación de la decisión tomada después de tal evaluación y a impugnar la decisión.**

Esto tiene implicaciones críticas en el uso de la AI, y puede llevar a cuestionar su viabilidad. No obstante, en palabras del Parlamento Europeo[23]:

> Ciertamente existe una tensión entre los principios tradicionales de protección de datos –limitación de la finalidad, minimización de los datos, tratamiento especial de los "datos sensibles", limitación de las decisiones automatizadas– y el pleno despliegue del poder de la AI y *big data*. Estos últimos implican la recopilación de cantidades ingentes de datos relativos a las personas y sus relaciones sociales y su tratamiento para fines que no estaban totalmente determinados en el momento de la recopilación. Sin embargo, hay formas de interpretar, aplicar y desarrollar los principios de protección de datos que son coherentes con los usos beneficiosos de la AI y de *big data*.

Y esto está en línea con el cuarto principio para el uso ético de la AI establecido por el Grupo de Alto Nivel sobre inteligencia artificial de la Comisión Europea[24]:

> Explicabilidad: los procesos algorítmicos deben ser transparentes, las capacidades y objetivos de los sistemas de AI deben comunicarse abiertamente, y las decisiones deben poder explicarse a los afectados directa e indirectamente.

En todo caso, GDPR tiene impactos relevantes en el uso de la AI, en el sentido de que las compañías están legalmente obligadas a poder explicar por qué un modelo de AI ha arrojado un determinado resultado, y esto tiene implicaciones críticas en el diseño y el análisis de interpretabilidad de los modelos de AI[25].

2. Artificial intelligence act (Parlamento Europeo)

El borrador de Reglamento de inteligencia artificial o *artificial intelligence act* (AI Act), publicada en 2021, es una propuesta para el uso de la inteligencia artificial en la Unión Europea que pretende garantizar un alto nivel de confianza en la AI y sus aplicaciones, al tiempo que sienta las bases para la innovación.

[22] GDPR (2018), Cons. 71.
[23] *European Parliamentary Research Service* (2020).
[24] Ibíd.
[25] En algunos países europeos se está analizando el nivel de cumplimiento de este tipo de IA (en particular, de los denominados *Large Language Models*) con la regulación de protección de datos, y en ciertos casos se ha prohibido el uso de algunos de estos modelos de forma provisional.

El Reglamento establece un marco regulador para los sistemas de AI en la UE, e incluye requisitos de desarrollo ético, transparencia, seguridad y precisión. También establece un sistema de gobernanza y supervisión de los sistemas de AI, así como normas de protección y gobernanza de datos.

Al tratarse de un Reglamento, cuando sea aprobado será de aplicación directa en los 27 países de la Unión[26], sin necesidad de ser traspuesto al ordenamiento jurídico de cada país.

Una de sus características fundamentales es que clasifica las aplicaciones de AI en niveles de riesgo[27]:

- **Prácticas prohibidas,** que denotan la categoría de mayor riesgo; estos sistemas están totalmente prohibidos. Entre ellos se incluyen:
  - Sistemas biométricos en tiempo real que pueden utilizarse para cualquier tipo de vigilancia, aunque se aplican excepciones para la prevención de delitos y las investigaciones criminales en contextos policiales y de seguridad nacional.
  - Algoritmos de puntuación social que pueden utilizarse para evaluar a los individuos basándose en características personales o en su comportamiento de una manera que pueda causar daño o conducir a un trato desfavorable a un individuo.
  - Sistemas manipuladores que explotan las vulnerabilidades de determinados individuos específicos para distorsionar su comportamiento de manera que pueda causar daño físico o psicológico.

- **Sistemas de AI de alto riesgo,** enumerados en el Anexo III y que probablemente constituyan la mayoría de los sistemas de AI. Entre ellos se incluyen:
  - Identificación biométrica y categorización de personas físicas […].
  - Gestión y funcionamiento de infraestructuras críticas […] [p. ej. tráfico].
  - Educación y formación profesional […].
  - Empleo y gestión de trabajadores […].
  - Acceso a servicios esenciales […], incluida la evaluación de la solvencia, la calificación crediticia o el establecimiento del orden de prioridad de acceso a dichos servicios. (Nota: este aspecto se aplica en particular a los sistemas de AI utilizados en el sector de los servicios financieros).
  - Fuerzas de seguridad […].
  - Gestión de controles fronterizos […].
  - Administración de justicia y procesos democráticos […].

- **Sistemas de AI de bajo riesgo** (o riesgo limitado), que incluyen sistemas que no utilizan datos personales ni hacen predicciones que puedan afectar directa o indirectamente a ninguna persona, como las aplicaciones industriales de mantenimiento predictivo.

En lo relativo a la interpretabilidad de los modelos de AI clasificados como de alto riesgo, el AI Act establece[28] en sus Artículos 13 y 14:

Art. 13. Transparencia y comunicación de información a usuarios

1. Los sistemas de AI de alto riesgo se diseñarán y desarrollarán de un modo que garantice que funcionan con un **nivel de transparencia suficiente para que los usuarios interpreten y usen correctamente su información de salida.** […]

2. Los sistemas de AI de alto riesgo irán acompañados de las instrucciones de uso correspondientes en un formato digital o de otro tipo adecuado, las cuales incluirán información concisa, completa, correcta y clara que sea pertinente, accesible y comprensible para los usuarios. […]

Art. 14. Vigilancia humana

1. Los sistemas de AI de alto riesgo se diseñarán y desarrollarán de modo que puedan ser vigilados de manera efectiva por personas físicas durante el período que estén en uso, lo que incluye dotarlos de una herramienta de interfaz humano-máquina adecuada, entre otras cosas.

[…]

4. Las medidas mencionadas […] permitirán que las personas a quienes se encomiende la vigilancia humana puedan, en función de las circunstancias:

   a. **Entender por completo las capacidades y limitaciones del sistema de AI de alto riesgo** y controlar debidamente su funcionamiento, de modo que puedan detectar indicios de anomalías, problemas de funcionamiento y comportamientos inesperados y ponerles solución lo antes posible;
   b. ser conscientes de la posible tendencia a confiar automáticamente o en exceso en la información de salida generada por un sistema de AI de alto riesgo («sesgo de automatización») […];
   c. interpretar correctamente la información de salida del sistema de AI de alto riesgo […];
   d. decidir, en cualquier situación concreta, no utilizar el sistema de AI de alto riesgo o desestimar, invalidar o revertir la información de salida que este genere;
   e. intervenir en el funcionamiento del sistema de AI de alto riesgo o interrumpir el sistema […].

Como se puede observar, el AI Act impone condiciones restrictivas sobre la interpretabilidad de los modelos de AI de alto riesgo (Fig. 4), que en breve serán de obligado cumplimiento en toda la Unión. Es previsible que esto

[26] Se prevé que entre en vigor a los 20 días desde su publicación en el Diario Oficial de la Unión Europea, y que sea de plena aplicación a los 24 meses desde su entrada en vigor.
[27] Floridi et al. (2022).
[28] Comisión Europea (2021).

desencadene una cantidad significativa de iniciativas de adaptación al Reglamento, incluyendo una documentación más exhaustiva de los modelos y de sus usos, la aplicación de técnicas de interpretabilidad, el desarrollo de cuadros de mando de seguimiento y alertas sobre los modelos, o la revisión del procedimiento integral de desarrollo, validación, implementación y uso de los modelos, entre otros.

3. Directrices éticas para una inteligencia artificial fiable (Comisión Europea)

En abril de 2019, el grupo de expertos de alto nivel sobre AI de la Comisión Europea presentó las Directrices Éticas para una AI fiable[29], tras un proceso de consulta con más de 500 respuestas de la industria.

Las Directrices proponen siete requisitos clave que deben cumplir los sistemas de AI para ser considerados fiables, que en resumen son: (i) acción y supervisión humanas, (ii) solidez técnica y seguridad, (iii) gestión de la privacidad y de los datos, (iv) transparencia, (v) diversidad, no discriminación y equidad, (vi) bienestar ambiental y social, y (vii) rendición de cuentas.

En concreto, en lo relativo a la interpretabilidad de los modelos de AI, las Directrices establecen lo siguiente dentro de su requisito de transparencia:

> 53. La explicabilidad es crucial para conseguir que los usuarios confíen en los sistemas de AI y para mantener dicha confianza. Esto significa que los procesos han de ser transparentes, que es preciso comunicar abiertamente las capacidades y la finalidad de los sistemas de AI y que las decisiones deben poder explicarse —en la medida de lo posible— a las partes que se vean afectadas por ellas de manera directa o indirecta. Sin esta información, no es posible impugnar adecuadamente una decisión.
>
> No siempre resulta posible explicar por qué un modelo ha generado un resultado o una decisión particular (ni qué combinación de factores contribuyeron a ello). Esos casos, que se denominan algoritmos de «caja negra», requieren especial atención.
>
> En tales circunstancias, puede ser necesario adoptar otras medidas relacionadas con la explicabilidad (por ejemplo, la trazabilidad, la auditabilidad y la comunicación transparente sobre las prestaciones del sistema), siempre y cuando el sistema en su conjunto respete los derechos fundamentales.
>
> El grado de necesidad de explicabilidad depende en gran medida del contexto y la gravedad de las consecuencias derivadas de un resultado erróneo o inadecuado.

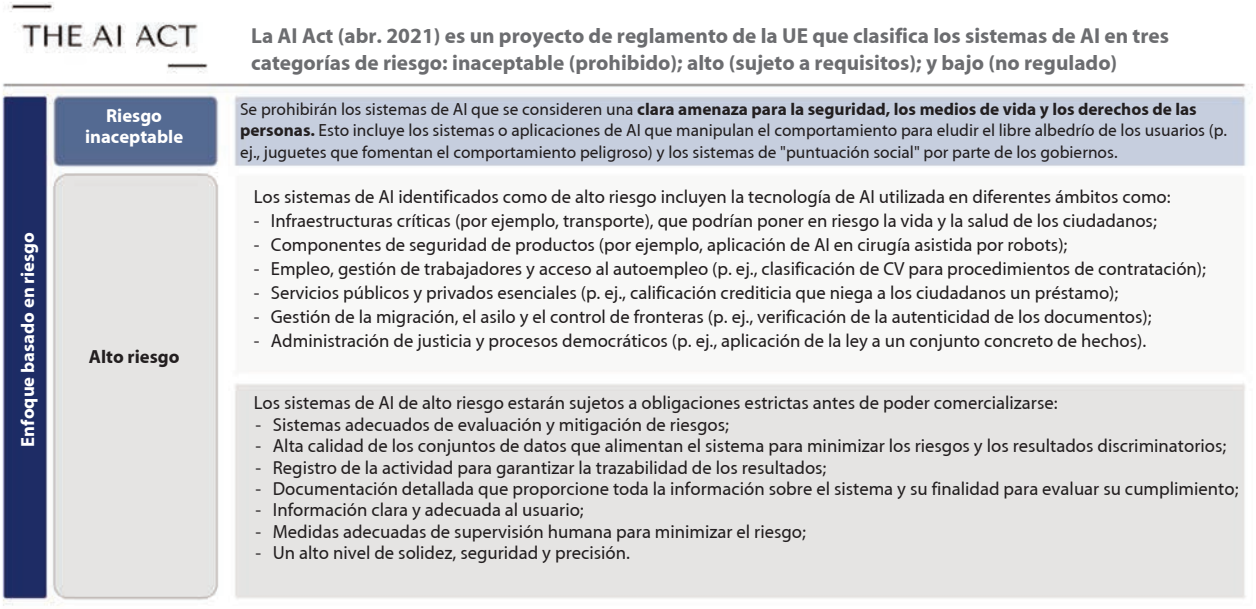
Como se puede apreciar, las Directrices apuntan en la misma dirección: el requerimiento (que se eleva al nivel de necesidad ética) de que los modelos de AI sean explicables.

Asimismo, lo que a primera vista podría parecer un requerimiento más relajado de interpretabilidad de los modelos de AI, por cuanto las Directrices reconocen que hay modelos de AI más difíciles de explicar, en realidad introduce una complejidad adicional: la necesidad de clasificar los modelos de AI según su riesgo y su potencial de ser interpretados, para aplicar un mayor o menor grado de esfuerzo en su explicación.

Por último, las Directrices están orientadas a evaluar en qué medida un modelo de AI cumple con estos siete requisitos, y para ello propone un listado de criterios de evaluación, que debe adaptarse a cada caso específico. En lo relativo a la explicabilidad, las Directrices formulan los siguientes criterios de evaluación[30], que deberían integrarse con otras herramientas de evaluación de las que ya dispongan las organizaciones:

[29] Comisión Europea (2019).
[30] Ibíd

Figura 4. Ámbitos de aplicación y requerimientos de la Artificial Intelligence Act.



- ¿Ha evaluado en qué medida son comprensibles las decisiones y, por tanto, el resultado producido por el sistema de AI?
- ¿Se ha asegurado de que se pueda elaborar una explicación comprensible para todos los usuarios que puedan desearla sobre las razones por las que un sistema adoptó una decisión determinada que diera lugar a un resultado específico?
- ¿Ha evaluado en qué medida la decisión del sistema influye en los procesos de adopción de decisiones de la organización?
- ¿Ha evaluado por qué se desplegó ese sistema en particular en esa área concreta?
- ¿Ha evaluado el modelo de negocio del sistema (por ejemplo, de qué modo crea valor para la organización)?
- ¿Ha diseñado el sistema de AI teniendo en mente desde el principio la interpretabilidad?
- ¿Ha investigado y tratado de utilizar el modelo más sencillo e interpretable posible para la aplicación en cuestión?
- ¿Ha evaluado si puede analizar sus datos relativos a la formación y los ensayos realizados? ¿Puede modificar y actualizar estos datos a lo largo del tiempo?
- ¿Ha evaluado si, tras la formación y el desarrollo del modelo, tiene alguna posibilidad de examinar su interpretabilidad o si dispone de acceso al flujo de trabajo interno del modelo?

4. Blueprint for an AI Bill of Rights (White House)

En octubre de 2022, la Casa Blanca propuso un Borrador de Declaración de Derechos sobre inteligencia artificial[31], impulsado por el presidente Joe Biden y desarrollado por la Oficina de Política Científica y Tecnológica (OSTP) de la Casa Blanca, y se acompaña de un manual *(From principles to practice)* sobre cómo implementarlo en la práctica.

El *AI Bill of Rights* establece cinco principios o derechos de los ciudadanos en lo relativo a la AI, que se resumen en[32]:

- Sistemas seguros y efectivos.
- Protección contra la discriminación de los algoritmos.
- Privacidad de los datos.
- Notificación y explicación.
- Alternativa, evaluación por un ser humano y proceso de corrección en caso de fallo de la AI *(fallback)*.

Dentro de su cuarto principio, en lo relativo a la explicabilidad de los modelos de AI, establece, entre otros, que[33]:

> Los diseñadores, desarrolladores e implantadores de sistemas automáticos deben proporcionar documentación en lenguaje sencillo y generalmente accesible que incluya descripciones claras del funcionamiento general del sistema. [...]
>
> Los sistemas automáticos deben acompañarse de explicaciones que sean técnicamente válidas, significativas y útiles para usted y para cualquier operador u otras personas que necesiten entender el sistema. [...]
>
> Los sistemas automáticos deben proporcionar notificaciones de uso demostrablemente claras, oportunas, comprensibles y accesibles, y **explicaciones sobre cómo y por qué el sistema ha tomado una decisión** o realizado una acción.

5. Principios sobre inteligencia artificial (OCDE)

Los Principios de la OCDE sobre inteligencia artificial promueven el uso de una AI digna de confianza y que respete los derechos humanos y los valores democráticos. Fueron adoptados en mayo de 2019 por los 38 países miembros de la OCDE. Fueron los primeros principios de este tipo suscritos por gobiernos, e incluyen recomendaciones concretas para la política y la estrategia públicas sobre AI.

Entre otros, establecen que "los responsables de la AI deben comprometerse con la transparencia y la divulgación responsable de los sistemas de AI. Para ello, deben proporcionar información significativa, adecuada al contexto y coherente con el estado de la técnica […] para que los afectados por un sistema de IA puedan comprender el resultado"[34]. El Observatorio de Políticas de AI de la OCDE, lanzado en febrero de 2020, tiene como objetivo ayudar a los responsables a aplicar estos Principios.

6. Discussion paper on machine learning for IRB models (EBA)

Por su relevancia en el sector bancario, es destacable el *Discussion paper on machine learning for IRB models,* de la Autoridad Bancaria Europea (EBA), publicado en noviembre de 2021 (Fig. 5).

El documento tiene como objetivo analizar la relevancia de los posibles obstáculos para la implementación de técnicas de aprendizaje automático en el ámbito del enfoque IRB de cálculo de capital en entidades financieras, incluye los desafíos y los beneficios potenciales del uso de estas técnicas, y establece ciertos principios y recomendaciones[35]. Un eje central del documento es, lógicamente, cómo hacer compatible el uso de estas técnicas con el cumplimiento de la regulación europea sobre capital (CRR[36]).

[31] White House OSTP (2022).
[32] Ibíd.
[33] Ibíd.
[34] OECD (2019).
[35] Ver un análisis detallado en Management Solutions (2021).
[36] CRR: Capital Requirements Regulation, regulación central sobre capital en entidades financieras en Europa.

En lo relativo a la interpretabilidad de los modelos, el documento lo aborda bajo el epígrafe de "Preocupaciones sobre el uso de las técnicas de aprendizaje automático", y afirma[37]:

> Las principales preocupaciones derivadas del análisis de los requisitos de la CRR se refieren a la complejidad y fiabilidad de los modelos de ML, en los que los principales retos parecen ser la interpretabilidad de los resultados, la gobernanza, con especial referencia a las mayores necesidades de formación del personal, y la dificultad de evaluar la capacidad de generalización de un modelo (es decir, evitar el sobreajuste).
>
> Para comprender las relaciones subyacentes entre las variables explotadas por el modelo, los profesionales han desarrollado varias técnicas de interpretabilidad [...] [y] la elección de cuál de estas técnicas utilizar puede plantear un reto en sí misma, ya que a menudo estas técnicas solo permiten una comprensión limitada de la lógica del modelo.

Más allá de esto, el documento introduce la necesidad de encontrar un equilibrio entre complejidad e interpretabilidad del modelo, y, a diferencia de otra regulación, baja a un nivel más técnico al recomendar a las entidades financieras:

a. Analizar de forma estadística: i) la relación de cada variable de entrada con la variable de salida, *ceteris paribus*; ii) el peso global de cada variable de entrada en la determinación de la variable de salida, para detectar qué variables influyen más en la predicción del modelo. Estos análisis son especialmente pertinentes cuando no es posible determinar una representación estrecha y puntual de la relación entre la variable de salida del modelo y las variables de entrada debido a la complejidad del modelo.

b. Evaluar la relación económica de cada variable de entrada con la variable de salida para garantizar que las estimaciones del modelo son plausibles e intuitivas.

c. Presentar un documento de síntesis que explique de forma sencilla el modelo a partir de los resultados de los análisis descritos en el punto a. El documento deberá describir como mínimo:
   i. Los factores clave del modelo.
   ii. Las principales relaciones entre las variables de entrada y las predicciones del modelo.

   Los destinatarios del documento son todas las partes interesadas, incluido el personal que utiliza el modelo con fines internos.

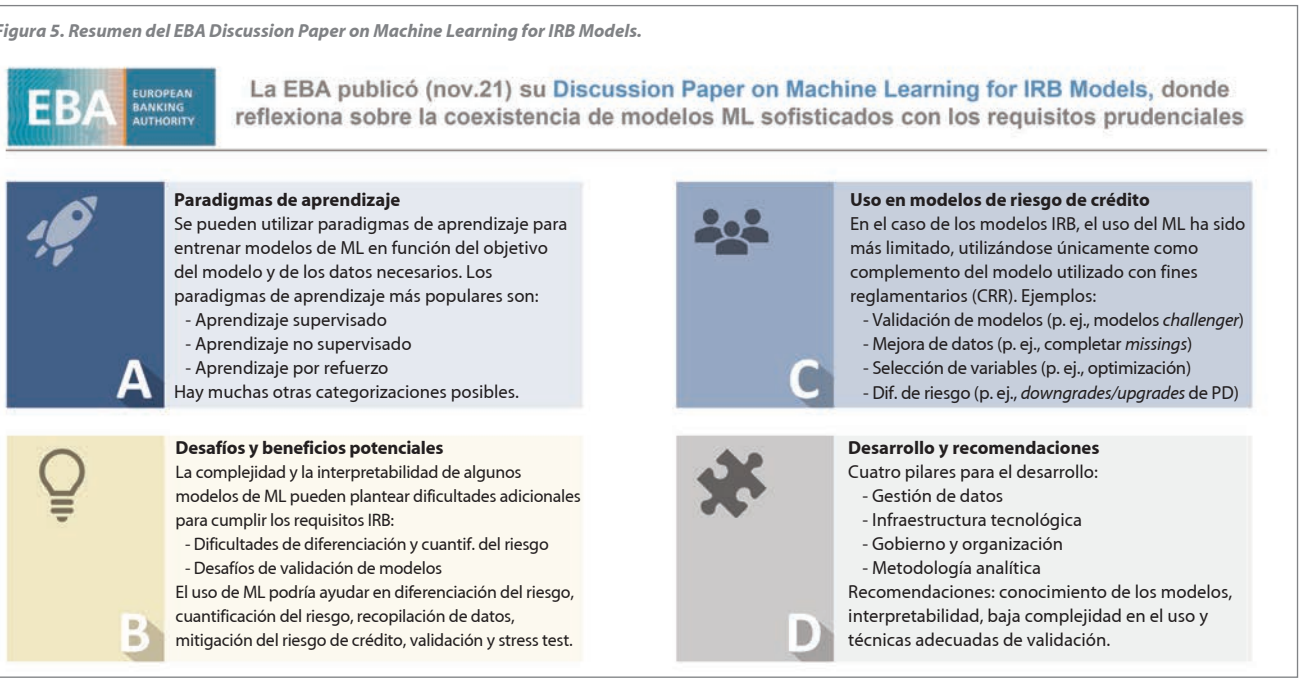
d. Garantizar la detección de posibles sesgos en el modelo (por ejemplo, un ajuste excesivo a la muestra de entrenamiento).

En la práctica, al tiempo que la industria bancaria espera la versión final del documento consultivo de la EBA, la mayor parte de entidades que utilizan técnicas de aprendizaje automático en sus modelos IRB ya están adaptando sus marcos de desarrollo, seguimiento y validación de modelos para asegurar su cumplimiento en el futuro.

Un elemento común a todas las referencias regulatorias mencionadas, como se puede apreciar, es la necesidad de proporcionar una explicación a los ciudadanos sobre el uso de la AI, y de hacerlo en dos niveles: la interpretabilidad y transparencia del modelo de AI en su conjunto, y la capacidad de explicar una decisión concreta del modelo, en caso de ser requerido.

[37] EBA (2021).

Figura 5. Resumen del EBA Discussion Paper on Machine Learning for IRB Models.



Más allá de las referencias regulatorias descritas, hay un gran número de publicaciones, principios, directrices y borradores de regulación en múltiples jurisdicciones que abordan la interpretabilidad de los modelos de AI, tanto de ámbito general como sectoriales, y tanto regionales como locales de cada país; la selección expuesta en esta sección incluye las consideradas de mayor ámbito y potencial influencia.

Impactos en organización y procesos

Un principio esencial de la XAI como disciplina es que, más allá del desarrollo de las técnicas específicas de explicabilidad o de la construcción de modelos inherentemente interpretables, esta explicabilidad e interpretabilidad se debe integrar en la organización y los procesos de la compañía.

Llevado a la práctica, este principio supone el desarrollo y la puesta en funcionamiento de un marco de XAI, que se puede estructurar en cuatro elementos:

1. Técnicas de interpretabilidad de los modelos de AI
2. Integración en los procesos de gestión del riesgo de modelo (MRM)
3. Soporte tecnológico
4. Factor humano

1. Técnicas de interpretabilidad de los modelos de AI

El núcleo de un marco de XAI lo constituyen las técnicas de interpretabilidad y explicabilidad, que de forma resumida se pueden clasificar en tres aspectos:

- Interpretabilidad del diseño del modelo: esto incluye analizar cómo se comportaría el modelo en diferentes escenarios (p. ej. ataques adversarios, escenarios extremos...), comprender cómo funcionan los submodelos y los conjuntos ("ensembles") de modelos, e integrar la interpretabilidad en el diseño del modelo aplicando restricciones durante su desarrollo.
- Interpretabilidad de los resultados del modelo: se refiere a detectar qué variables y cómo influyen en la predicción del modelo a través de la interpretabilidad local (LIME, SHAP, etc.) y global (PDP, importancia de las variables, modelos subrogados, análisis de sensibilidad); a evaluar el sentido económico de cada variable (p. ej., análisis de casos de uso de una muestra representativa de datos), y a asegurar que la documentación del modelo lo describe correctamente, incluidas las variables de entrada y su relación con los resultados.
- Otros aspectos: garantizar la detección de posibles sesgos en el modelo (p. ej., sobreentrenamiento, datos de entrada sesgados, errores en los datos) y supervisar periódicamente el modelo, especialmente cuando cambie su alcance o cuando se aplique a datos distintos de los de desarrollo.

Por su importancia, las principales técnicas de interpretabilidad y explicabilidad se desarrollan en la siguiente sección.

2. Integración en los procesos de gestión del riesgo de modelo (MRM)

La interpretabilidad de los modelos de AI es una característica que trasciende el desarrollo e impacta a lo largo de toda la cadena del ciclo de vida de los modelos, y por tanto en todo el marco de la gestión del riesgo de modelo. Un resumen no exhaustivo de la incorporación de XAI en el marco de MRM de una compañía incluye revisar los siguientes elementos:

- **Gobierno:** actualizar el marco de organización y gobierno para incorporar XAI; evaluar el impacto de la regulación aplicable a los modelos de AI; actualizar el sistema de tiering de los modelos para contemplar la falta de interpretabilidad como un mayor riesgo; actualizar el inventario y los procedimientos de inventariado de los modelos para incorporar los elementos de XAI (p. ej. atributos específicos para modelos de AI).
- **Desarrollo:** actualizar las políticas y procedimientos de desarrollo de los modelos, así como los requisitos de documentación; evaluar imparcialidad (*fairness*) y sesgos, interpretabilidad de inputs, diseño y resultados, datos, riesgo de proveedores, métricas de capacidad predictiva, límites al uso de los modelos de AI, etc.; realizar análisis de sensibilidad de los modelos de AI para identificar vulnerabilidades; incluir en el marco de desarrollo tests específicos para XAI.
- **Seguimiento:** actualizar el marco de seguimiento de los modelos y completarlo con tests específicos de XAI; revisar los umbrales y las acciones derivadas de su incumplimiento; desarrollar sistemas de alerta temprana para detectar cambios en los modelos de AI; revisar el cumplimiento del apetito al riesgo de modelo; valorar la necesidad de desarrollar un módulo de seguimiento ad hoc para modelos de aprendizaje dinámico (i.e. que se recalibran automáticamente sin intervención humana).

- **Validación:** actualizar el marco de validación interna para detectar posibles riesgos asociados a los modelos de AI e incorporar tests de XAI; establecer un marco de validación cruzada para garantizar la calidad de los modelos de AI; evaluar el impacto de los cambios en el entorno de producción en los modelos de AI.

- **Implementación:** actualizar el proceso de implementación del modelo para incorporar tests propios de las características de XAI; actualizar, en su caso, la plataforma tecnológica para permitir la puesta en producción de los modelos de AI.

- **Uso:** actualizar los procedimientos de uso de los modelos de AI para determinar su adecuación al contexto en que se van a emplear; revisar y completar la formación a usuarios respecto a los modelos de AI; actualizar los protocolos para detectar posibles situaciones de mal uso o explotación de los modelos.

- **Auditoría:** implementar un marco de auditoría de los modelos de AI para asegurar su adecuada implementación y uso; establecer tests de XAI para la auditoría de los modelos de AI; evaluar la adecuación de los sistemas de control interno para garantizar la calidad de los modelos de AI; analizar los registros de auditoría para detectar posibles riesgos asociados a los modelos de AI.

Por tanto, el uso de modelos de AI conlleva una revisión completa de las políticas y procedimientos a lo largo de todo el ciclo de vida del modelo para incorporar, como mínimo, los elementos propios de XAI.

3. Soporte tecnológico

La implementación de un marco de XAI tiende a comenzar por herramientas departamentales, y tan pronto como alcanza un mínimo nivel de madurez, requiere de soluciones tecnológicas profesionales para dar soporte a los aspectos propios de la interpretabilidad de los modelos de AI.

Estas soluciones pueden clasificarse en dos grupos:

- Interpretabilidad: desarrollo de sistemas que implementen las técnicas de interpretabilidad de forma estandarizada y homogénea. Deben permitir realizar la interpretación de los modelos de forma automática, fácilmente configurable y con una alta calidad, incorporando las técnicas más comunes y proporcionando flexibilidad para añadir nuevas técnicas conforme se desarrollen[38].

- Gobierno de modelos: desarrollo o actualización de los sistemas de gobierno de modelos para dar soporte a los aspectos de XAI en MRM (inventario, *tiering*, documentación, etc.), asegurando así que los modelos disponibles cumplan los requisitos de calidad, seguridad y explicabilidad requeridos[39].

Más allá de esto, es recomendable una aproximación holística que abarque todos los aspectos del marco de XAI. Esto incluye el uso de herramientas de análisis de datos, el desarrollo de APIs para la integración de los sistemas de interpretabilidad y gobierno de modelos antes descritos, la creación de mecanismos de seguridad y auditoría, y la definición de protocolos para garantizar el cumplimiento de los estándares de calidad y explicabilidad.

4. Factor humano

Un cuarto elemento en la integración de XAI en la organización y procesos es la consideración del factor humano. Esto incluye, entre otros:

- **Captación y retención del talento:** desarrollo de programas para la captación y retención del talento especializado en XAI, para asegurar la presencia de profesionales con los conocimientos técnicos y la experiencia necesaria para aplicar XAI en la compañía, lo que es especialmente relevante en un mercado laboral con escasez de este perfil profesional.

- **Formación:** desarrollo de programas de formación para equipos de desarrollo de modelos de AI, equipos de gobierno de modelos y usuarios de los modelos de AI, con el fin de asegurar que todos los involucrados comprendan los principios básicos de XAI y cómo aplicarlos en el contexto específico de la compañía.

- **Cultura:** desarrollo de una cultura en la compañía que potencie el uso y la explotación de la explicabilidad y la interpretabilidad de los modelos de AI. Esto puede incluir la adopción de metodologías ágiles para el desarrollo de modelos de AI, la creación de una cultura de colaboración entre los equipos de desarrollo de modelos y de gobierno de modelos, y la consideración de la explicabilidad como un factor crítico en la aprobación de los modelos de AI.

- **Gestión del cambio:** desarrollo de programas de gestión del cambio para asegurar la adecuada adopción de XAI por parte de los equipos de la compañía que trabajan con modelos de AI. Esto incluye motivar a los equipos de desarrollo, el análisis de los costes y beneficios de la explicabilidad, la definición de protocolos de comunicación con terceros, etc.

En conclusión, la explicabilidad y la interpretabilidad de los modelos de AI son aspectos clave que deben integrarse en la organización y los procesos de la compañía mediante un marco apropiado y completo de XAI, lo que resulta esencial para garantizar el uso de estos modelos conforme a la regulación y las buenas prácticas.

[38] A este respecto, Management Solutions dispone de ModelCraft™, un sistema propietario de AutoML y modelización por componentes, que incorpora un módulo completo de interpretabilidad. Ver Management Solutions (2023).

[39] Management Solutions dispone de Gamma™, un sistema propietario de gobierno de modelos que cubre todos los aspectos mencionados. Ver Management Solutions (2022).