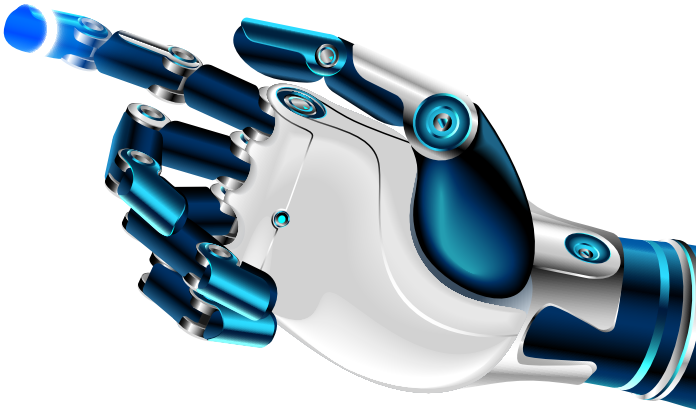
BBVA
Creando Oportunidades

Inteligencia Artificial Responsable
desde el diseño

Esther de la Torre Gordaliza

Índice



Durante los últimos años, tanto grupos de expertos, como instituciones, han planteado cuáles son los retos de una inteligencia artificial (IA) ética. En BBVA, los equipos de DATA, Client Solutions y Responsible Business hemos reflexionado sobre cómo llevar estos retos a cabo, integrando la responsabilidad de forma natural en las principales fases de un proyecto de IA: Ideación, recopilación de datos, desarrollo del sistema y lanzamiento, control y retroalimentación.

Existen diversos Sistemas de Inteligencia Artificial **(IA)**, entendidos como lo hace el proyecto de reglamento de la Comisión Europea[1]: desde el que se basa en un modelo de regresión lineal, que sería un sistema sencillo, hasta el que utiliza una red neuronal, que sería un sistema complejo. La responsabilidad aplicada a los sistemas de IA supone, independientemente del tipo de modelo utilizado:

Poner a las personas en el centro

Teniendo en cuenta el respeto de los derechos humanos, la protección de los datos personales, el fomento de la igualdad de oportunidades, la transparencia y la libertad de elección, así como las necesidades y expectativas de las personas

Mantener el rigor y la prudencia

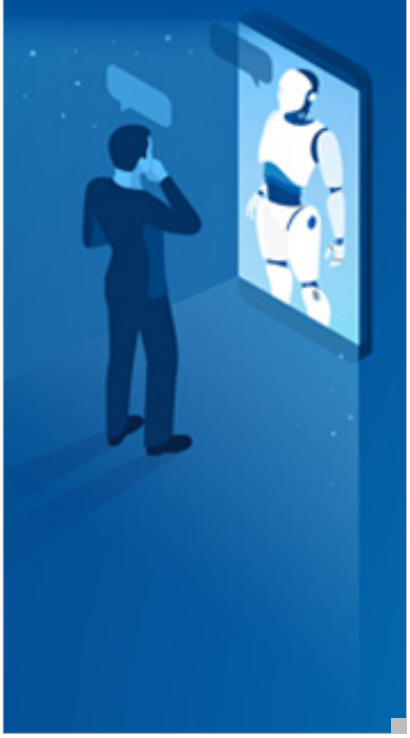
Aspirando a la excelencia a la hora de desarrollar, mantener y mejorar los sistemas de IA así como los mecanismos de control.

[1] Sistema de inteligencia artificial ": software desarrollado utilizando una o más de las técnicas y enfoques enumerados en el anexo I del Reglamento y que puede, para un conjunto dado de objetivos definidos por el ser humano, generar resultados como contenido, predicciones, recomendaciones o decisiones que influyan los entornos con los que interactúan.Proposal for a Regulation laying down harmonised rules on artificial intelligence EU Commision

01

Ideación

Para establecer **requerimientos y controles posteriores,** en esta fase es importante abordar determinadas **acciones y decisiones** como:



A) Analizar el caso de uso y determinar su nivel de impacto potencial en la vida de las personas.

- El caso de uso es aquello para lo que vamos a utilizar un sistema de IA. Por ejemplo, para recomendar películas, para ayudar a los clientes de un banco a manejar sus finanzas, o para llevar a cabo un diagnóstico clínico por análisis de imágenes.

- Para su definición es recomendable tener en cuenta aspectos como las necesidades del usuario, las del negocio, los posibles riesgos y si el caso de uso es ético y está alineado con los valores de la compañía. Por ejemplo, una empresa que fomenta la salud financiera tendría bajo su paraguas casos de uso como ayudar a las personas a manejar mejor sus finanzas.

- También es importante medir el impacto que pueden tener un error en el modelo, ya sea un falso positivo o un falso negativo[2], o posibles sesgos injustos. El impacto de un error en un modelo que recomienda una película, es menor que el de un modelo que permite acceder a un crédito. En este segundo caso, el impacto es importante tanto si el error no permite acceder al crédito a una persona que sí tiene capacidad de asumirlo, como cuando el error permite acceder al crédito a una persona que está en situación de sobreendeudamiento.

- Analizar estos aspectos nos ayudará a establecer un nivel de precisión y una serie de requerimientos y controles posteriores. En líneas generales éstos serán mayores cuanto mayor sea el impacto potencial del sistema en la vida de las personas.

[2] Dado un modelo que analiza si los accesos a un correo electrónico web son o no ciberataques, puede tener un falso positivo cuando bloquea el acceso al usuario legítimo,si se conecta por ejemplo desde otro país.

B) Determinar el grado de explicabilidad necesario

- La explicabilidad, dicho de forma sencilla, es disponer de las explicaciones necesarias para **entender por qué un modelo ha llegado a determinados resultados.** Hay modelos sencillos que son explicables, y otros más complejos en los que es necesario realizar acciones para que lo sean.

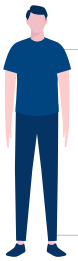
- Los motivos por los que es necesario un mínimo grado de explicabilidad en función de la aplicación del modelo son, entre otros: **conseguir transparencia en la experiencia de usuario, ejercer control sobre el modelo y cumplir la regulación.**

- **Determinar el grado de explicabilidad adecuado nos ayudará a decidir qué tipo de modelo elegiremos para el sistema, y a planificar tareas de explicabilidad necesarias.** En términos generales, y siempre en alineamiento con la regulación aplicable, a mayor impacto en las personas del caso de uso, más necesidad de entender el comportamiento del modelo.

C) Analizar el grado de automatización y control humano adecuado

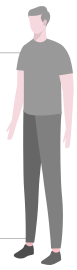
- **La automatización es muy positiva para maximizar la rapidez y la eficiencia,** pero siempre teniendo en cuenta **el caso de uso y los riesgos potenciales.**

- Existen tres grandes niveles de **control humano en los procesos automatizados**:



CONTROL HUMANO ALTO

El modelo proporciona recomendaciones y el humano decide.



CONTROL HUMANO MEDIO

El humano ajusta parámetros durante la ejecución del modelo.

CONTROL HUMANO BAJO

El humano efectúa controles posteriores a la ejecución del modelo.

- Este análisis nos ayudará a decidir qué nivel de automatización requiere el caso de uso. En general, aquellos **casos de uso de mayor impacto en las personas, precisarán un mayor control humano en el proceso.**

02

Recopilación de datos

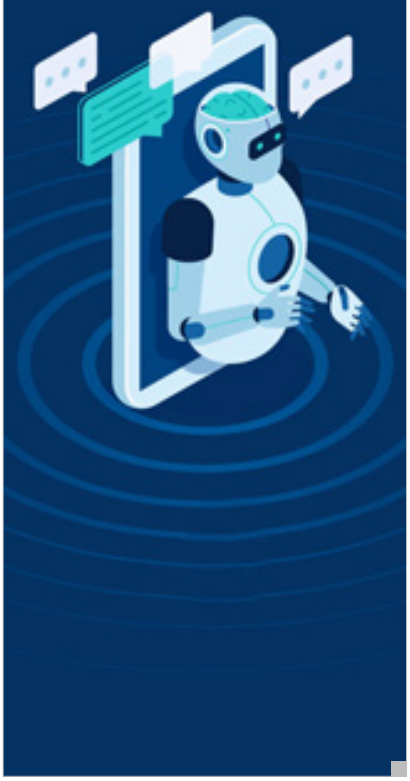
Obtener los datos necesarios y confiables es clave para el funcionamiento del modelo. En función, por ejemplo, del caso de uso y del grado de precisión establecidos en la fase anterior, debemos asegurarnos de que los datos son adecuados para los objetivos del proyecto, tienen la calidad necesaria, cumplen con los requerimientos de protección de datos y privacidad, están correctamente etiquetados, son representativos, y disponemos de una cantidad suficiente ya que la IA necesita gran cantidad de datos para su entrenamiento.

Para evitar sesgos injustos, hay que analizar si el conjunto de datos representa de manera equitativa la realidad que queremos representar y nos ayuda a detectar desviaciones y lo que nos permitiría evitar arrastrarlas en el desarrollo de los modelos.

03

Diseño y desarrollo del sistema

Diseñar el sistema de IA abarca tanto la **creación del modelo analítico como el diseño de la interfaz de usuario que lo integra;** algunas de las acciones a realizar serían las siguientes:



A) Elegir y desarrollar el modelo adecuado para el caso de uso

- De la fase de ideación tenemos el nivel de precisión y las necesidades de explicabilidad requeridas.
- Elegir el modelo adecuado supone acercarnos a un equilibrio entre maximizar la precisión, la adecuación a la expectativa del usuario, la explicabilidad, así como minimizar la complejidad y el nivel de riesgo.

B) Analizar si cumplimos con la expectativa de precisión definida

- En la fase de ideación, se ha marcado una precisión necesaria dado, entre otras cuestiones, el caso de uso y el impacto en la vida de las personas. En esta fase de desarrollo del modelo podemos comprobar que sus resultados nos dan el grado de precisión establecido.
- Esto nos permitirá ajustar lo necesario antes del lanzamiento.

C) Realizar un test de equidad con los resultados del modelo

- En la fase de recopilación de datos se analiza si el conjunto de los datos de origen están libres de sesgos, y por tanto son equitativos. En esta fase tenemos que hacer ese análisis sobre los resultados del modelo. Puede ser que en el proceso de depuración y preparación de los datos se haya perdido representatividad o, incluso, que, aunque hayamos excluido información sensible en los datos de origen, el modelo realice correlaciones entre otros datos, por ejemplo, compras de un determinado establecimiento, y género y produzca resultados sesgados o discriminatorios.
- Este test nos servirá para prevenir que los resultados del modelo repliquen sesgos injustos.

D) Conseguir el grado de explicabilidad establecido

- Si el caso de uso, ya sea por temas normativos, o de impacto del sistema requiere cierto grado de explicabilidad, se debe trabajar en ello en esta fase.
- Nos ayudará a entender por qué el modelo ha llegado a determinados resultados.

E) Recibir la opinión de los usuarios acerca del sistema

- Es necesario testar la interfaz con la que interactúan los usuarios antes del lanzamiento. Puede ser mediante pruebas y entrevistas personales, y es importante incluir a usuarios con distintos perfiles y capacidades con el objetivo de que la solución sea inclusiva. Entre los objetivos de investigación estaría saber si:

1 | El caso de uso y su ejecución cumplen con las expectativas del usuario

- El objetivo es explorar si el usuario percibe que el sistema está **en línea con los valores de la compañía,** le aporta **valor añadido** y da **respuesta a sus necesidades** y si los mensajes y el diseño le trasladan **claridad y transparencia.**
- Asimismo, es interesante validar con el usuario su **percepción del impacto de un error (como un falso positivo o un falso negativo en un modelo de clasificación),** y del **nivel de explicaciones acerca de los resultados** que espera, por si tuviésemos que realizar algún ajuste.

2 | El sistema da la autonomía necesaria al usuario

- Los clics, o la navegación del usuario, si contamos con su consentimiento, se pueden utilizar para retroalimentar el algoritmo y personalizar sus resultados para que **estén cada vez más alineados con las preferencias o el perfil del usuario.** Esto explica, que algunos buscadores nos den resultados de noticias u ofertas, en función de búsquedas anteriores.

- Es un reto encontrar el **equilibrio entre la facilidad, la personalización y las opciones de elección por parte del usuario.** Normalmente, a mayor personalización y menor número de opciones que dé una interfaz, más **facilidad de elección.** Pero dar menos opciones, o no utilizar adecuadamente el perfilado de clientes, puede restar capacidad de elección al usuario. Es fundamental **cumplir con la regulación vigente,** ofrecer la **información adecuada** para que el usuario pueda **tomar decisiones informadas y buscar el equilibrio** con sus preferencias personales.

3 | Las predicciones se trasladan de forma clara a los usuarios:

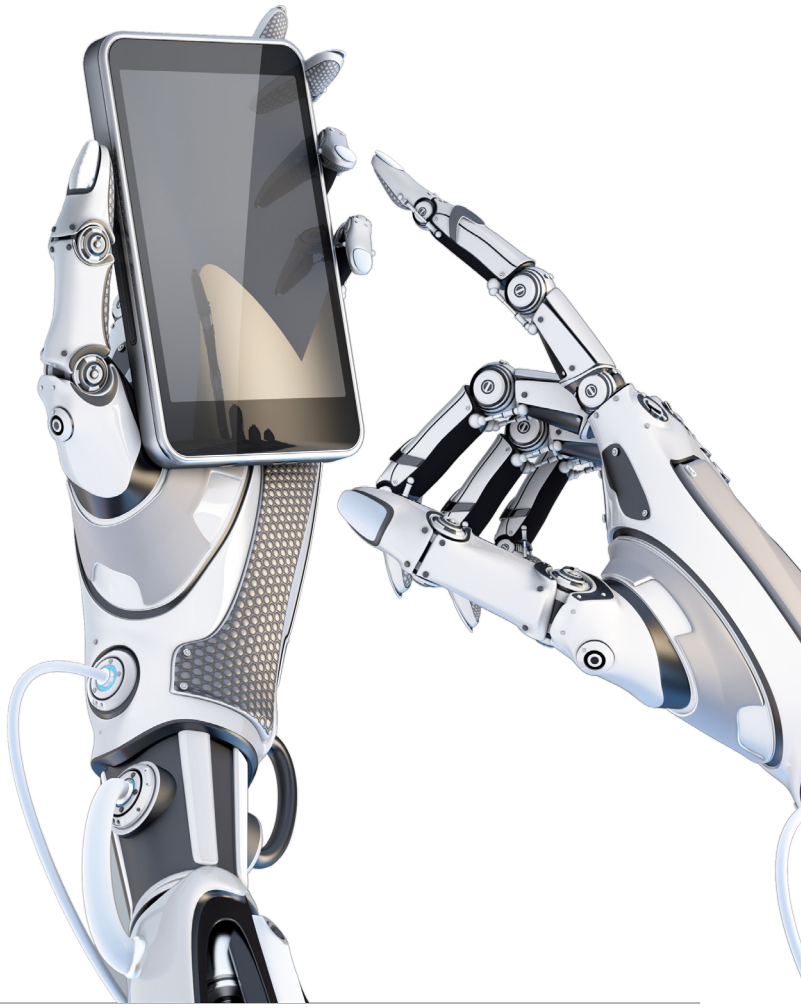
- Los modelos dan resultados que pueden cumplirse con un determinado grado de certidumbre.

- Es importante analizar **en cada caso de uso,** si **la forma de presentar las predicciones es comprensible, si percibe que están sustentadas**, si queda claro que se trata **de una predicción y no de un hecho seguro,** y si transmitimos a los usuarios que la **predicción supone una información adicional para ayudarles a tomar decisiones.**

4 | El sistema es inclusivo y accesible

- **El sistema debe ser accesible para personas con algún tipo de discapacidad, temporal o permanente, y considerar la diversidad en los niveles de comprensión lectora o digitalización.**

- Un test de accesibilidad nos dirá si la interfaz de usuarios presenta tanto a nivel de diseño, como de desarrollo, **barreras de accesibilidad**.



04

Lanzamiento, control y retroalimentación

Es recomendable fijar **mecanismos de control** que se mantengan durante la vida del sistema.

- En lo relativo a los modelos, es relevante monitorear y revisar sus resultados para asegurar que mantienen, entre otros aspectos, el grado de precisión, explicabilidad, y equidad establecidos. Especialmente cuanto mayor es el nivel de automatización y si utilizamos modelos con aprendizaje automático por refuerzo cuyo funcionamiento evoluciona en el tiempo.

- Es importante recabar la opinión del usuario, y que esta llegue de forma adecuada a las áreas y personas involucradas para ajustar lo necesario, cerrando un ciclo de mejora continua. La opinión de los usuarios sirve tanto para mantener su satisfacción, como para mejorar el comportamiento del modelo y de la interfaz.

- Este flujo de trabajo debe mantenerse actualizado, así como tener una gobernanza y unas medidas de remediación.

Conclusiones

Conseguir IA responsable supone **poner a las personas en el centro y actuar con rigor y prudencia.**

- En función del caso de uso, y de su impacto en la vida de las personas, es recomendable realizar una serie de acciones y establecer en mayor o menor medida una serie de requerimientos y controles a plantear desde el diseño del sistema.
- Las principales cuestiones a tener en cuenta para poner a las personas en el centro, son: el fin ético del caso de uso, el respeto de los derechos humanos, el fomento de la igualdad de oportunidades, la libertad de elección y la transparencia, y la búsqueda de la mejor experiencia del usuario.
- Para ello es interesante tener el control de los datos de origen y de los modelos, controlando la calidad de ambos, establecer un grado de automatización adecuado, medir el correcto funcionamiento del sistema y diseñar de forma inclusiva, teniendo en cuenta, las necesidades de las personas.

Referencias

EU Commission

_ Ethics Guidelines on Trustworthy AI

_ Proposal for a Regulation laying down harmonised rules on artificial intelligence

OCDE

_Recommendation of the Council on Artificial Intelligence also referred by

G20

_Ministerial Statement on Trade and Digital Economy

IEEE

_AI Ethic Aligned Design

FSB

_"Artificial intelligence and machine learning in financial services"

Monetary Authority of Singapore

_ Principles for AI and Data Analytics in Singapore's Financial Sector

PDPC

_Model AI Governance Framework PDPC

Canadá Government

_Directive on Automated Decision-Making

WEF

_Empowering AI Leadership

Asilomar

_AI Principles

IIF

_Bias and ethical implications in ML

EBA
Final Report on big data and Advanced Analytics

AEPD

_AEPD AI Guidelines

_Requisitos auditorías para tratamientos que incluyan IA

ACM
code for Ethics and professional Conduct

Université de Montréal

_Montreal Déclaration

European Institute for Science, Media and Democracy

_Ai4people

BBVA

Creando Oportunidades