Sabadell Hub Empresa

#SabadellHubEmpresa #Valencia

IA en la empresa: regulación y ética, claves para la innovación responsable

5 de julio de 2024

Carlota Mañas
Data Scientist | Banco Sabadell

Yago Casal
Partner | Technology & Transformation

IA en la Empresa: regulación y ética, claves para la innovación responsable

Analítica e Inteligencia Artificial

Carlota Mañas Bagué, Data Scientist

Julio 2024

ÍNDICE

1

Ética en IA: Un Camino Hacia la Responsabilidad

RELEVANCIA DE LA ÉTICA EN LA IA

A lo largo de los años se ha aumentado el uso de soluciones basadas en **Inteligencia Artificial** y ha puesto sobre la mesa la necesidad de implementar una **metodología ética** para evitar que las empresas implementen modelos erróneos en cuestiones de protección de datos y sesgos en algoritmos.

La ética en los modelos de IA es fundamental para garantizar que las tecnologías se desarrollen y apliquen de manera justa y responsable.

Es importante porque …

- **Impulsa la confianza de nuestros clientes**
- **Promueve la justicia y la equidad**
- **Garantiza la transparencia y responsabilidad**
- **Fomenta la innovación responsable**

REUTERS

Amazon scraps secret AI recruiting tool that showed bias against women

La automatización del proceso de contratación de Amazon destapó un problema intrínseco de los datos. Al tener muchos más currículums de hombres que de mujeres, el modelo daba preferencia a contratar hombres.

REUTERS

Microsoft's AI Twitter bot goes dark after racist, sexist tweets

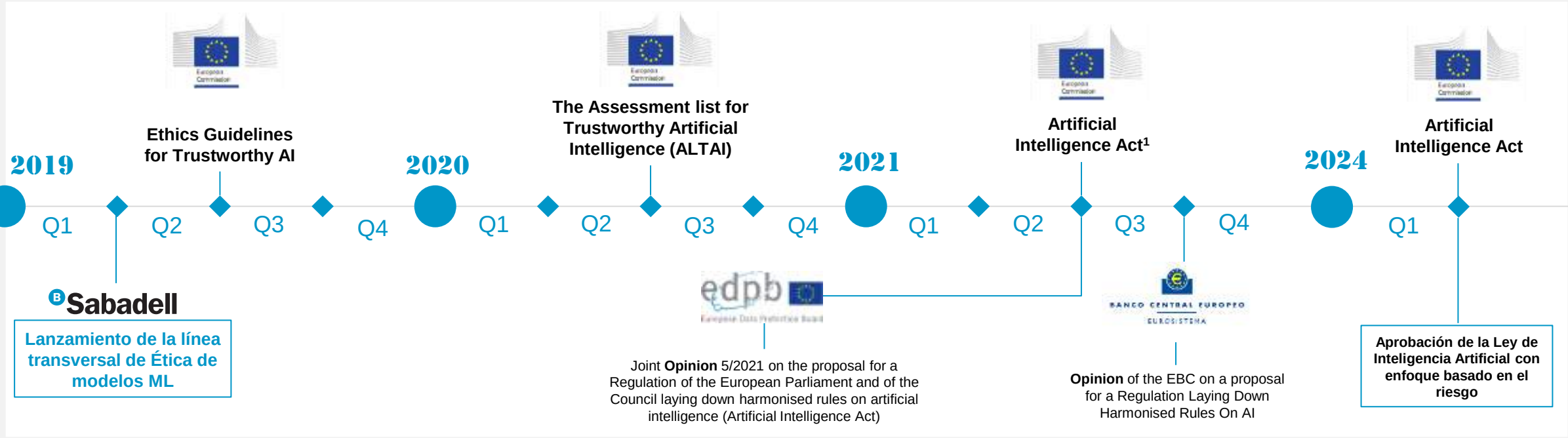
Microsoft lanzó un Bot en Twitter que aprendía a medida que otros usuarios interactuaban con él. Varios usuarios se coordinaron para “tweetear” mensajes ofensivos, por lo que los datos con los que aprendía el Bot no eran adecuados.

Un mal tratamiento de datos conlleva a multas cuantiosas por incumplir GDPR. Si no se aborda correctamente la regulación en IA y no se dispone de un buen marco de **gobierno de modelos**, puede haber el mismo tipo de multas económicas y efectos reputacionales.

ÉTICA EN LA VANGUARDIA DEL ML

En Banco Sabadell empezamos en 2019 el viaje en el análisis de la ética de modelos y datos a raíz de notícias de grandes empresas tecnológicas que experimentaron sesgos notorios en sus algoritmos, anticipándonos a las publicaciones de la Comisión Europea.

Nos inspiraron para actuar tras observar estos incidentes y se puso en relieve la importancia de diseñar una metodología ética y definir un gobierno.



[1] *Tiene carácter regulatorio y aplicará a todos los estados miembros de la UE. Por el momento, se trata de un* ***borrador***

PRINCIPIOS ÉTICOS BANCO SABADELL

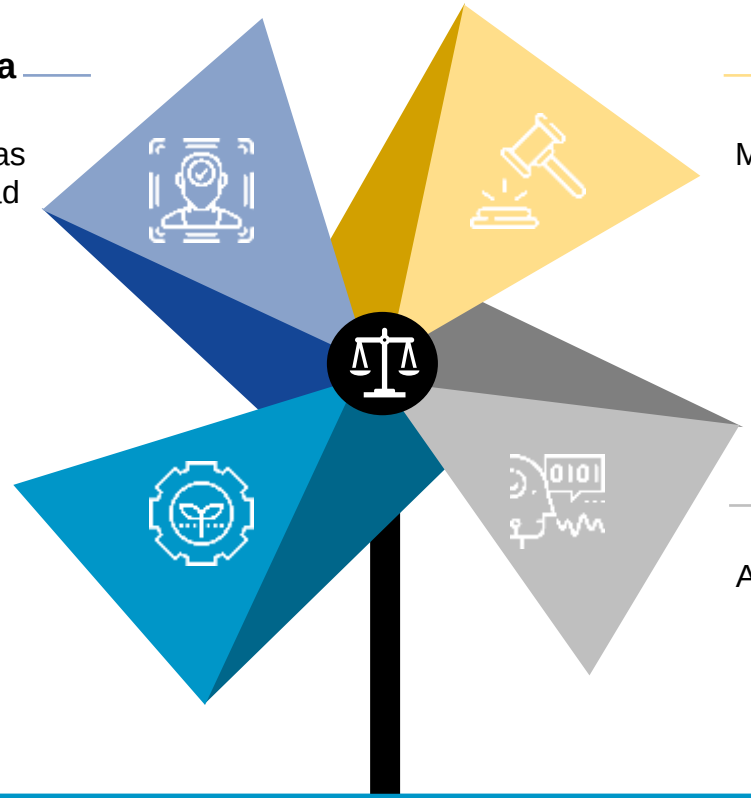
Desde **BS** se ha desarrollado una **propuesta de principios éticos propios**, **excluyendo** todos aquellos **ámbitos** ya **recogidos en la regulación actual vigente.** Se consideran la base de todo proyecto que incluya algoritmos de Inteligencia Artificial o de analítica y se diseñan con el propósito de cumplir con los requerimientos de la regulación europea.

Supervisión y monitorización humana

Garantizar que todas las **soluciones** diseñadas estén **revisadas** por trabajadores de la entidad y asegurar un exhaustivo seguimiento que cerciore el **buen funcionamiento** de los **algoritmos**

Justicia

Medidas orientadas a garantizar que los **algoritmos** sean **equitativos**, **inclusivos** y **libres de sesgos** y **prejuicios**



Sostenibilidad

Diseño de **algoritmos responsables** con el **medio ambiente**

Transparencia y explicabilidad

Asegurar que los **algoritmos** sean **trazables**, **claros** y puedan **justificarse** todas las **decisiones** tomadas en base a ellos

ACCIONES RELACIONADAS CON IA EN BS

Sabadell Hub Empresa

Sabadell

Aproximación a Data Ethics

Analítica para sostenibilidad

Checklist Ética en el ciclo de vida de datos y modelos analíticos

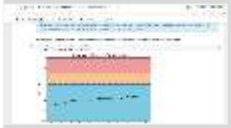
Ciclo de vida de datos y modelos analíticos

Survey on the Trustworthy AI Assessment List

AI Ethics White Book

Formaciones monitorización funcional y explicabilidad

Gobierno de la inteligencia



Notebooks monitorización funcional y explicabilidad

Dashboard monitorización funcional y explicabilidad

Benchmark. Análisis de proveedores de soluciones de Ética

Benchmark. Análisis de situación en BS, competencia y tecnologías

Monitorización funcional. Strategy Room

Publicaciones sobre Ética en IA Noemí Brito

Situación ética en IA

Cuadro de mando de métricas de ética

Librería de métricas en R

Librería de métricas en Python

Paper monitorización funcional

Paper explicabilidad

Resumen regulación IA

Repositorio de ética

2019/2020 | **2021 Q1-Q2** | **2021 Q3-Q4** | **2022 Q1-Q2** | **2022 Q3-Q4**

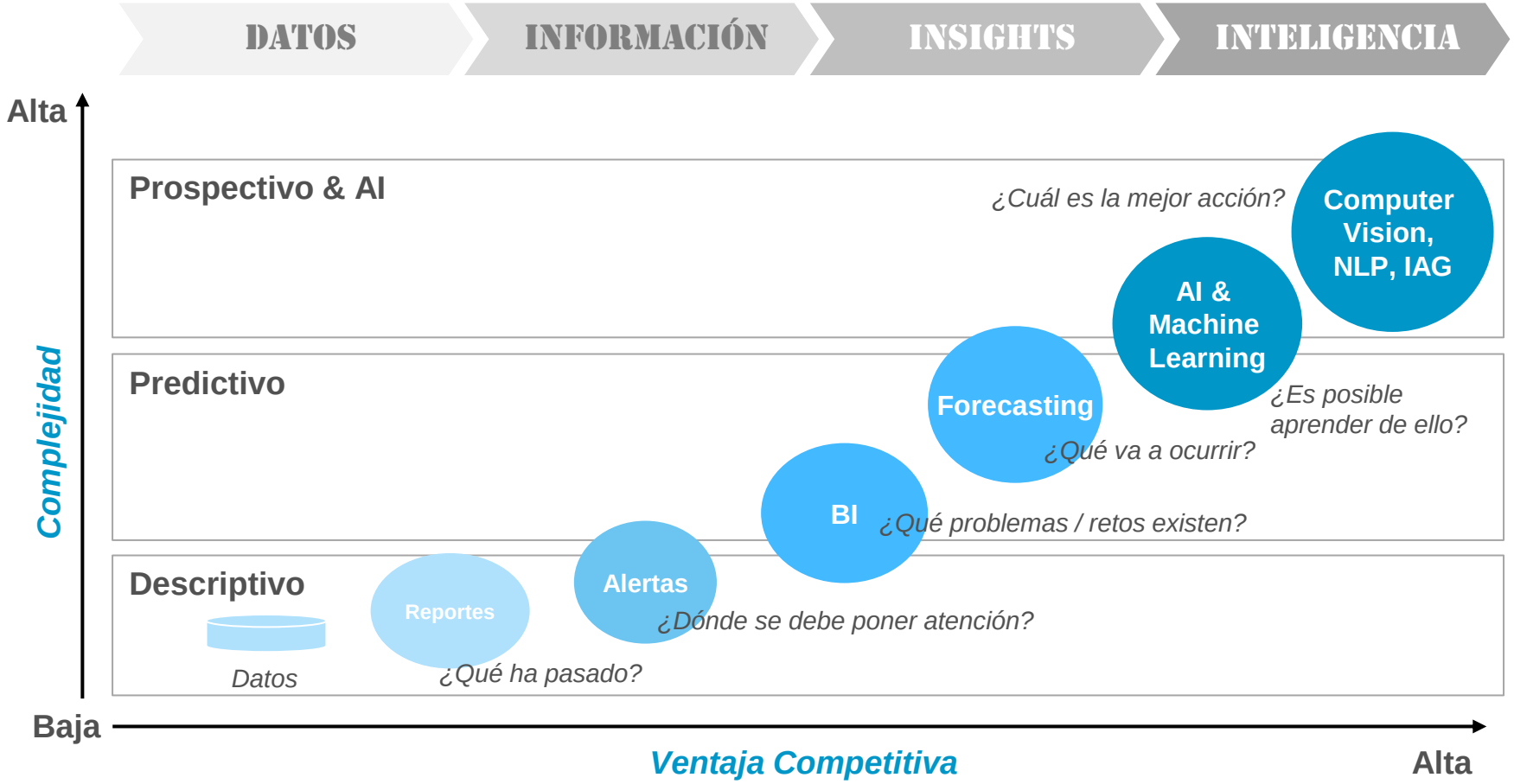
2

La Ética en acción: un caso de éxito

APORTANDO VALOR A TRAVÉS DEL USO DE LA IA

A pesar de que el uso de técnicas de analítica avanzada conlleva una **mayor complejidad**: tanto en algoritmos, procesamiento de la información, …

… su aplicación también deriva en una **mayor ventaja competitiva** y un importante **incremento** de **valor** para la Entidad



SITUACIÓN ACTUAL EN A&IA

Gobierno

- Implantación de un gobierno en el del **ciclo de vida** de los modelos: desde la identificación de la necesidad, hasta su industrialización o decomisado.
- **Adaptación del modelo de riesgos y desarrollo actual**, desde los desarrolladores hasta los comités de aprobación, y definición de nuevos elementos a tener en cuenta
- Diseño de nuevos **parámetros y principios** que permitan ampliar el modelo de gestión y gobierno de proyectos a la nueva normativa de manera unificada y transversal

Metodología

- Construcción de un **repositorio único** que refleje el pipeline de modelos del Banco, permitiendo:
  - Visualizar la **situación** en la que se encuentra la **entidad**
  - Reportar el estado al regulador y disponiendo del registro de actividades (evitando posibles **sanciones**)
  - Asegurar un **correcto uso de la IA** mediante la supervisión y monitorización de modelos

Técnico

- **Investigación** e **implementación** continua de nuevas técnicas de análisis. Se trabajan los tres ejes técnicos más relevantes: **Functional Monitoring, Explicabilidad y Sesgo**
- Creación de una **librería** que recoge las variables de interés para poder disponer de un repositorio
- Disposición de **herramientas** de análisis y de seguimiento de los modelos
- Conformación de un equipo con ***expertise*** suficiente para abordar los retos inherentes al desarrollo de una IA éticamente responsable

Divulgación y estado del arte

- Generación de **material científico** para la difusión interna
- **Formación** continua a las distintas áreas del Banco
- Transmisión de la metodología y la importancia de la ética en la IA a través de la creación de **foros/eventos/ encuentros divulgativos**
- **Difusión** del posicionamiento de Banco Sabadell por medio de la **publicación** de **entrevistas** a empleados **referentes** en el **campo de la ética**

CASO DE USO DE NEGOCIO BANCARIO

Fraude en Transferencias Internacionales

- Modelo desarrollado para el área de negocio **Fraude**
- El fraude bancario está **evolucionando hacia entornos cada vez más complejos** que hacen que las técnicas tradicionales de prevención y detección de fraude no consigan detectar todo el fraude que sufre una Entidad ni anticiparse a él. El departamento de Hechos Delictivos necesita poder **renovar, eficientar y automatizar** sus procesos de detección de fraude

Objetivo — ***Predecir transferencias internacionales fraudulentas realizadas por clientes particulares de la entidad***

Alcance

Del total de transferencias internacionales entre los años 2019 y 2020, se han detectado como fraudulentas el 0.2%

Caso de uso

Scoring de casos fraudulentos. Además, proporciona al departamento de fraude información adicional para robustecer sus propios procesos de toma de decisiones

Alcance y caso de uso

¿Qué ha aportado la explicabilidad?

1. Creación de nuevas variables
2. Coherencia entre el significado de la variable y el fraude en las predicciones según su valor
3. Explicar localmente por qué una transferencia concreta se considera fraudulenta

Explicabilidad

¿Qué beneficios se obtienen de este modelo?

Del total de transferencias enviadas y revisadas por el equipo de fraude (en total160) un 34% fueron fraude en Q4 2022

Detección de 365 transferencias fraudulentas (55 por el modelo) evitando el robo de 124.000€ en Q4 2022

Detección de un intento de fraude, evitando el robo de 73.000 € en abril de 2023

Resultados

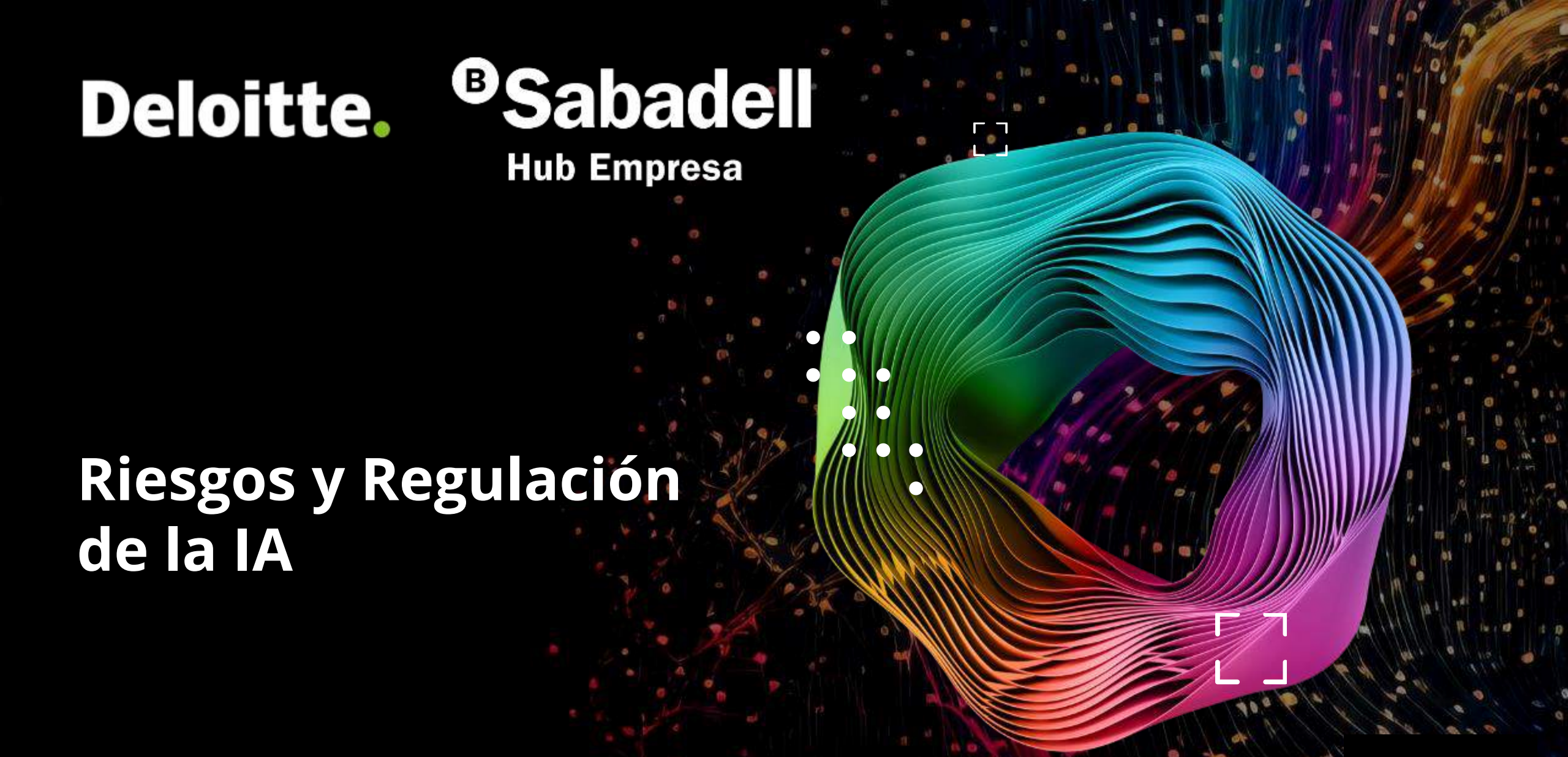
CONCLUSIONES

- Las técnicas de explicabilidad han ayudado a las áreas de negocio a hacer **más comprensibles las variables de entrada** de los modelos y a crear variables auxiliares que aportan **poder predictivo** a los algoritmos
- El caso de uso presentado se considera como un gran **éxito**, dando pie a nuevos proyectos en los que se seguirán aplicando estas metodologías éticas
- La **confianza** que negocio va depositando en este tipo de algoritmos permite ir llevando a cabo una sustitución progresiva de reglas sencillas por modelos de *Machine Learning* mucho más eficientes y adaptables

La **explicabilidad** es un área importante dentro del campo de la Inteligencia Artificial, como así lo recoge la CE. El uso de estas técnicas es cada día más relevante y **debería ser un paso fundamental** en cualquier proyecto de ciencia de datos que persiga avanzar desde modelos tradicionales hacia ML

Deloitte.

Sabadell Hub Empresa



Riesgos y Regulación de la IA

Julio, 2024

ÍNDICE DE CONTENIDOS

Deloitte.

01. Riesgos y Marco Regulatorio

Contexto regulatorio actual

La UE está decidida en liderar la próxima revolución tecnológica diseñando el entorno del futuro de la economía digital, **los datos y la inteligencia artificial de confianza pueden ayudar a encontrar soluciones y aportar muchos beneficios para la sociedad**, desde la atención sanitaria, pasando por el transporte y automoción y la fabricación limpia y sostenible.

AI

RGPD

DATA

Directrices de la CE

Directrices éticas para una IA confiable

08/04/2019

Artículo de la CE

Una estrategia europea para los datos

19/02/2020

Guía de la AEPD

Adecuación al RGPD de tratamientos que incorporan Inteligencia Artificial

13/02/2020

Informe de la CE

Implicaciones en materia de seguridad y responsabilidad de la IA, el IoT y la robótica

19/02/2020

Artículo de la CE

Libro blanco de la inteligencia artificial

19/02/2020

Estudio del PE

Inteligencia Artificial y aplicación de la Ley

13/07/2020

Estudio del PE

Inteligencia Artificial y responsabilidad civil (Aspectos Legales)

13/07/2020

Estudio del PE

El impacto del RGPD en la inteligencia artificial

15/07/2020

Guía de la AEPD

Requisitos para Auditorías de Tratamientos que incluyan IA

Enero de 2021

Propuesta regulación Comisión Europea

Un enfoque europeo para la inteligencia artificial (AI Act)

21/04/2021

Opinión del EBF

Sobre la propuesta de la CE de reglamento sobre la IA (AI Act)

27/09/2021

Opinión del ECB

Sobre la propuesta de la CE de reglamento sobre la IA (AI Act)

29/12/2021

Versión del Consejo Europeo

Un enfoque europeo para la inteligencia artificial (AI Act)

06/12/2022

Propuesta de regulación CE

Data Act

23/02/2022

Fin Trílogos

Acuerdo para la ley de inteligencia artificial (AI Act)

09/12/2023

Versión del Parlamento Europeo

Un enfoque europeo para la inteligencia artificial (AI Act)

07/06/2023

Artificial Intelligence Act

Texto final aprobado

13/03/2024

2019 2020 2021 2022 2023 2024

Mensajes principales

El 21 de Abril de 2021, la Comisión Europea propuso el primer marco jurídico sobre Inteligencia Artificial (IA) de la historia, que aborda los riesgos de la IA y sitúa a la Unión Europea en una posición de liderazgo mundial.

A lo largo de los últimos años esta propuesta ha ido evolucionando. A continuación, se ofrece una visión general de la última actualización:

¿En qué se centra?

En el ser **humano**, con un enfoque basado en **riesgos** y que **clasifica** los **sistemas de IA** en varios grupos.

Foco en la **seguridad, salud y derechos fundamentales.**

¿A quién aplica?

Proveedores, responsables del despliegue, importadores y distribuidores de sistemas de IA dentro de la UE.

¿Cuándo se aplicará?

2 años desde la publicación en el Boletín Oficial de la UE en términos generales, pero con ciertos matices:

- ***Prohibidos*** → *6 meses*
- ***GPAIS*** → *12 meses*
- ***Regulación general*** → *24 meses*
- ***HRAIS Anexo III*** → *24 meses*
- ***HRAIS Anexo I*** → *36 meses*

¿Qué margen tengo?

Para poder utilizar un sistema de IA de alto riesgo será necesario registrar un número de **certificado** en una base de datos pública, que habrá emitido un **Organismo Noticado** tras la evaluación completa de la **Evaluación de Conformidad** o de la **autoevaluación.**

Régimen sancionador

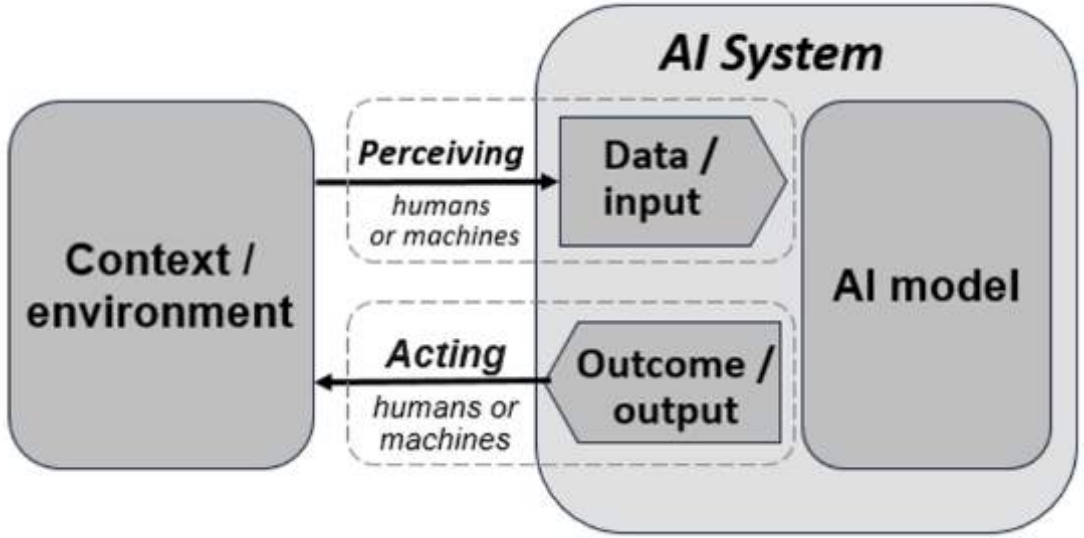
El **incumplimiento** puede **acarrear tasas** de hasta 35.000.000 euros o el 7% de la facturación global anual

Definición general de la IA

¿Qué sistemas tiene una organización que el AI Act considera como Inteligencia Artificial?

Sistema de IA

Un sistema **basado en una máquina** diseñado para funcionar con **distintos niveles de autonomía**, que puede mostrar **capacidad de adaptación** tras el despliegue y que, para **objetivos explícitos o implícitos**, **infiere** de la información de entrada que recibe la manera de **generar información de salida**, como predicciones, contenidos, recomendaciones o decisiones, que puede influir en entornos físicos o virtuales



Clasificación de los sistemas de IA según su riesgo

Sistemas de IA de Riesgo Inaceptable (Art. 5)

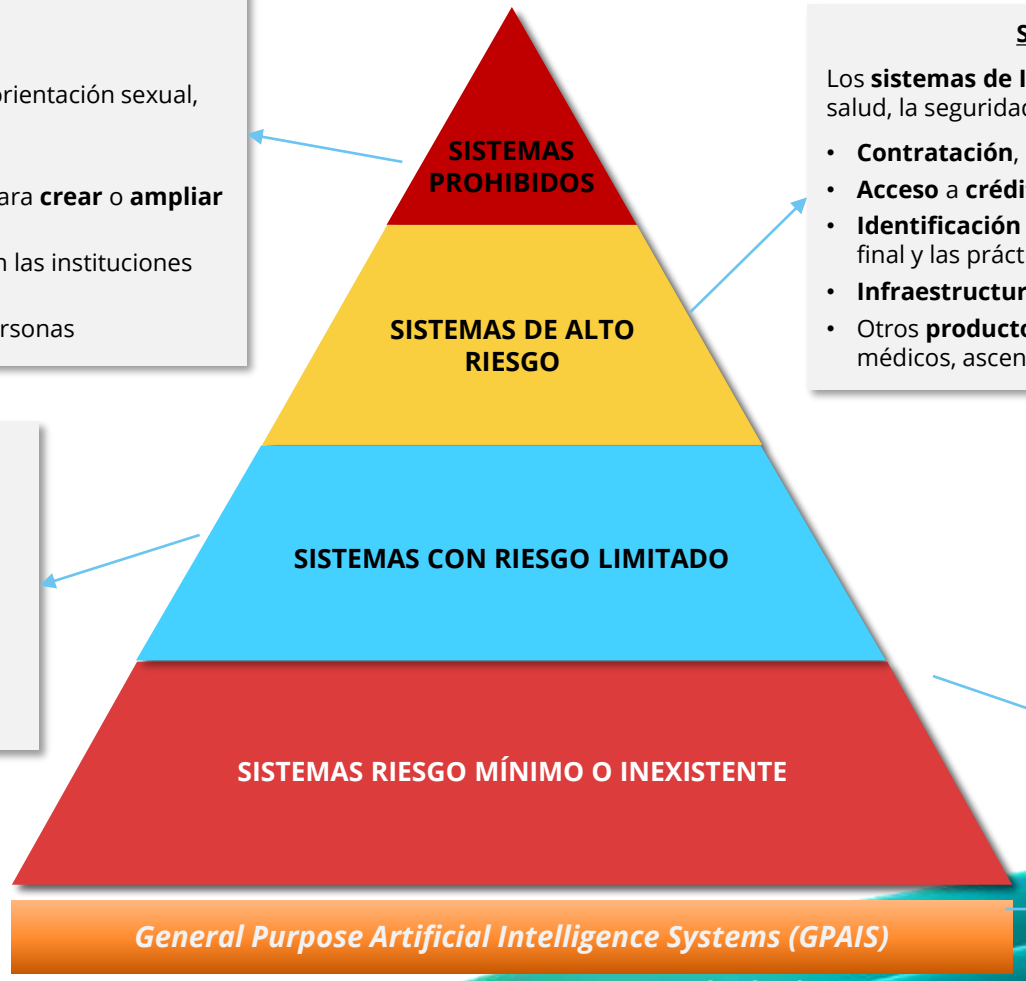
Se **prohíbe** el uso de **sistemas de IA como:**

- **Manipulación** cognitivo-conductual
- **Explotación** de **vulnerabilidades** de personas
- **Categorización biométrica** para inferir **datos sensibles** (orientación sexual, creencias, etc.)
- **Puntuación** y/o **clasificación social**
- **Extracción** no selectiva de **imágenes** faciales de Internet para **crear** o **ampliar bases de datos** de reconocimiento facial
- **Reconocimiento** de **emociones** en el lugar de **trabajo** y en las instituciones **educativas**
- Algunos casos de **vigilancia policial predictiva** para las personas

Sistemas de IA con obligaciones de transparencia específicas (Art.50)

Permitidos pero sujetos a una serie de **obligaciones** de **transparencia,** sistemas relacionados con:

- La interacción con humanos
- La generación de contenidos manipulados
- **Control mitigante:** Revelar que el contenido ha sido generado por IA



Sistemas de IA de Alto Riesgo (HRAIS, Art.6)

Los **sistemas de IA** que pueden conducir a un **riesgo significativo** para la salud, la seguridad o los derechos fundamentales, por ejemplo:

- **Contratación**, **promoción, evaluación**
- **Acceso** a **créditos** y **seguros de vida y salud**
- **Identificación biométrica** (a excepción de la mera identificación del usuario final y las prácticas prohibidas)
- **Infraestructuras críticas** según Directiva UE 2022/2557
- Otros **productos** ya regulados por **normas armonizadas UE** (dispositivos médicos, ascensores, vehículo autónomo, etc.)

Sistemas de IA de Riesgo Inexistente o Mínimo

Permitidos sin restricciones (ejemplo: mantenimiento predictivo)

Sistemas sin un propósito definido, no se puede categorizar en función del nivel de riesgo de la finalidad prevista, pero se establecen obligaciones específicas

Requisitos para los sistemas de IA de alto riesgo

Leyenda

- Gestión
- Buen uso por parte del proveedor y resp. desp.
- Técnicos

Evaluación de impacto sobre los derechos fundamentales (FRIA) antes de la introducción en el mercado de un HRAIS

Datos y gobernanza de datos

- Conjuntos de datos con buena calidad y gobierno del dato
- Los datos deben ser relevantes, representativos, libres de error y completos
- Se deben aplicar técnicas de Gobierno del Dato apropiadas

Registros

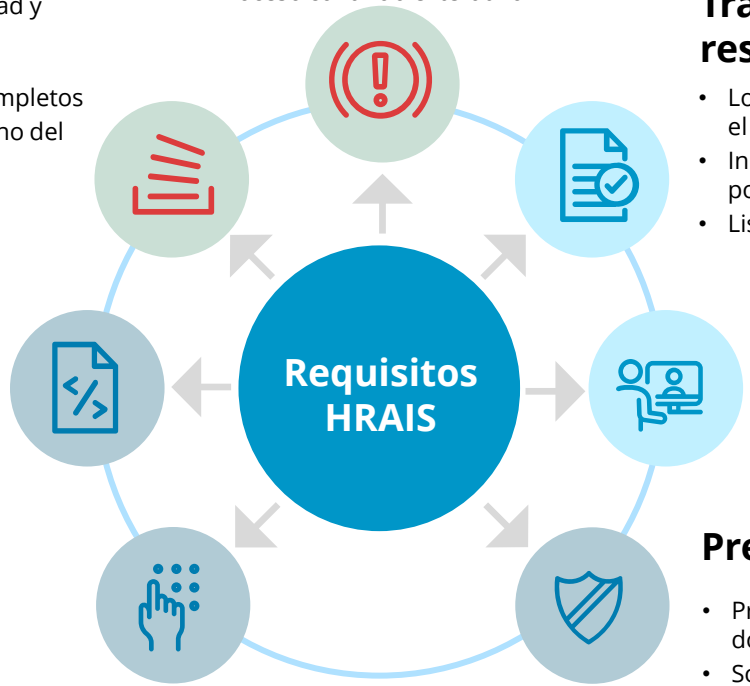
- Se guardan suficientes logs
- Debe ser mantenida documentación técnica suficiente para poder verificar la conformidad con la Regulación
- Explicabilidad

Documentación técnica

- Documentación de todos los requisitos
- Actualización continua
- Antes de la comercialización

Sistema de gestión de riesgos

- Foco en la gestión de los riesgos que puedan afectar a la salud, seguridad o derechos fundamentales de las personas
- Proceso continuo e iterativo



Transparencia y comunicación a los responsables del despliegue

- Los responsables del despliegue deben poder entender y controlar el HRAIS
- Incluye información concisa, clara, no técnica, accesible y entendible por el implementador
- Lista de lo que debe incluir la documentación

Vigilancia humana

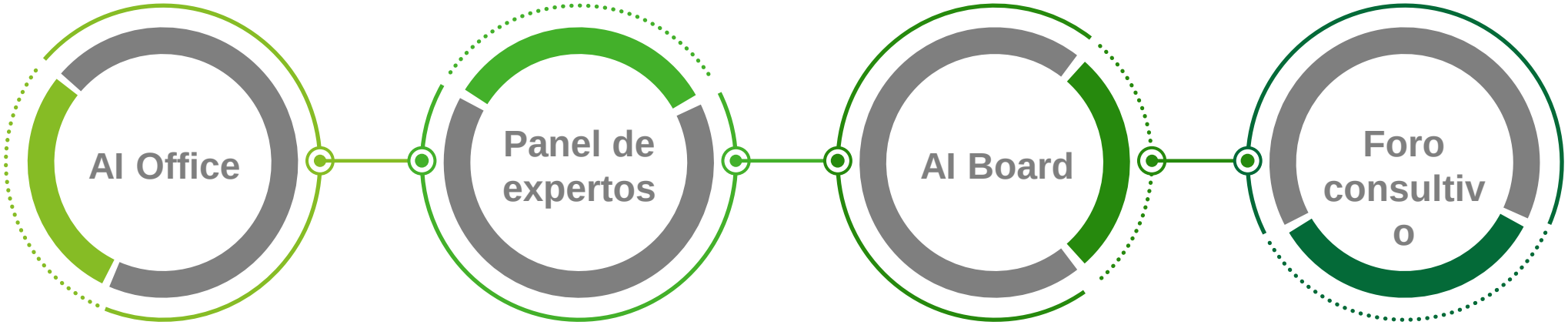
- Límite en la delegación
- Se pueda decidir no usar el HRAIS o sus outputs
- Posibilidad de que el **humano** pueda i) de manera segura e instantánea **interrumpir** la operación y ii) descartar, corregir o revertir el output

Precisión, solidez y ciberseguridad

- Precisión: el resultado es preciso y se conoce por la documentación el nivel de precisión que tiene
- Solidez: resiliencia a fallos, errores o inconsistencias.
- Ciberseguridad: resiliencia a ataques

Obligaciones específicas para proveedores GPAIS y GPAIS con riesgo sistémico

Gobernanza a nivel europeo



- Supervisión de modelos de IA más avanzados.
- Fomentar normas y prácticas de prueba.
- Hacer cumplir las normas comunes en todos los Estados miembros.

- Asesoramiento a la Oficina de IA sobre los modelos GPAI.
- Desarrollo de metodologías para evaluar las capacidades de los modelos fundacionales.
- Designación de modelos fundacionales de alto impacto
- Monitorización de los posibles riesgos de seguridad relacionados con los modelos fundacionales.

- Compuesto por representantes de los Estados miembros.
- Plataforma de coordinación y un órgano consultivo de la Comisión
- Otorgará un papel importante a los Estados miembros en la aplicación del Reglamento, incluido el diseño de códigos de buenas prácticas para los modelos fundacionales

- Foro consultivo para las partes interesadas, con el fin de proporcionar conocimientos técnicos al Consejo de IA.

Estados Miembro, Autoridades Nacionales Competentes y Organismos Notificados

Estado Miembro

- Papel clave en la **aplicación** y el **cumplimiento** del reglamento
- **Designa** a las autoridades nacionales competentes (**ANCs**)

La Autoridad de VIGILANCIA es la entidad más importante.
Ejemplos:
- AESIA
- BCE/BdE
- AEMPS

Autoridades Nacionales Competentes (ANC)

Autoridad de Vigilancia del Mercado (AVM)

- Supervisa actividades del mercado
- Informa a las autoridades nacionales en caso de incumplimiento de las obligaciones
- Realiza actividades y toma medidas en virtud del Reglamento (UE) 2019/1020

Autoridad Notificante (AN)

- Proporciona y ejecuta los procesos de evaluación, designación y notificación de los organismos de evaluación de la conformidad y su supervisión

Su labor es más puntual y administrativa.
Ejemplos:
- DG Industria
- DG Transporte
- DG Consumo
- Ministerio Sanidad
- Ministerio Economía

Los **Organismos de Evaluación de Conformidad** solicitan la notificación y se convierten en organismos notificados

Organismo Notificado (ON)*

- Realiza la evaluación de conformidad, las pruebas, la certificación y la inspección
- Coopera con las autoridades nacionales competentes
- Un Organismo de Evaluación de Conformidad puede ser ON en varios sectores

Es la entidad que hace la EVALUACIÓN los sistemas antes de ponerlo en mercado.
Ejemplos:
- Privado: AENOR, Applus, etc.
- Público: CNCPS

**En los casos en los que se prevea una auto evaluación de conformidad no será necesario un Organismo Notificado*

Deloitte.

02. Iniciativas nacionales e internacionales

Iniciativas de la Secretaría de Estado de Digitalización e Inteligencia Artificial

La **SEDIA** presenta en **Bruselas** en **Junio del 2022** su propuesta de **Sandbox Regulatorio**, un entorno contralado donde se espera que diferentes empresas y organizaciones **pongan a prueba los requisitos regulatorios**.

Desde **Marzo de 2022**, y hasta **diciembre de 2025**, la **SEDIA** está desarrollando las siguientes **iniciativas** en torno al **Sandbox regulatorio**:

Diseño y desarrollo de las bases para la elaboración del sello de la IA y la ejecución del Sandbox Regulatorio

Principales ámbitos de actuación

AESIA

Definición de **procedimientos de autoridad de supervisión** y **herramienta web de auditoría**.

SELLO

Mecanismo para **fomentar la confianza** de la sociedad española en los **sistemas de IA que no son de alto riesgo**.

SANDBOX

Piloto de aplicación del Reglamento IA Europeo a sistemas reales de Alto Riesgo, con participantes de la empresa y las AAPP:

- Optimiza guías para facilitar cumplir el Reglamento.
- Prepara operativa de seguimiento y asesoría
- Produce conclusiones sobre desafíos y oportunidades

INVESTIGACIÓN

Estudio de **Sellos de IA a nivel internacional**, **futura normativa europea** y **coordinación con organismos de normalización**.

DIVULGACIÓN

Divulgación del nuevo sello (portal web y actividades) y **formación específica por cada sector**.

Elaboración de las guías que ayudarán a las empresas a entender los requisitos del AI Act y establecen las medidas para cumplir con éstos

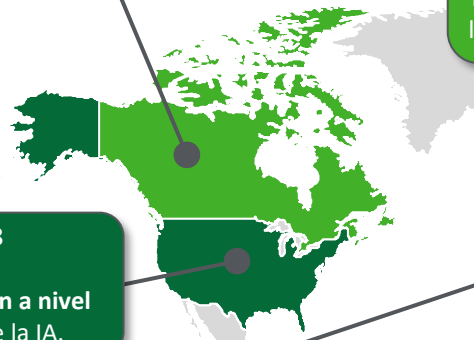
Guías técnicas

- **Artículo 09 –** Sistema de gestión de riesgos
- **Artículo 10 –** Datos y gobernanza de datos
- **Artículo 11 –** Documentación técnica
- **Artículo 12 –** Conservación de registros
- **Artículo 13 –** Transparencia y provisión de información a los responsables del despliegue
- **Artículo 14 –** Vigilancia humana
- **Artículo 15.I –** Precisión
- **Artículo 15.II –** Solidez
- **Artículo 15.III –** Ciberseguridad
- **Artículo 17 –** Sistema de gestión de calidad
- **Artículo 19 –** Evaluación de la conformidad
- **Artículo 61 –** Seguimiento posterior a la comercialización

Stakeholders relevantes

- Secretaría de Estado de Digitalización e Inteligencia Artificial (SEDIA)
- Asociación Española de Normalización (UNE)
- Agencia Española de Protección de Datos (AEPD)
- Joint Research Centre (JRC)
- DG CONNECT (Comisión Europea)

Iniciativas regulatorias a nivel internacional



Canadá ha propuesto el Artificial Intelligence and Data Act (AIDA) **un acta de evaluación de la IA** y sus riesgos.

Reino Unido implementará un **Framework** para aportar **claridad** y **coherencia** al panorama **regulador** de la **IA** garantizando la **innovación responsable**. Tiene como **meta** fortalecer la **posición** de **UK** como líder mundial en IA

Corea del Sur, no hay una regulación integral de la IA. El Gobierno está trabajando para proponer una Ley de IA

China ha establecido una regulación basada en **24 guidelines**, pionera en la incorporación de especificidades para IA Generativa.

Estados Unidos, Orden Ejecutiva **8 principios y prioridades**, complementándose con **regulación a nivel estatal** para la gestión y control de la IA.

Israel, Proyecto de política de regulación y ética en IA

UE, Framework regulatorio de la IA enfocado en los riesgos para la salud, seguridad y derechos fundamentales de las personas. Enfoque basado en el riesgo para casos de uso de IA

Brasil, Proyecto de ley sobre IA que hace hincapié en los derechos humanos y crea un régimen de responsabilidad civil para los desarrolladores de IA

India, No hay una regulación integral de la IA. Propuesta de Ley de IA en India para regular los sistemas de alto riesgo.

Australia, No hay una regulación integral, el gobierno hace hincapié en la aplicación de los marcos existentes

Sin propuestas definidas

Marcos de referencia voluntarios

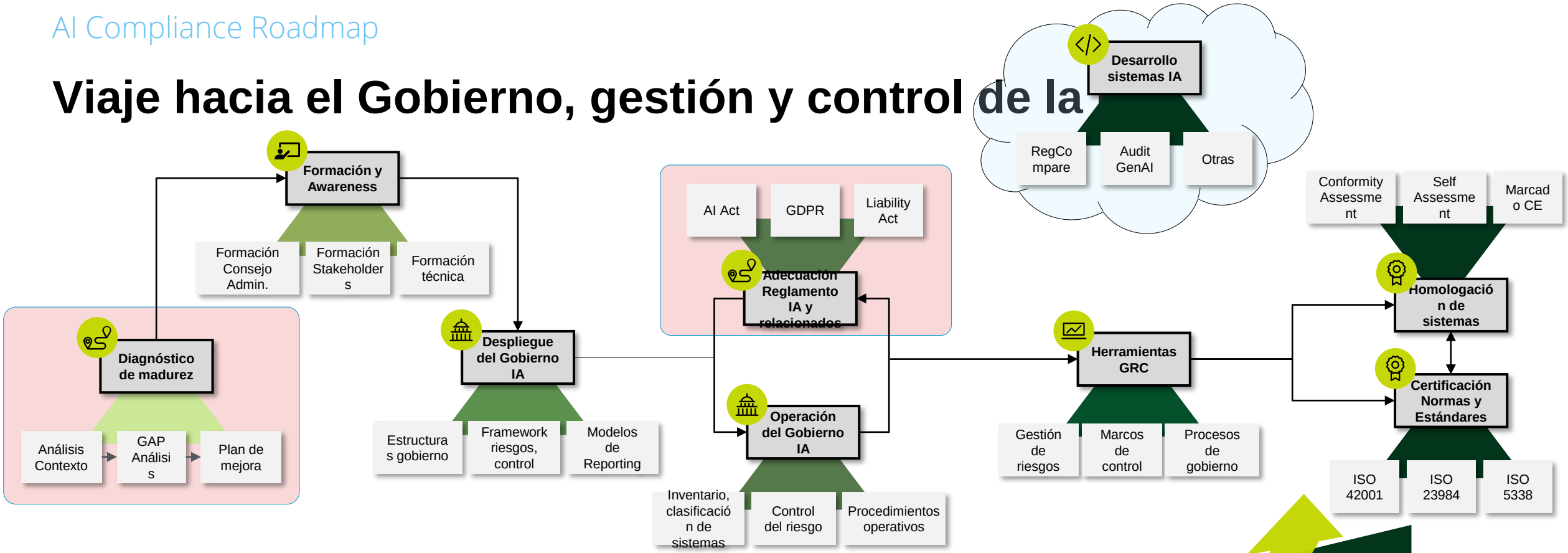
Proyectos integrales para regular IA

Regulación en vigor

Deloitte.

03. AI Compliance Roadmap

Viaje hacia el Gobierno, gestión y control de la



"Viaje de una organización hacia el Gobierno, gestión y control de la IA"

Escala de progreso

Gobierno y control de la IA

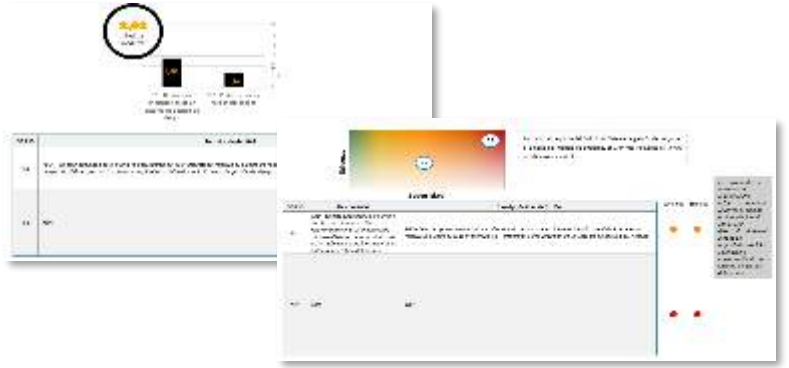
Metodología de análisis de madurez de adaptación al AI Act

Análisis de madurez y cumplimiento con los requisitos del Reglamento Europeo de IA

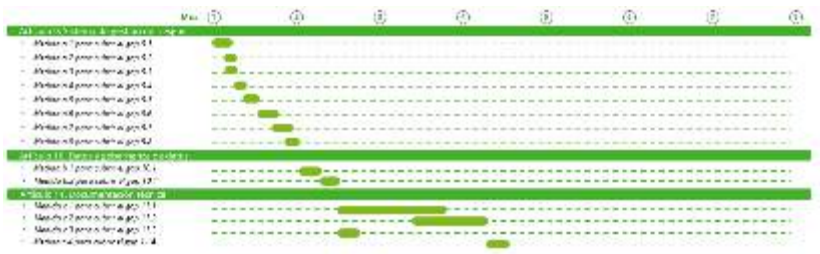
1. Análisis de madurez

2. Identificación de "gaps"

3. Definición de medidas

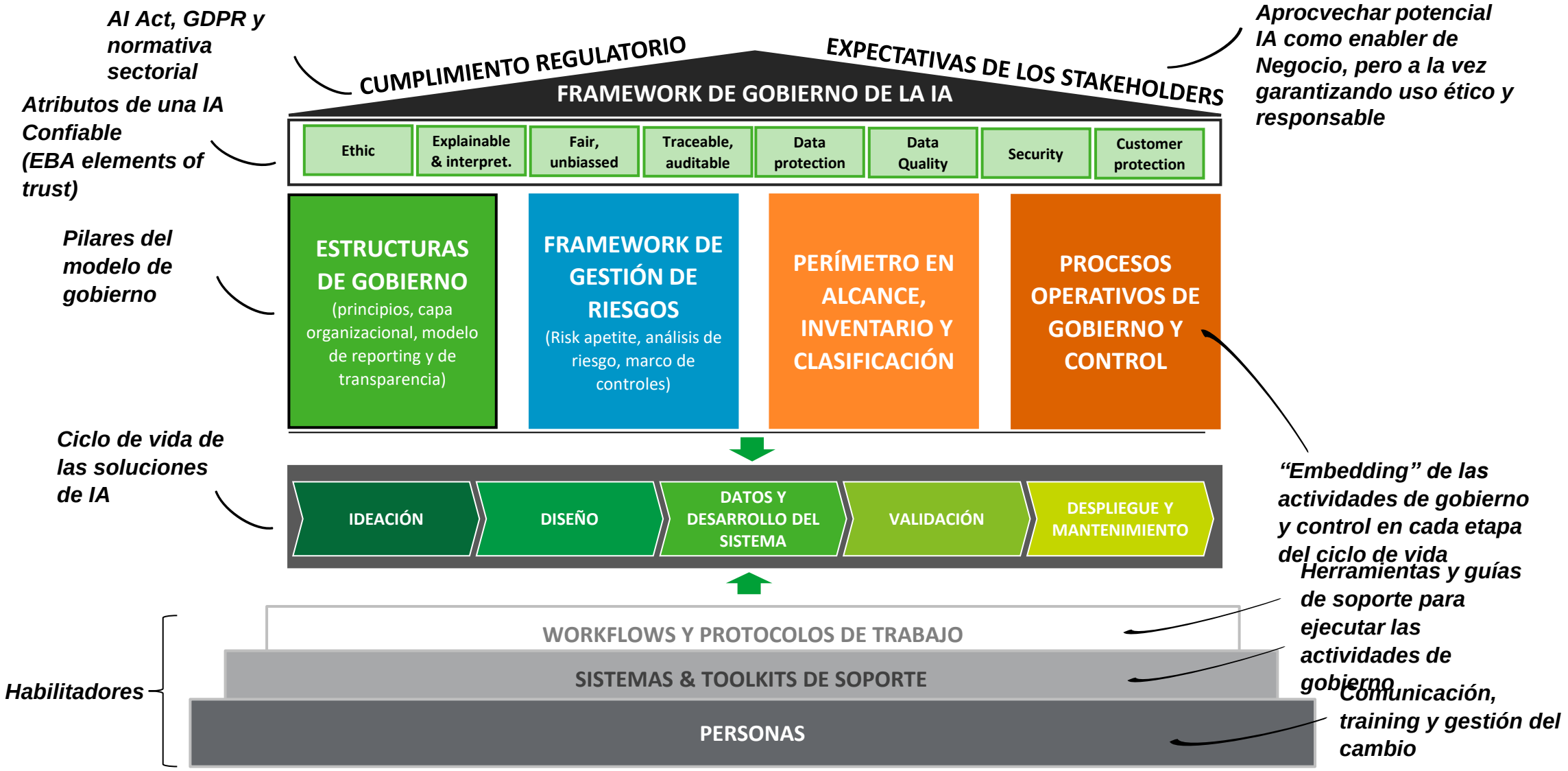


4. Plan de implementación



Gobierno y control de la IA

Modelo de Gobierno de la IA: Visión general



Deloitte.