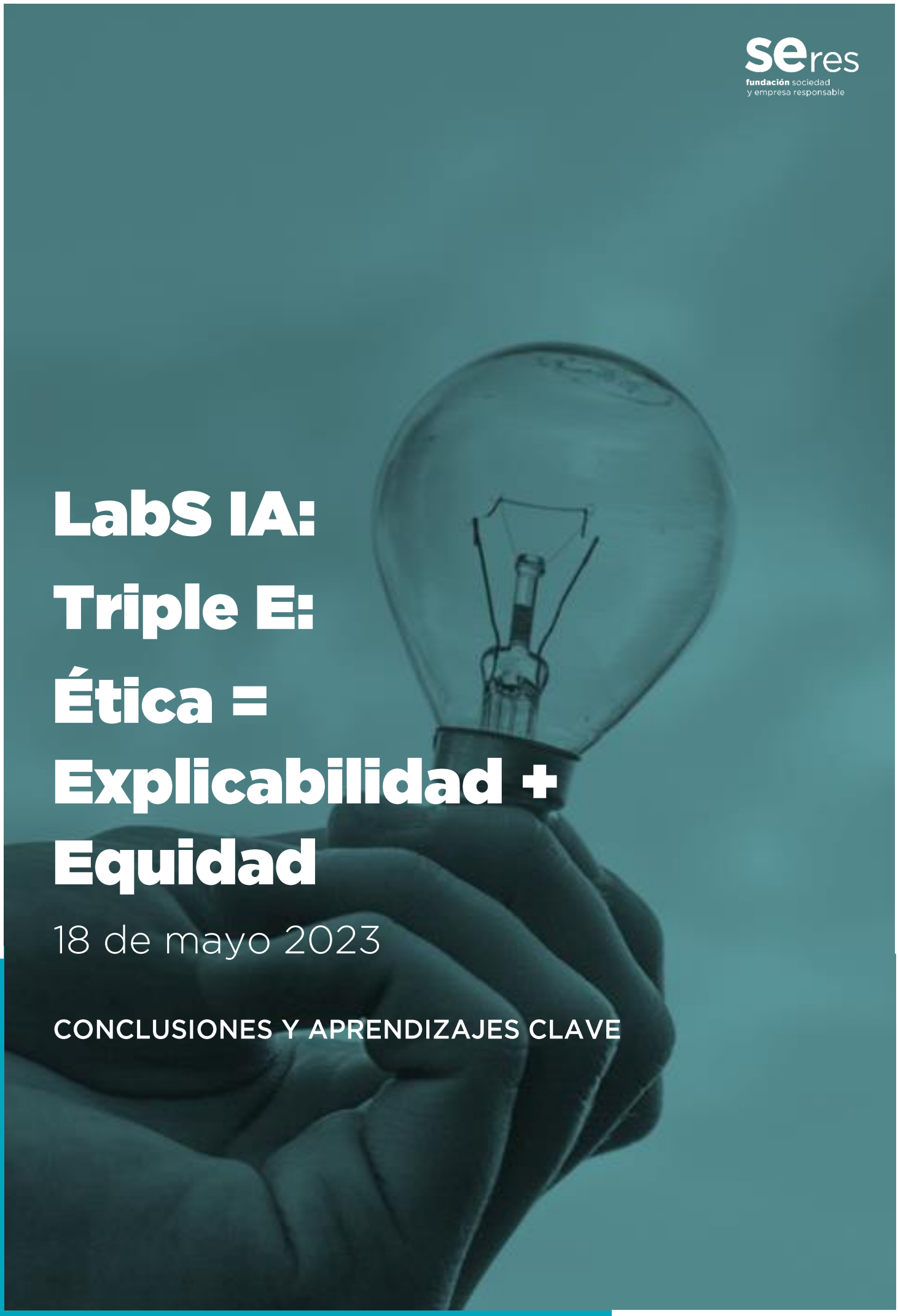

seres
fundación sociedad y empresa responsable

# LabS IA: Triple E: Ética = Explicabilidad + Equidad

18 de mayo 2023

## CONCLUSIONES Y APRENDIZAJES CLAVE

# AGENDA

| | |
|---|---|
| **10:00 – 10:05** | **Bienvenida y contexto**<br>Paula de la Cal, Project Manager en Fundación SERES |
| **10:05 – 10:10** | **Introducción y marco regulatorio**<br>Rose Barragán, AI Advisory Consultant en NTT Data |
| **10:10 – 10:30** | **Explicabilidad: Comprender el proceso de toma de decisiones de los algoritmos**<br>Alicia de Manuel, AI Expert Analyst en NTT Data |
| **10:30 – 10:50** | **Equidad: Garantizar resultados justos para todas las personas**<br>Alicia de Manuel, AI Expert Analyst en NTT Data |
| **10:50 – 11:00** | **Hacia una organización enfocada en explicabilidad y equidad**<br>Marc Sanguesa, D&I CoE Manager en NTT Data |
| **11:00 – 11:20** | **Caso real: Intervención de Banc Sabadell**<br>Raquel Pérez, Analytics Vertical Leader & Responsable de IA Confiable en Banc Sabadell |
| **11:20 – 11:30** | **Q&A** |

# RESUMEN DE LA SESIÓN

El desarrollo de modelos de IA generativa cada vez más sofitiscados, ha provocado un debate acerca de las implicaciones que esta tecnología tendrá para el ser humano en los próximos años a nivel económico, productivo y social.

La gestión de esta transformación digital implica, entre otras cosas, hacer frente a un reto ético y regulatorio: ¿Qué principios queremos que rijan las relaciones entre las personas y las máquinas? y ¿Con qué normas vamos a proteger dichos principios?

La Unión Europea está impulsando diferentes propuestas regulatorias con el objetivo de lograr una IA que sea robusta, confiable y que promueva los derechos de las personas.

Entre ellas, destacan el **Libro blanco de la IA (2020), el AI Act (2021) o el Data Act (2022)**, que pretenden facilitar una serie de directrices a las organizaciones acerca de cómo emplear la IA.

La **explicabilidad** y la **equidad** son dos paradigmas recurrentes en estos documentos.  En la sesión, profundizamos en cada uno de ellos y sus implicaciones para las compañías.

**La explicabilidad** se refiere a comprender las decisiones de los modelos de IA, garantizando que son seguros, imparciales y respetuosos con la privacidad.

En ocasiones, la complejidad de algunos algoritmos hace que sea imposible descrifrar cómo han tomado sus decisiones (*black box*). Dicha complejidad se relaciona con la explicabilidad y con la precisión o capacidad para hacer predicciones exactas del algoritmo.

En función de estas dos variables, los modelos pueden clasificarse en **interpretables** (sencillos de comprender, pero menos exactos en sus resultados) y **modelos explicables** (más precisos, pero requieren técnicas específicas para su comprensión).

Es importante escoger el modelo más adecuado para cada caso según nuestros objetivos y crear **metodologías de auditoría** que permitan conocer las decisiones del sistema. Contar con procesos de documentación es clave para facilitar la trazabilidad y auditabilidad.

**La equidad** busca abordar los sesgos y errores de los modelos de IA para mejorar la calidad de los datos y garantizar que las decisiones que toman son justas.

A lo largo de su ciclo de vida, los algoritmos  pueden generar sesgos que hacen que tomen decisiones discriminatorias hacia ciertos grupos de personas. Estos sesgos pueden clasificarse en función de la fase en la que se generan:

**Fase 1: Oportunidad de negocio o conceptualización del modelo**

- Sesgo de confirmación: Seleccionar únicamente datos que respalden la oportunidad de negocio, mientras se ignoran aquellos que la contradicen.
- Desalineamiento de la iniciativa: Distanciamiento del objetivo original con el que se crea el modelo, que podría dar lugar a resultados no éticos.

**Fase 2: Adquisición y descubrimiento de los datos que entrenarán el modelo**

- Sesgo de calidad de los datos: Uso de bases de datos desactualizadas o poco precisas.
- Sesgo de muestreo: Los datos utilizados no son representativos de la población.
- Sesgo histórico: Los datos reflejan desigualdades históricas.

**Fase 3: Creación y desarrollo del modelo**

- Sesgo del atributo: Uso de atributos irrelevantes para el fin del algoritmo.
- Sesgo del programador: Reproducción de los sesgos inconscientes del equipo que programa.

**Fase 4: Evaluación y despliegue del modelo**

- Sesgo de retroalimentación: El modelo se retroalimenta de manera sesgada.
- Sesgo de popularidad: Se favorecen unos resultados sobre otros por ser más populares.

**Fase 5: Monitorización del modelo**

- 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

Establecer medidas de mitigación en cada una de las fases del modelo de vida y garantizar que los datos que se utilizan tienen en cuenta multitud de expresiones culturales es clave para mitigar sesgos.

Para ello, NTT DATA ha creado el ***CDO Journey*** para aterrizar los principios de explicabilidad y equidad a través de un enfoque *End-to-End* en las organizaciones; el ***AI Audit Tool*** para la evaluacion y auditoria de los modelos de IA; el ***AI Risk Control Management Model*** para el diseño de modelos estandarizados de supervisión y control de sesgos y riesgos, y **programas de formación ad-hoc** para empleados en ética de la IA.

# LECCIONES APRENDIDAS

**Explicabilidad:**

- Entender cómo funciona un algoritmo nos permite confiar en él. La explicabilidad mejora la propuesta de valor de las empresas al aumentar la confianza de los clientes, dar mejor respuesta a los requisitos regulatorios, aumentar la calidad del servicio, mejorar la transparencia de los procesos y ayudar a producir nuevos modelos.
- Contar con metodologías de auditoría y programas de formación ad-hoc que permitan explicar las decisiones del sistema es clave para garantizar la explicabilidad.
- Los procesos de documentación dentro de los proyectos de IA son clave para la auditabilidad y la trazabilidad.
- Las personas tienen derecho a saber cuándo están interactuando con un sistema de IA y deben ser informados sobre sus capacidades y limitaciones.

**Equidad:**

- Los datos de los algoritmos deben tener en cuenta distintas expresiones culturales y sociales para garantizar la inclusión y reflejar la diversidad de los individuos y grupos sociales.
- Los procedimientos de prevención y mitigación de sesgos deben establecerse en todas las etapas del ciclo de vida.
- Es importante formar a los equipos en materia de diversidad e inclusión y tener en cuenta a distintos stakeholders para prevenir y mitigar la inclusión de sesgos.

**Aplicación en la empresa:**

- El cambio cultural  y los procesos de concienciación son fundamentales para priorizar el uso ético de la IA en toda la organización.
- El uso de técnicas de explicabilidad y equidad debería ser un paso fundamental en cualquier proyecto de ciencia de datos que persiga avanzar desde modelos tradicionales hacia técnicas de IA como el Machine Learning.
- Las técnicas de explicabilidad ayudan a las áreas de negocio a hacer más comprensibles las variables de entrada de los modelos y a crear nuevas variables que aportan poder predictivo a los algoritmos.
- La confianza que el negocio deposita en los algoritmos permite una sustitución progresiva de reglas sencillas por modelos más complejos de IA más eficientes y adaptables. Por ejemplo, el uso de modelos de IA avanzados para calcular préstamos o identificar fraudes fiscales.

# PRÓXIMOS PASOS

En los próximos meses, publicaremos un informe elaborado a partir del contenido trabajado en la sesión, que será presentado en la próxima edición del LabS.

# EMPRESAS PARTICIPANTES