

Investigación en Inteligencia Artificial

Dr. Pablo Moreno Ger

Dr. Ismael Sagredo Olivenza

Dr. Luis Miguel Garay Gallastegui

Dr. Ricardo Alonso

Tema 12 – Implicaciones filosóficas, éticas y legales en la aplicación de la inteligencia artificial

De qué vamos a hablar...

- ▶ Repaso de situación
- ▶ Tema 12 (la semana pasada)
 - Introducción y objetivos
 - Contexto legal aplicable a proyectos de inteligencia artificial
- ▶ Tema 12 (HOY)
 - Sesgos
 - Seguridad
 - Explicabilidad algoritmos

Tema 12

El problema de que un algoritmo decida

Fuentes de los sesgos

► Sesgos cognitivos

- Atajos mentales que evolutivamente han sido útiles para tomar decisiones de forma más rápida
- Estos errores sistemáticos en los procesos cognitivos (pensamiento, percepción, memoria, entre otros) nos producen una desviación en el procesamiento mental y nos pueden alejar de la racionalidad o nublar nuestro juicio
- Son más habituales de lo que creemos o reconocemos











<https://www.titlemax.com/discovery-center/lifestyle/50-cognitive-biases-to-be-aware-of-so-you-can-be-the-very-best-version-of-you/>

Fuentes de los sesgos

► Sesgos cognitivos: ejemplos



50 SESGOS COGNITIVOS A TENER EN CUENTA PARA SER LA MEJOR VERSIÓN DE TI

Memoria	Social	Aprendizaje	Creencia	Dinero	Política
Sesgo de correspondencia Juzgamos a los demás por su personalidad, pero a nosotros mismos por la situación  María llega tarde a clase; es perezosa. Yo llevo tarde; tengo un día malo	Sesgo por interés personal Nuestros fracasos son circunstanciales, pero nuestros éxitos son nuestra responsabilidad  Ganas un premio por tu trabajo duro, no por suerte. Pero si fallas es porque no has dormido suficiente	Favoritismo del endogrupo Cuando favorecemos a las personas que están en nuestro grupo en lugar de las que están fuera  Francis está en tu iglesia, así que te gusta más Francis que María	Efecto arrastre Es la tendencia a hacer o creer en algo porque muchas personas lo hacen o lo creen  María cree que los spinners ayudan a sus hijos. Francis también lo cree.	Pensamiento de grupo El deseo de armonía en el grupo lleva a tomar decisiones irracionales  María quiere helado y Francis comprar camisetas. Entonces tu sugieres comprar camisetas con fotos de helados	
Efecto halo Si ves a una persona con un rasgo positivo, esa impresión positiva se extenderá a otros rasgos  «María nunca podría ser mala; ¡es tan guapa!»	Suerte moral Tendencia a atribuir una mayor o menor posición moral basada en el resultado de un evento.  La cultura X ganó la guerra X porque eran moralmente superiores a los perdedores	Efecto del falso consenso Cuando creemos que hay más gente que está de acuerdo con nosotros de lo que realmente es el caso  ¡Todo el mundo piensa eso!	Maldición del conocimiento Una vez que sabemos algo, asumimos que todos los demás también lo saben  Alicia es una profesora y lucha por entender la perspectiva de sus nuevos estudiantes	Efecto Spotlight Cuando pensamos que un acto o elemento propio resulta muy llamativo y todo el mundo va a verlo  María está preocupada de que todos se den cuenta de lo patética que es su camiseta de helado	

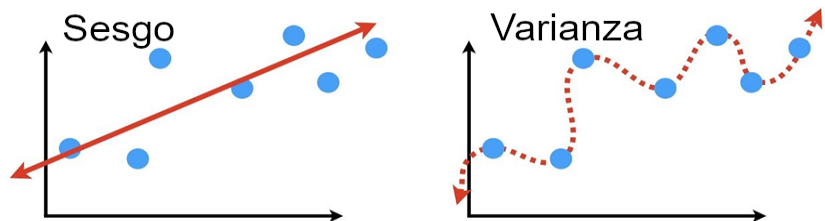
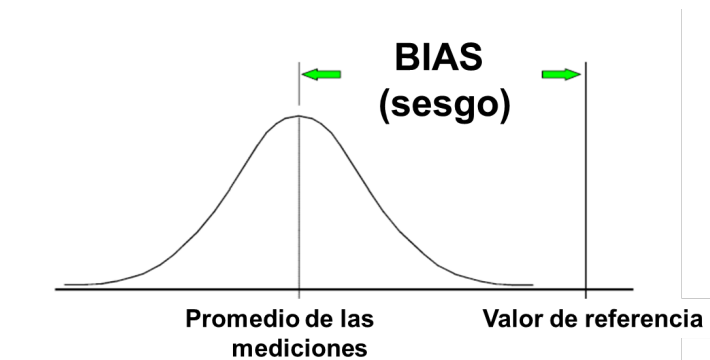
El problema del sesgo

► Los algoritmos usan sesgos para aprender

- Y por tanto magnifican fenómenos pequeños si tienen características únicas.
- Hay que guiar el aprendizaje para evitarlos.

► Sesgo

- Recordando el Tema 9: Aprendizaje Automático
- Sesgo es característica innata de la propia simplificación



Building trust in machine learning and AI

If we understand how discrimination and bias enter algorithmic decision making, we have an opportunity to eliminate them

They can find patterns in data that elude us, patterns that might reveal important relationships that improve the accuracy of the algorithm. They can recover patterns and relationships that we as human beings want to ignore. But they can just as easily fail to discover important relationships and produce bad recommendations, even dangerous ones.

A well-known example of the latter involved research to see whether machine learning could guide the treatment of pneumonia patients. The team was trying to predict the risk of complications in pneumonia patients where low-risk patients could receive outpatient treatment. A rule-based machine learning system decided that pneumonia patients who also had asthma could be sent home—because they experienced few complications from pneumonia. However, the reason patients with asthma and pneumonia experienced few complications was because they received intensive care at the hospital. The important connection between patient condition and quality of care was not reflected by the machine learning algorithm.

<https://www.infoworld.com/article/3255948/machine-learning/building-trust-in-machine-learning-and-ai.html>

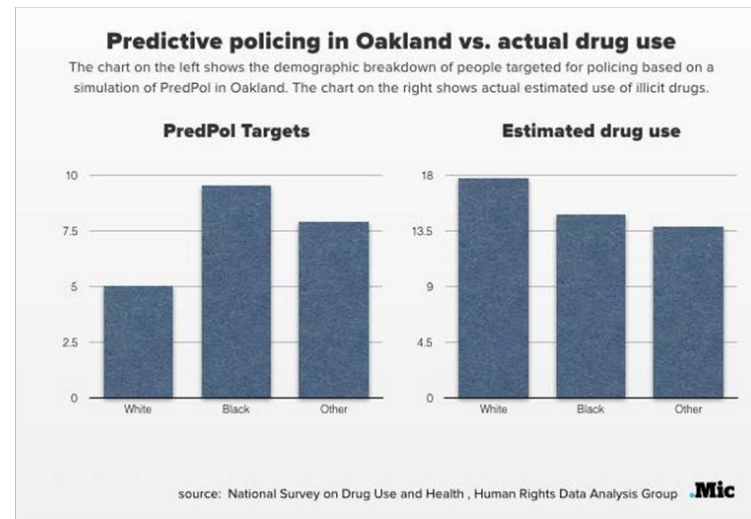
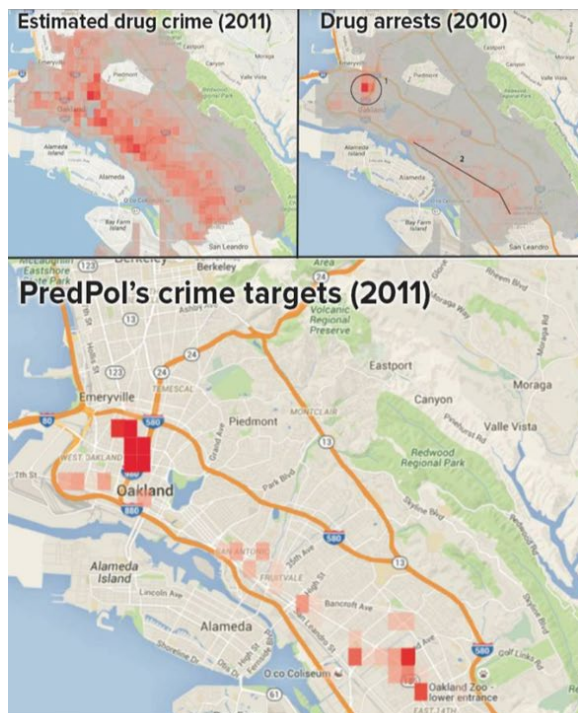
Fuentes de los sesgos

- ▶ Algoritmos basados en bases de datos con sesgos
 - Reconocimiento de caras sesgado hacia internet (caucásicos)
 - Datos de criminalidad sesgados por condición social
 - Datos médicos sesgados hacia hospitales privados o públicos
 - El mundo digital sesgado hacia el primer mundo

Fuentes de los sesgos

► Algoritmos basados en bases de datos con sesgos

**(Exclusive) Crime-prediction tool
PredPol amplifies racially biased
policing, study shows**



IA vs. legislación

Bank Accused Of Racist Lending Practices Settles Suit With New York State

Discrimination by algorithm: scientists
devise test to detect AI bias

Researchers devise test to determine whether machine learning algorithms are
introducing gender or racial biases into decision-making

Companies must act

How to comply

New Compliance Rules

Regulations take effect

Law Changes

How rules affect you



Code is the new law

- ▶ Si todos los coches son autónomos...
 - ¿Qué hacemos con el código de circulación?
- ▶ Si todos los traders de bolsa son agentes
 - ¿Podemos regular el mercado?
- ▶ Si un algoritmo comete un delito
 - ¿De quién es la responsabilidad?

Code is the new law

- On Wall Street today, more than 60% of all trades are executed by AI with little or no real time oversight from humans



IA vs. sociedad

No todo es binario

- ▶ Si cada vez usamos más algoritmos para resolver problemas, porque lo hacen mejor...
 - Automáticamente las decisiones que toman esos modelos son interpretadas como correctas.
 - Pero ... ¿Lo son ?

Ejercicio de discusión

- ▶ ¿Si una IA que selecciona candidatos para un puesto de trabajo siempre hombres, implica que los hombres son mejores para ese puesto de trabajo?
 - ¿Existen trabajos para los que en promedio los hombres sean objetivamente mejores?
- ▶ ¿Y si eligiese solamente mujeres?
 - ¿Existen trabajos para los que en promedio las mujeres sean objetivamente mejores?
- ▶ ¿Y si eligiese solamente personas de color?
 - ¿Existen trabajos para los que en promedio las personas de color sean objetivamente mejores?

Ejercicio de discusión

- ▶ ¿Si una IA que selecciona candidatos para un puesto de trabajo siempre hombres, implica que los hombres son mejores para ese puesto de trabajo?
 - ¿Existen trabajos para los que en promedio los hombres sean objetivamente mejores?
- ▶ ¿Y si eligiese solamente mujeres?
 - ¿Existen trabajos para los que en promedio las mujeres sean objetivamente mejores?
- ▶ ¿Y si eligiese solamente personas de color?
 - ¿Existen trabajos para los que en promedio las personas de color sean objetivamente mejores?
- ▶ ¿Y si eligiese solamente personas con muy altas capacidades matemáticas?

Ejercicio de discusión

► Repetimos la pregunta

- ¿Si una IA que selecciona candidatos para un puesto de trabajo siempre hombres, implica que los hombres son mejores para ese puesto de trabajo?

Pero incluso si lo fuesen...

¿es ése el mundo que queremos construir?



Ejercicio de discusión

► Otras reflexiones interesantes

- Los candidatos no estándar pueden tener grandes fortalezas ocultas (contra viento y marea...)
- Incluso si un perfil específico *fuese* objetivamente mejor, ¿queremos vivir en ese tipo de sociedad?
- ¿Es posible modelar factores emocionales en una IA? (e.g. inclusividad, dar oportunidades, tener fe, confiar en el instinto...)

Discusiones adicionales

¿Qué otras zonas grises se os ocurren?

Discusiones adicionales

- ▶ Las decisiones del vehículo autónomo
- ▶ Contenidos de ocio seleccionados por IA
- ▶ IA vs. Educación
- ▶ IA vs. Guerra (*detachment*)
- ▶ La IA como poder legislador y judicial

IA vs. seguridad

Vulnerabilidades de la IA

- ▶ Los algoritmos de IA se pueden “atacar”
 - Técnicas de engaño a redes neuronales
 - Inducción de sesgos
 - Disipación en forma de ruido

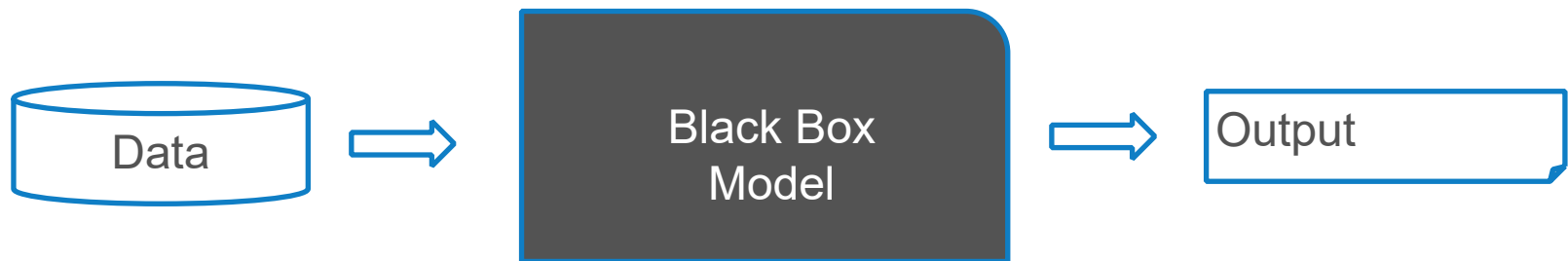
Explicabilidad

Explicabilidad

- ▶ Si queremos supervisión... necesitamos explicabilidad.
- ▶ La explicabilidad como respuesta a muchas preguntas
 - ¿Por qué me recomiendan que compre un producto?
 - ¿Por qué no me conceden un préstamo?
 - ¿Por qué recomendamos un tratamiento a un paciente?
 - ¿Por qué este alumno necesita examinarse?
 - ¿Por qué frenó bruscamente el coche?

Explicabilidad

- ▶ Modelos de caja negra. Veo entradas y salidas.
 - Deep Learning es el caso más característico.
 - Razonamiento basado en casos: El caso es muy parecido... pero ¿Puedo deducir el modelo subyacente?
 - Árboles de decisión bien. Random Forest, no tanto.
 - Y encima son los que están funcionando mejor...



Explicabilidad. ¿Por qué la queremos?

► Confiabilidad.

- ¿Pondría mi vida en manos de un agente que toma decisiones que no me puede explicar?

► Adquirir nuevo conocimiento

- Si el modelo infiere nuevo conocimiento, nosotros queremos saber cómo ha llegado a esa conclusión.

► Detección de fallos

- Si el modelo tiene fallos y conocemos el modelo, podremos predecirlos.

Explicabilidad

► Ejemplos

- Los diseñadores de videojuegos temen usar machine learning para modelar el comportamiento de los NPCs porque tienen miedo a comportamientos emergentes.
- En medicina se siguen usando sistemas expertos porque se basan en el conocimiento de los expertos y generan diagnósticos explicables.

► ¿Donde importa menos?

- Algoritmo de Youtube.
- Recomendaciones de Netflix
- Algoritmos de escalado de imágenes por Deep Learning. ¿Por que este pixel es verde y no rojo...?

Es un aspecto que preocupa y en el que
ya hay frameworks de trabajo

Trustworthy AI EU

► Ethics Guideline

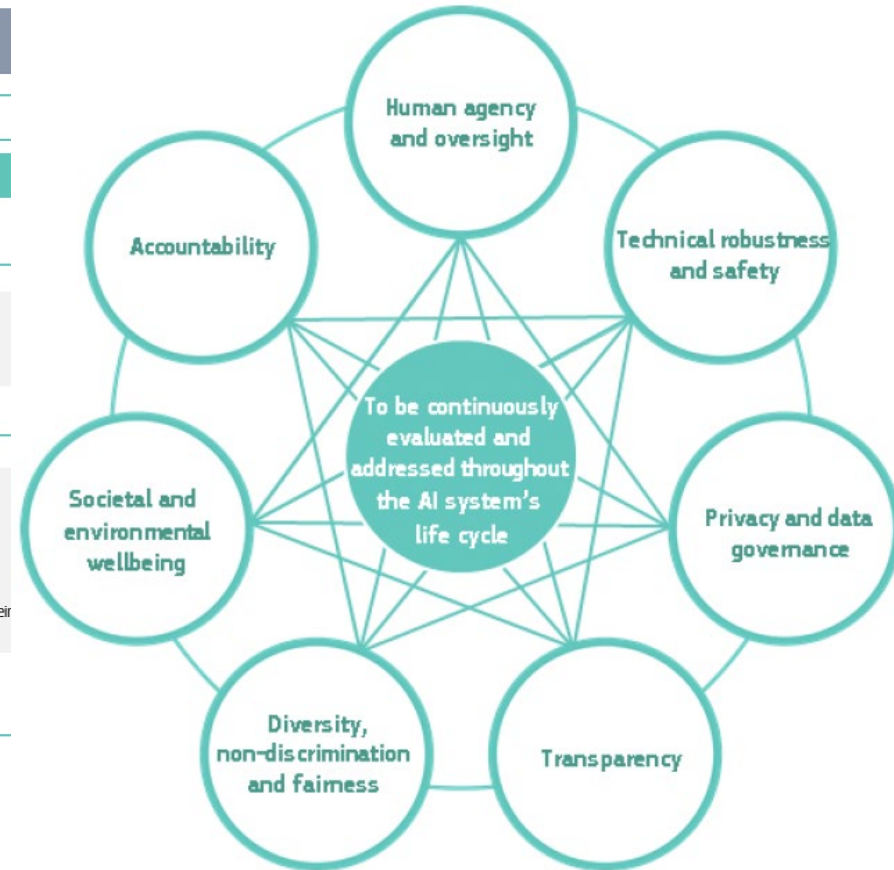
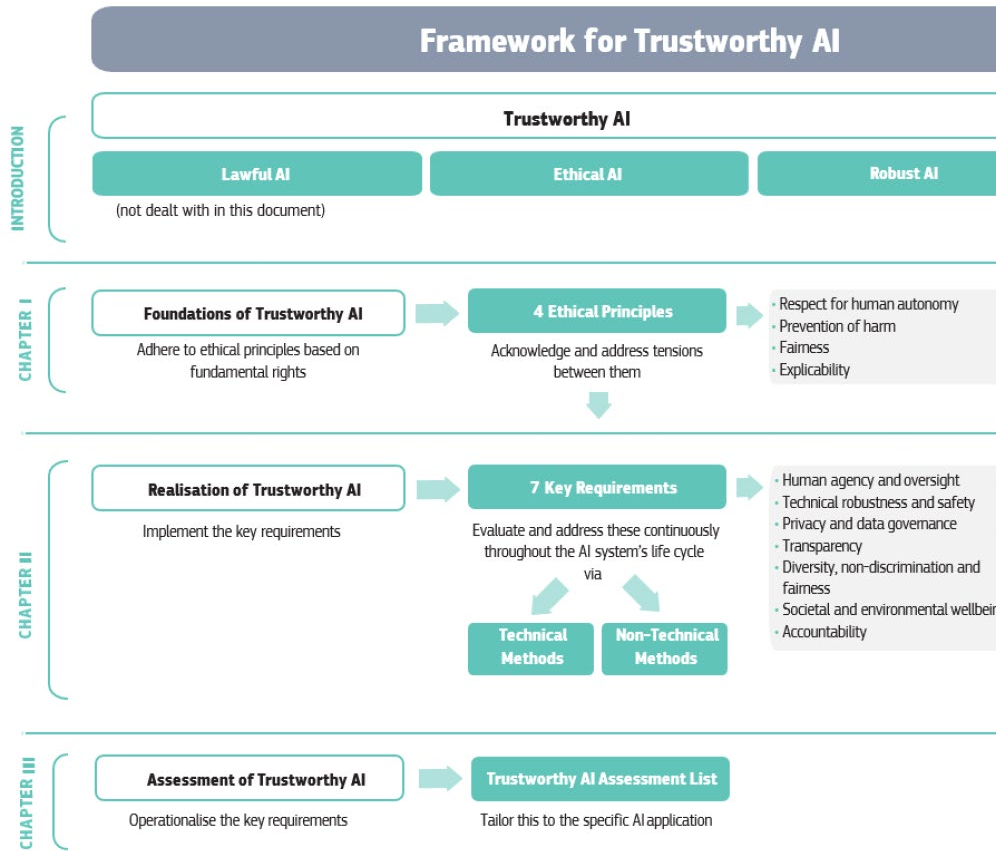


Figure 1. Ethics Guideline for Trustworthy AI EU

Y hasta aquí llegamos...

Porque todo lo bueno se acaba

¿Preguntas?



www.unir.net