

Universidad de Lleida; Massive Data Processing

DATA AND AI ETHICS

12 de febrero de 2024



The logo for Bluetab, an IBM Company. It features the word "/bluetab" in a bold, sans-serif font. The "b" is blue, the "l" is purple, and the rest of the letters are red. Below this, the text "an IBM Company" is written in a smaller, black, sans-serif font.

ÍNDICE

1. ? ¿Por qué es necesario hablar de esto?
2. 🦸 Ética
3. 📚 Data ethics
4. 📰 La IA para informar (y desinformar)
5. ⚖️ Data bias y sus efectos
6. 🔍 Feedback loop
7. 🤔 Futuro
8. 📲 GDPR

¿POR QUÉ...

... es necesario hablar de esto ?



- Vivimos en un “mundo de datos”
- Los datos y qué se hace con ellos puede afectar a miles de personas
- Los humanos se alejan del análisis, y dejamos que las máquinas de encarguen del trabajo



ÉTICA

Definición

- La rama de la filosofía que estudia la conducta humana, lo correcto y lo incorrecto, lo bueno y lo malo, la moral... [wikipedia]
- Conjunto de normas morales que rigen la conducta de la persona en cualquier ámbito de la vida [RAE]
- Ramas:
 - / **Metaética:** estudia las teorías éticas en sí mismas
 - / **Ética normativa** o deontología: establece principios para guiar los sistemas de normas y deberes en ámbitos de interés común.
 - / **Ética aplicada:** analiza la aplicación de las normas éticas y morales a situaciones concretas.



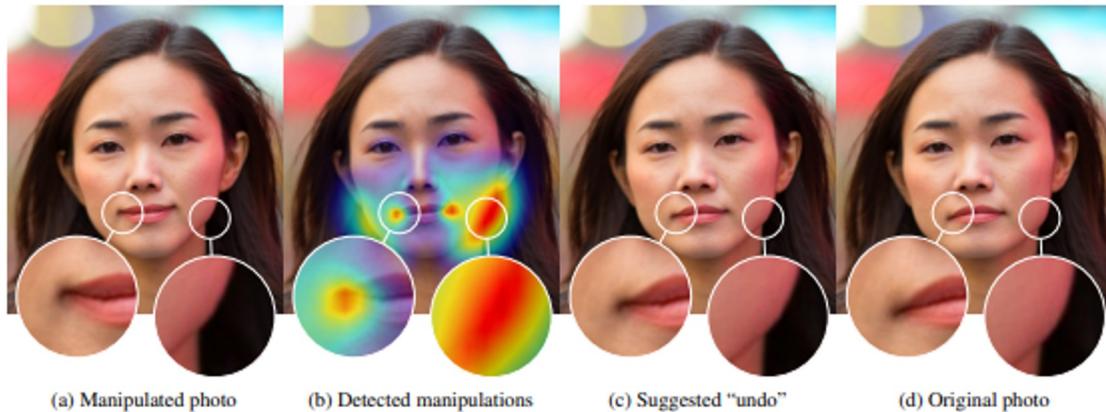
Definición

- Nueva rama de la ética
- Estudia y evalúa los problemas morales relacionados con los datos, los algoritmos y las correspondientes prácticas.
- Objetivo: formular y apoyar soluciones moralmente buenas



La IA para informar

- La IA puede ayudarnos a detectar noticias falsas, imágenes falsas o información errónea
 - / BotSentinel (<https://botsentinel.com/>)
 - / Detectar plagio (<https://www.markcopy.ai/>)
 - / Detectar imágenes retocadas (FALdetector:
<https://arxiv.org/pdf/1906.05856.pdf>)





La IA para informar (y desinformar)

- La IA puede ayudarnos a detectar noticias falsas, imágenes falsas o información errónea, pero también puede usarse para generar información falsa.

- / Crear tuits falsos (<https://github.com/adityagaydhani/deeptweets>)
- / FSGAN (faceswap)
- / VoiceSwap (<https://www.youtube.com/watch?v=u>



DeepTweets @lexfridman

Fake AI-generated tweets & tweet rewrites that mimic the style of real Twitter accounts.

Kanye West ● @kanyewest tweet bot
America is built on love. Love is the most powerful force in the universe.

Donald J. Trump ● @realDonaldTrump tweet bot
America is the land of vision, not facts.

Andrew Yang ● @AndrewYang tweet bot
America is stronger than Twitter is.

Elon Musk ● @elonmusk tweet bot
America is the land of short term risk and long term magic.

Jordan Peterson ● @jordanbpeterson tweet bot
America is built on diversity. Diversity of outcome.

Dwayne Johnson ● @TheRock tweet bot
America is my home. You come in, you don't stop. #WheresTheCoffee?



Data bias y sus efectos

- Sesgo histórico
- Sesgo de representación
- Sesgo de medición
- Sesgo de agregación
- Sesgo de evaluación
- Sesgo de implementación



Data bias y sus efectos

Historical Bias

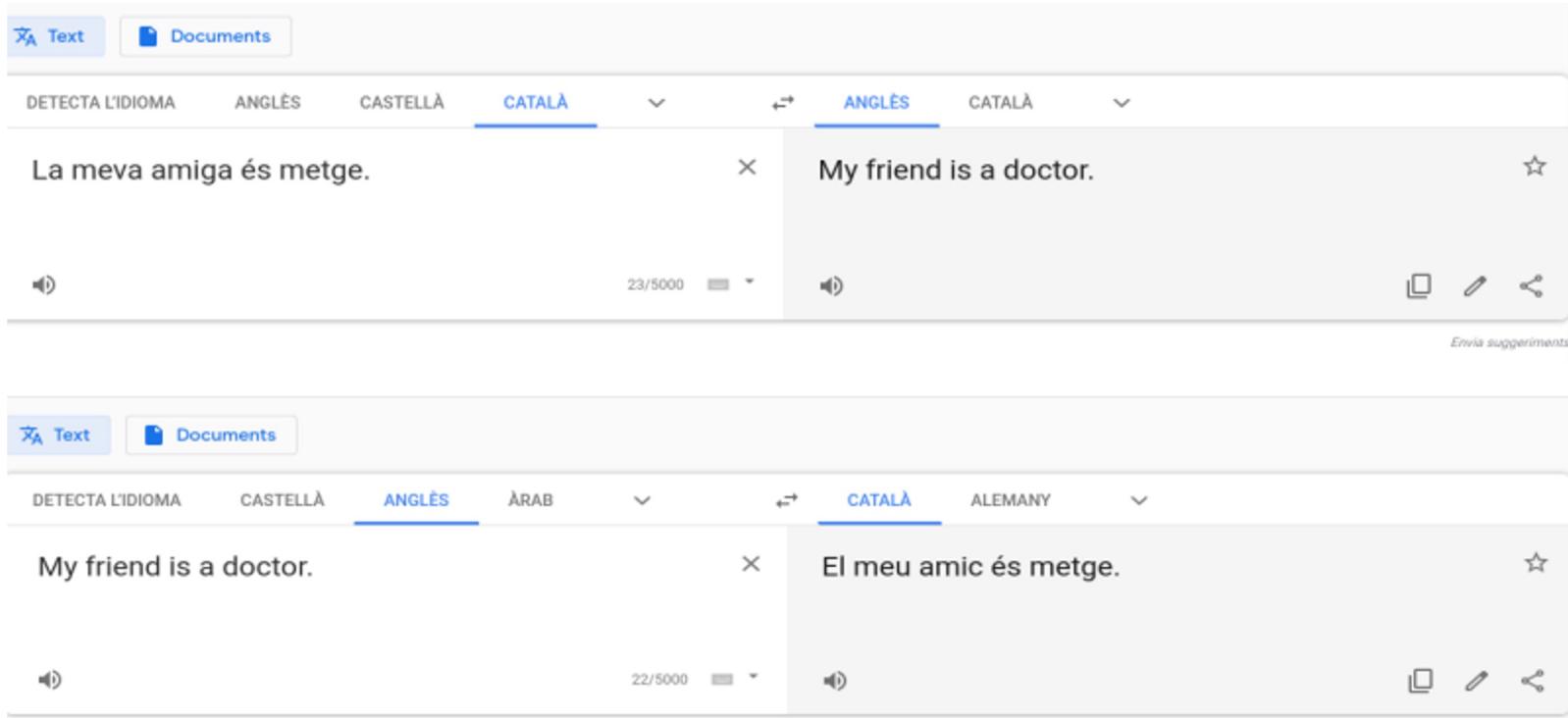
- **COMPAS**: algoritmo utilizado para ayudar a decidir sentencias y fianzas en USA. Este algoritmo se entrenó con datos de USA con datos de a partir del siglo XX.

	WHITE	AFRICAN AMERICAN
Labeled Higher Risk, But Didn't Re-Offend	23.5%	44.9%
Labeled Lower Risk, Yet Did Re-Offend	47.7%	28.0%

Data bias y sus efectos

Gender bias (representación)

- **Google Translate:** todos lo conocemos. La verdad, es que en estos últimos dos años se han puesto las pilas, y este caso en concreto lo han arreglado



The image shows two separate instances of the Google Translate interface side-by-side.

Top Translation: The source text is "La meva amiga és metge." and the target text is "My friend is a doctor." The source language is set to CATALÀ and the target language is set to ANGLÈS. The translation is accurate and gender-neutral.

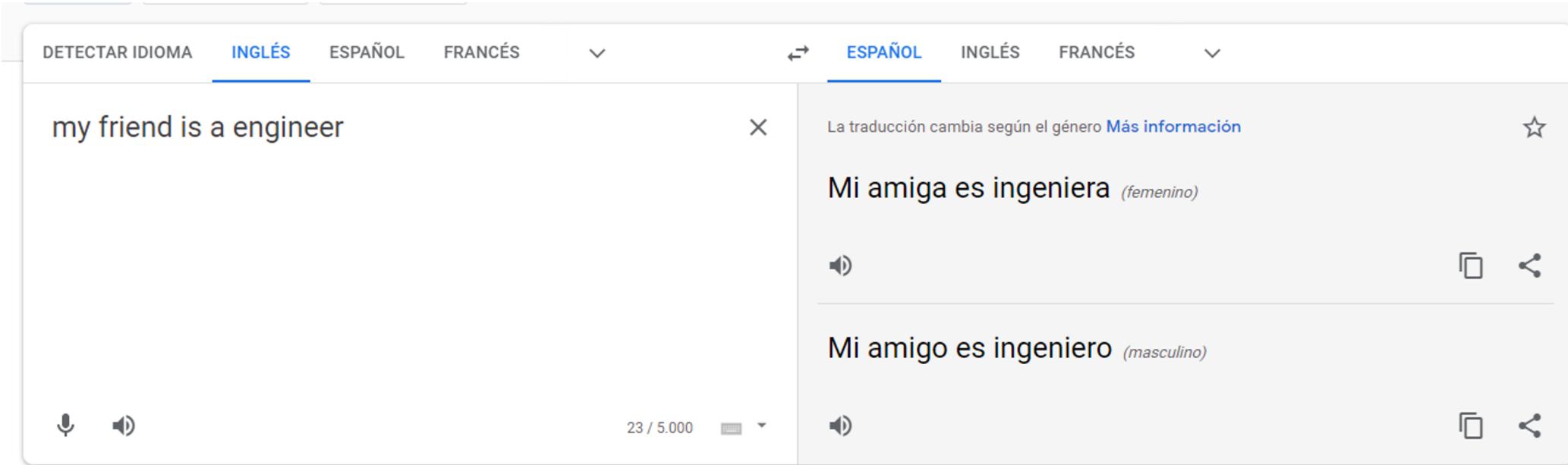
Bottom Translation: The source text is "My friend is a doctor." and the target text is "El meu amic és metge." The source language is set to ANGLÈS and the target language is set to CATALÀ. The translation is inaccurate, reflecting a gender bias by translating "my friend" as "El meu amic" (male friend) instead of "la meva amiga" (female friend).



Data bias y sus efectos

Gender bias (representación)

- **Google Translate:** todos lo conocemos. La verdad, es que en estos últimos dos años se han puesto las pilas, y este caso en concreto lo han arreglado

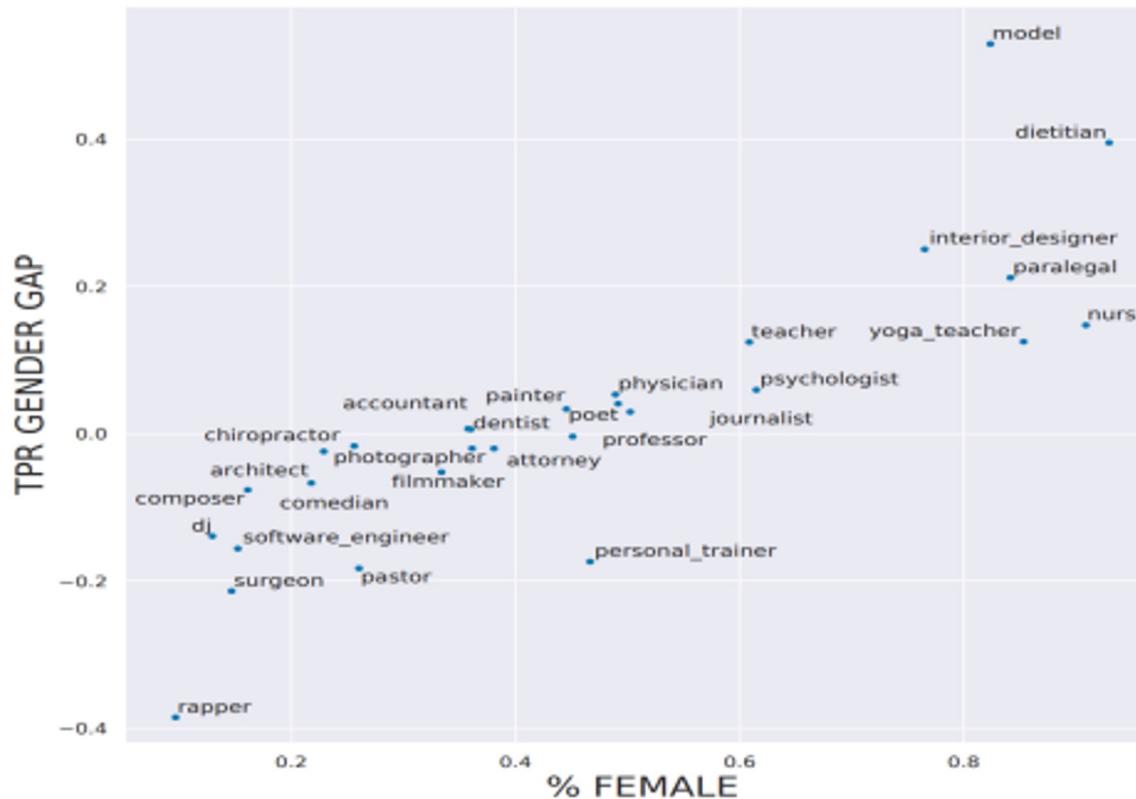


The screenshot shows the Google Translate interface. On the left, the input text "my friend is a engineer" is entered. The source language is set to "DETECTOR IDIOMA" (detection), and the target language is "INGLÉS". On the right, the output is shown in "ESPAÑOL". A note above the translations states: "La traducción cambia según el género [Más información](#)". Two options are provided: "Mi amiga es ingeniera" (feminine) and "Mi amigo es ingeniero" (masculine). Each option has a speaker icon for audio playback and a share icon.

Data bias y sus efectos

Gender bias (representación)

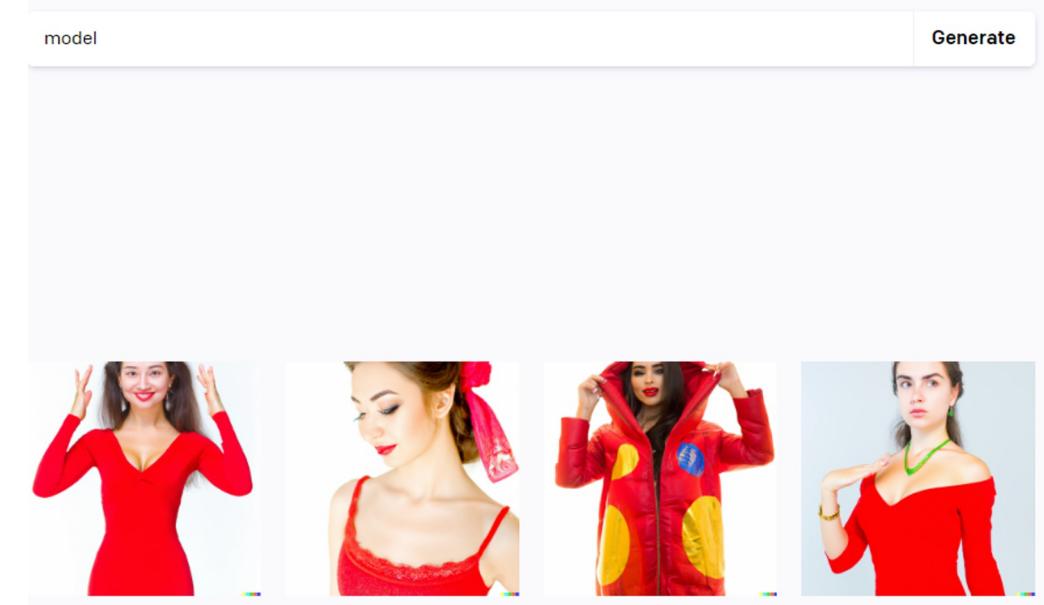
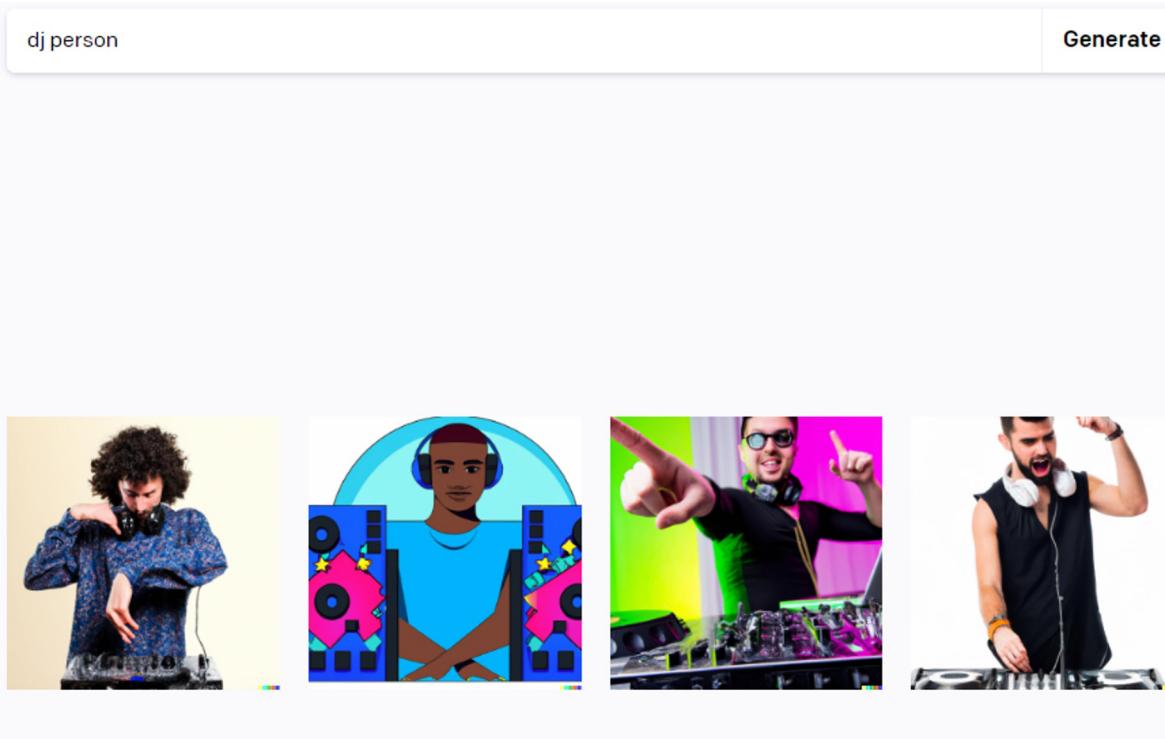
- **Google Translate:** la explicación es, que cuando el sesgo está en los datos de entrada, se magnifica en los modelos de Inteligencia Artificial



Data bias y sus efectos

Gender bias (representación)

- DALLE: este mismo ejemplo, también aplica a generadores de imágenes como DALLE



Data bias y sus efectos

Gender bias (representación)

- Amazon: la aplicación de reclutamiento de Amazon

 Reuters

Amazon scraps secret AI recruiting tool that showed bias against women

Automation has been key to Amazon's e-commerce dominance, be it inside warehouses or driving pricing decisions. The company's experimental...

10 oct 2018



Data bias y sus efectos

Race bias (representación)

- Apple: y su sistema de desbloqueo en iPhone X



The screenshot shows a news article from the New York Post. At the top, there's a red header bar with a search icon, an envelope icon, and the "NEW YORK POST" logo. Below the header, a red banner has the word "TECH" in white. The main title of the article is "Chinese users claim iPhone X face recognition can't tell them apart". Below the title, it says "By Guy Birchall, Tom Michael, The Sun" and "December 21, 2017 | 3:11pm | Updated". To the right of the title, there are social media sharing icons for Facebook, Twitter, Flipboard, Email, and LinkedIn.



Data bias y sus efectos

Race bias (representación)

- UnitedHealth : y su sistema algoritmo de recomendación médica

Compañía de seguros de salud UnitedHealth enfrenta investigación sobre discriminación racial en su algoritmo

TITULAR 05 NOV. 2019



El organismo regulador de seguros del estado de Nueva York está iniciando una investigación sobre la empresa UnitedHealth Group, luego de que un estudio mostró que un algoritmo de UnitedHealth priorizaba brindar atención médica a pacientes blancos más sanos que a pacientes negros más enfermos. Un estudio publicado en la revista



Data bias y sus efectos

Representation bias

- Midjourney: dando cositas por hecho



Midjourney mostró a mujeres en puestos no especializados, como periodistas (derecha).

También mostró solo a hombres mayores (pero no a mujeres mayores) en funciones especializadas como analista de noticias (izquierda).



Data bias y sus efectos

Representation bias

- Midjourney: dando cositas por hecho



Midjourney generó imágenes con personas de piel clara exclusivamente para todos los títulos laborales utilizados en las peticiones, incluidos de comentarista de noticias (izquierda) y reportera (derecha).



Data bias y sus efectos

¿Culpable?

- Antes de asignar la culpa, de todos modos, hay que tener en cuenta que el propósito de la minería de datos/aprendizaje automático es discriminar
 - / quién obtiene el préstamo
 - / quién recibe la oferta especial
 - / qué precio pagar por nuestro seguro
- Por lo tanto, la discriminación (**clasificación**) está en el centro de la mayoría de las técnicas de aprendizaje automático.
- Algunos tipos de discriminación **no son éticos** (incluso son ilegales). Pero depende del contexto, puede ser ético (clasificaciones médicas).

Data bias y sus efectos

Evitar problemas con el BIAS

- Un truco para evitar problemas éticos, es anonimizar los datos
/ Y ojito...



⌚ Feedback loop

Definición

- Los bucles de retroalimentación en IA ocurren cuando los datos que recibe como entrada a nuestro modelo también son generados o influenciados por su modelo.
- Usado más veces de lo que pensamos
 - / Amazon
 - / Youtube
- PERO....

♐ Feedback loop

Problemas

The New York Times

THE SHIFT

YouTube Unleashed a Conspiracy Theory Boom. Can It Be Contained?

Or take Pizzagate, a right-wing conspiracy theory that alleged that Hillary Clinton and other Democrats were secretly running a child-sex ring. The theory, which was spread in a variety of videos on YouTube and other platforms, might have remained an internet oddity. But it became a menace when a believer [showed up](#) at a pizza restaurant in Washington, D.C., with an assault rifle, vowing to save the children he believed were locked in the basement.

⌚ Feedback loop

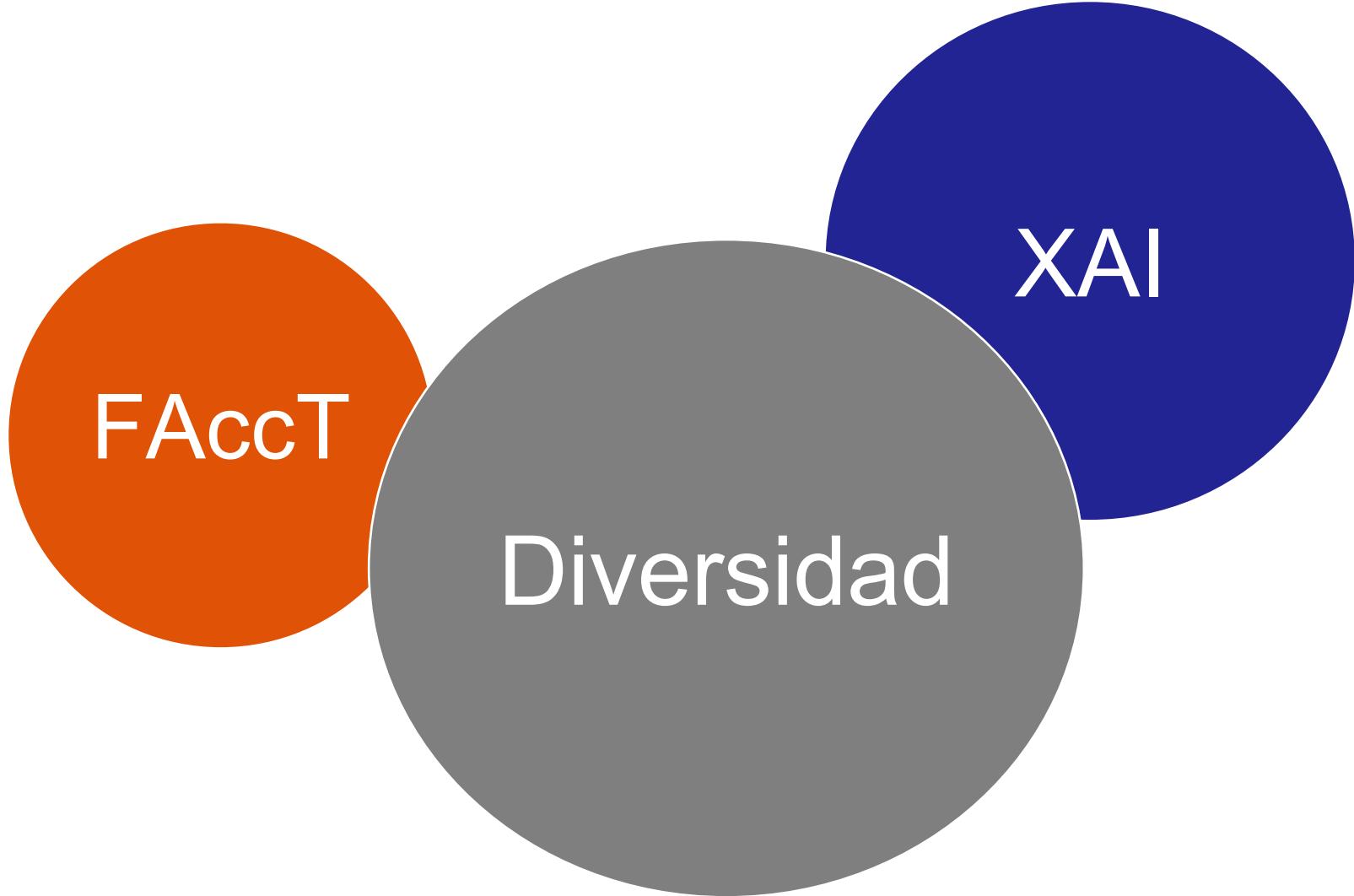
Problemas

THE INTERPRETER

*On YouTube's Digital Playground, an
Open Gate for Pedophiles*



- El problema, parece ser que el algoritmo recomienda, cada vez, videos más extremos o gráficos o más largos o más enfocados en el tema y donde tus contactos o personas relacionadas ya interactuaron (ver, dar me gusta o comentar).





- General Data Protection Regulation
- Más información: <https://gdpr-info.eu/>



Leyes y legalidad

Resumen de la GDPR

- Las leyes de privacidad de la información (UE) generalmente se pueden **resumir** como:
 - / Solo podemos recopilar datos para solo el propósito anunciado.
 - / No podemos dar esa información a otros sin consentimiento explícito.
 - / Los datos deben ser lo más precisos posible.
 - / Los ciudadanos deberían poder revisar los datos que hemos recopilado sobre ellos.
 - / Cuando los datos ya no sean necesarios, deben eliminarse.
 - / Los datos deben conservarse únicamente cuando pueda garantizarse su protección.
 - / Hay datos que no podemos recopilar, demasiado sensibles, altamente regulados:
 - Orientación sexual
 - Religión
 - Salud
 - etc.



Definiciones GDPR

Datos personales

- Cualquier información relacionada con un individuo que puede ser identificado directa o indirectamente.
- Los nombres, teléfonos y direcciones de correo electrónico son datos personales.
- La geubicación, los datos étnicos, el género, la biometría, la religión, las cookies web, las opiniones laborales y políticas y las preferencias sexuales son datos personales.
- Los seudónimos/anonimizados también pueden ser datos personales si es fácil identificar a las personas.



Definiciones GDPR

Procesamiento de datos

- Acciones (automáticas o no) realizadas sobre los datos.
- La ley cita (sin excluir otras):
 - / recopilar
 - / registrar
 - / organizar
 - / estructurar
 - / almacenar
 - / usar
 - / borrar



Definiciones GDPR

Sujeto de datos

- Persona cuyos datos están siendo tratados.
- Incluso los visitantes del sitio web.

Controlador de los datos

- Persona que decide sobre el tratamiento de datos personales (por qué y cómo).
- Todos los que manejan datos (ya sean propietarios o empleados).

Procesador de datos

- Tratamiento de datos por parte de terceros a petición del responsable del tratamiento.
- Esto incluye servidores en la nube, proveedores de correo electrónico, empresas de consultoría, etc.



GDPR describe sus principios en el artículo 5.1-2:

Principios

- Legalidad, equidad y transparencia
- Limitación de propósito
- Minimización de datos
- Exactitud
- Limitación de almacenamiento
- Responsabilidad
- Integridad y confidencialidad

Además: *Artículo 25: Protección de datos por diseño y por defecto.*

Seguridad

GDPR también proporciona protecciones de seguridad de datos:

- El propietario de los datos está obligado a implementar las medidas técnicas y organizativas apropiadas.
- Las medidas técnicas pueden ser:
 - / Protecciones criptográficas.
 - / Autenticación más difícil.
- Medidas organizativas:
 - / Formación de los empleados.
 - / Limitación de acceso a datos a empleados requeridos.
- Incluso está obligado a notificar (límite de 72 horas, a menos que utilice protecciones criptográficas) cualquier violación de datos.

Limitaciones del procesamiento de los datos

GDPR impone limitaciones sobre cuándo podemos procesar datos:

- Se debe contar con el consentimiento específico e inequívoco del sujeto.
- Es necesario un contrato en el que el interesado es parte.
- Para cumplir con una obligación legal.
- Para salvar la vida de alguien.
- Para realizar una tarea de interés público o una función oficial.
- Los derechos y libertades fundamentales del sujeto siempre prevalecen sobre nuestros intereses.



Derechos de privacidad

GDPR reconoce a las personas una gran cantidad de derechos de privacidad:

- Derecho a ser informado
- Derecho de acceso
- Derecho de rectificación
- Derecho de borrado
- Derecho a restringir el procesamiento
- Derecho a la portabilidad de los datos
- Derecho de oposición
- **Derechos en relación con la toma de decisiones automatizada y elaboración de perfiles.**

Bibliografía

- [1] Mateu, Carles. "Massive data processing", Universitat de Lleida. Lleida. 2021.
- [2] <https://royalsocietypublishing.org/doi/10.1098/rsta.2016.0360>
- [3] <https://gdpr-info.eu/>

PREGUNTAS



¡Gracias!

alba.lamas@bluetab.net

¡Síguenos!



<https://bluetab.net/>



<https://www.linkedin.com/company/bluetab/>

