Desafíos éticos de la Inteligencia Artificial

Angel Rubio, Ph D.

Profesor de Inteligencia artifical en TECNUN

(Universidad de Navarra)

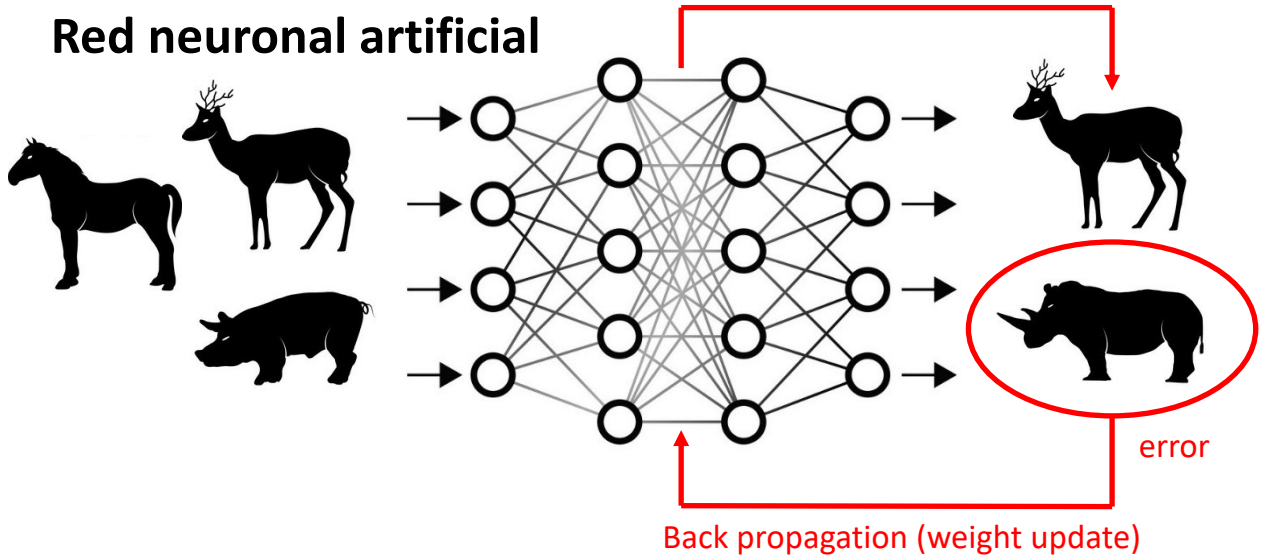
¿Qué es la inteligencia artificial?

- La inteligencia artificial (IA) es una disciplina que se centra en la creación de sistemas informáticos y de algoritmos capaces de imitar la inteligencia humana para **realizar tareas** y **mejorar** según recopilan información.
- Pueden llevar a cabo **procesos propios de la inteligencia humana**, como aprender*, razonar* o autocorregirse.
- Inteligencia artificial **débil** → tareas específicas
- Inteligencia artificial general o **fuerte** → IA hipotética que excediese las capacidades humanas y pudiera sustituir a una persona en sus tareas.
  - Es el objetivo último de OpenAI, Anthropic, o DeepMind.

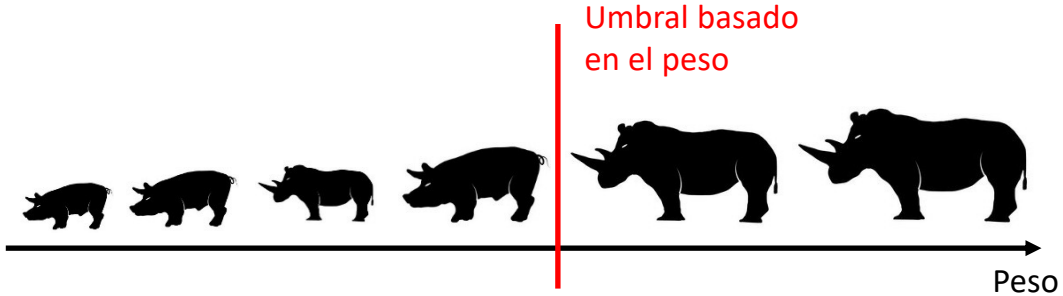
Tipos de datos de entrenamiento

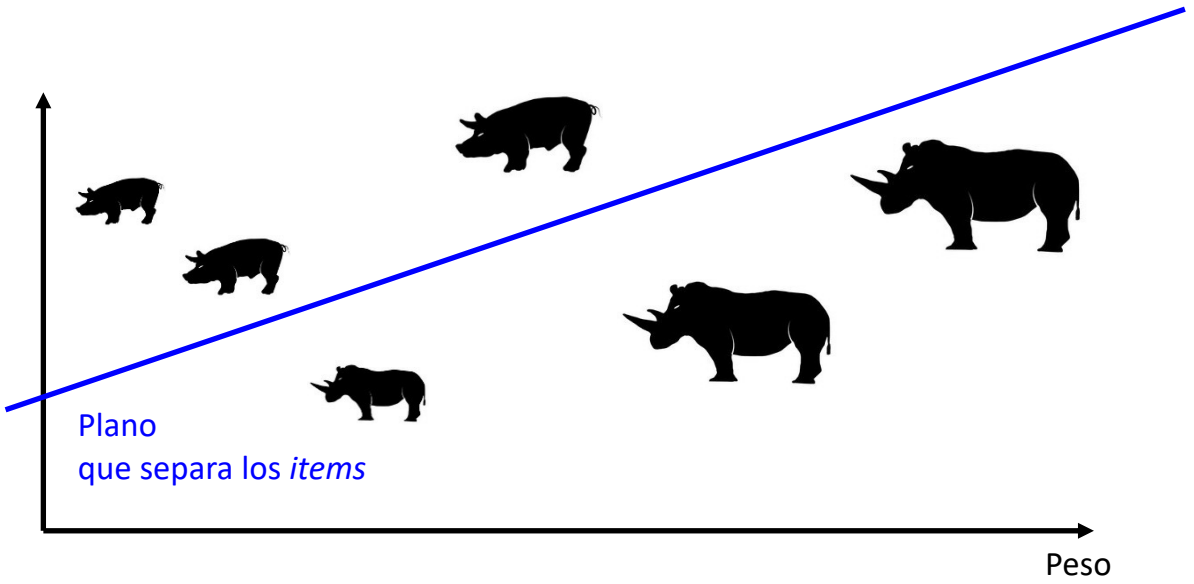
- Estructurados:
  - Los datos de entrenamiento son una o varias “hojas Excel” de números. Algunas columnas se marcan como entradas y otras como salidas.
  - A pesar de ser “estructurados”, tiene muchísimas aplicaciones: análisis de imágenes, reconocimiento de voz, interpretación de caracteres manuscritos, predicción en bolsa…
  - Utilizo los datos de entrada para hacer que mi IA, al introducirle las entradas, me devuelva las salidas.
    - Se usan diversas técnicas para asegurar que IA es capaz de generalizar bien: es decir, que al darle nuevos datos las predicciones sean razonables.
- No estructurados
  - Las datos de entrenamientos son textos, imágenes, sonidos, vídeos, etc. que no se han etiquetado de forma explícita cuáles son las entradas y cuáles son las salidas. Se pasa toda esta información a IA y ”aprende” de ahí. ¿Qué aprende? Un poco de todo ☺
  - En el caso de la IA fuerte, los datos son obviamente desestructurados.

¿Cómo funcionan con datos estructurados?



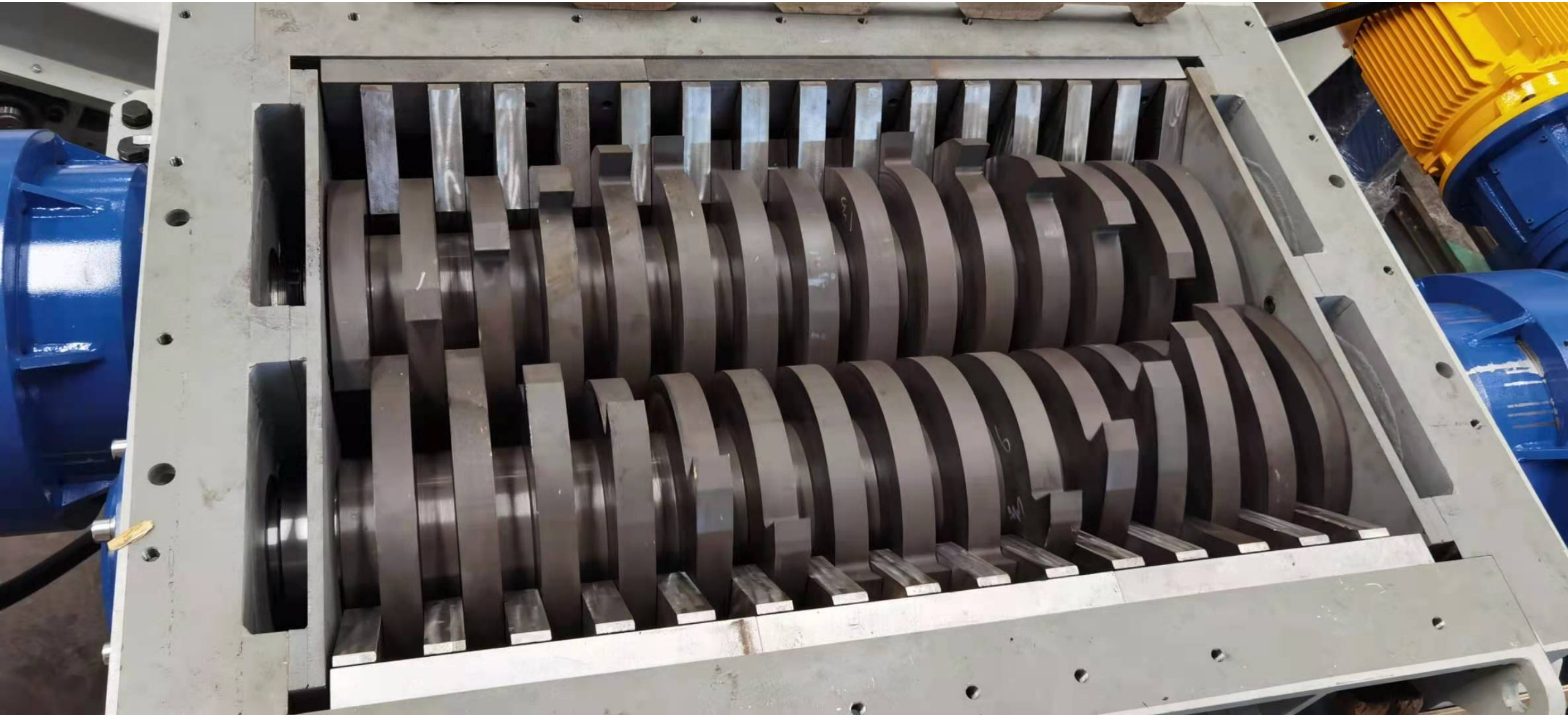
- Para aprender automáticamente se necesitan **ejemplos etiquetados a mano** (aprendizaje supervisado).
- Para que la tasa de fallos sea pequeña se requieren **muchísimos ejemplos** etiquetados.





¿Cómo funcionan con datos no estructurados?

tecnun
Universidad de Navarra

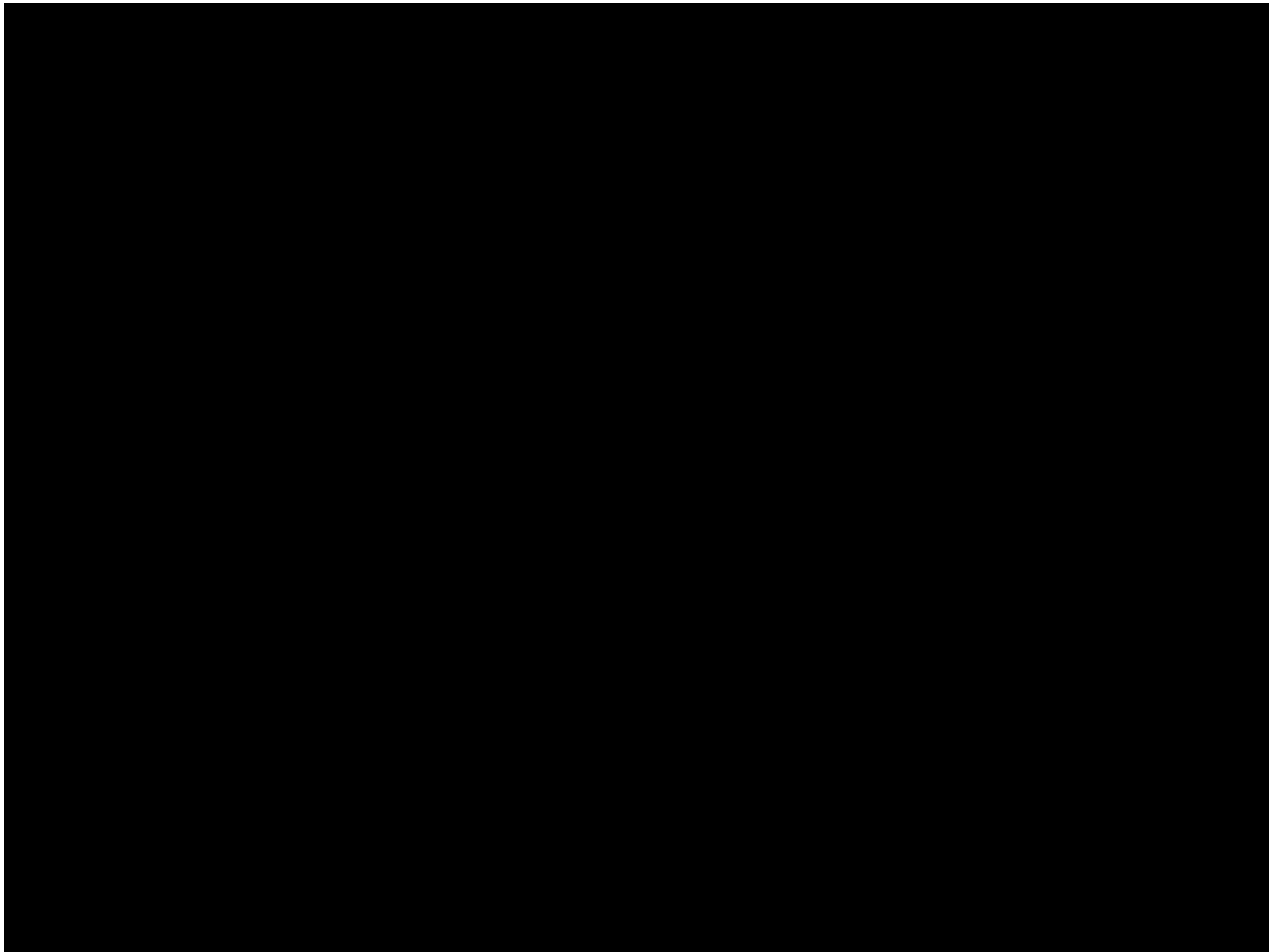


¿Cómo funcionan con datos no estructurados?

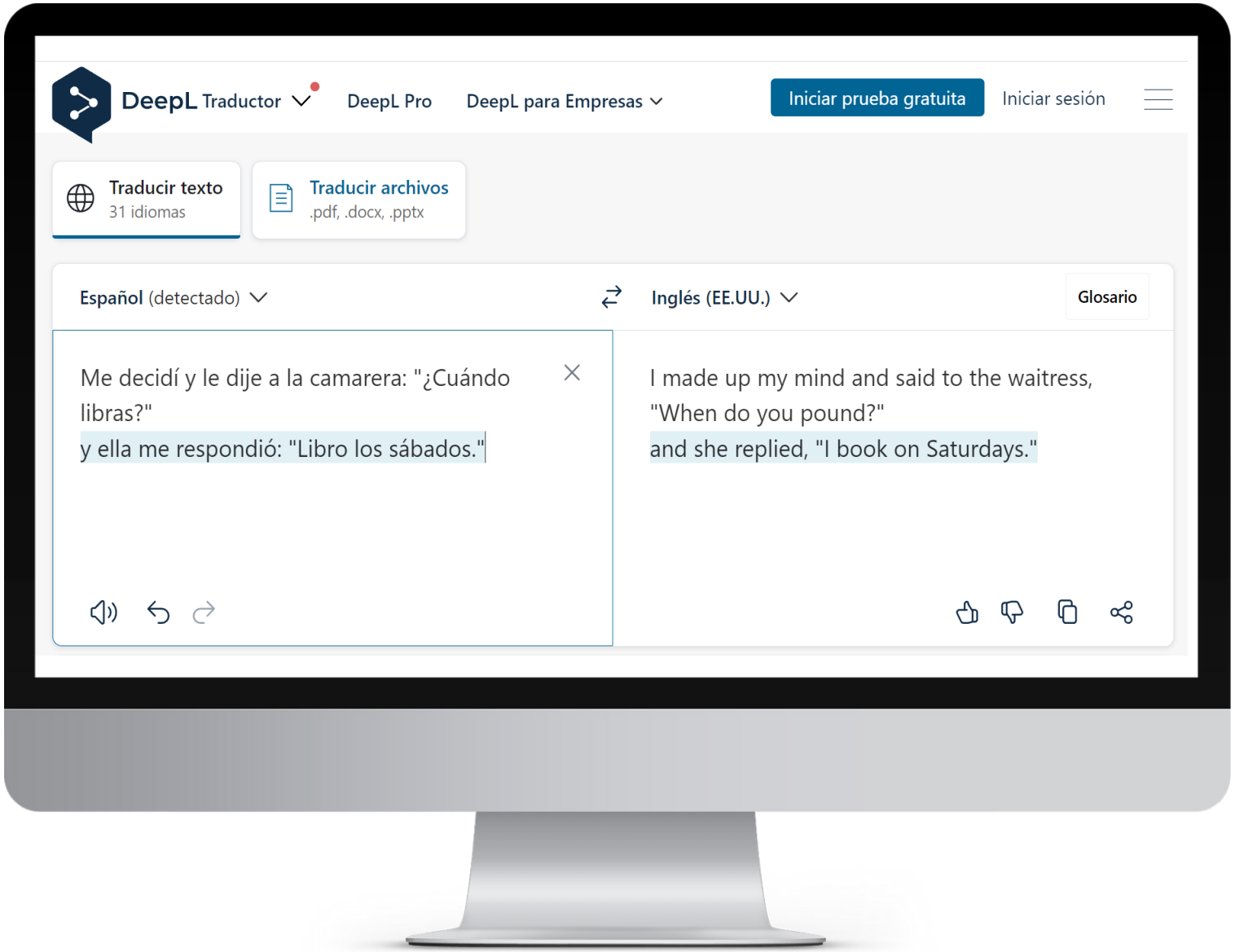
- La IA es alimentada con datos (si son de diferente tipo: texto, imágenes, vídeos, sonidos, se la denomina multimodal).
  - Google acaba de anunciar “Lumiere” una IA capaz de producer videos con instrucciones.
- En primer lugar hace una “estructuración”
  - Por ejemplo en el caso de texto, las palabras se codifican en vectores de números.
  - Por decirlo así se “traduce” el texto al lenguaje del ordenador.
  - Si las palabras son parecidas en significado, los vectores serán parecidos entre sí.
- En los chatBots, se analiza el contexto.
  - Un chatbot es un texto predictivo del móvil llevado al extremo: con lo que se le ha escrito, y lo que lleva respondido, basándose en los datos de entrenamiento, predice la siguiente palabra

Academias de inglés ☺

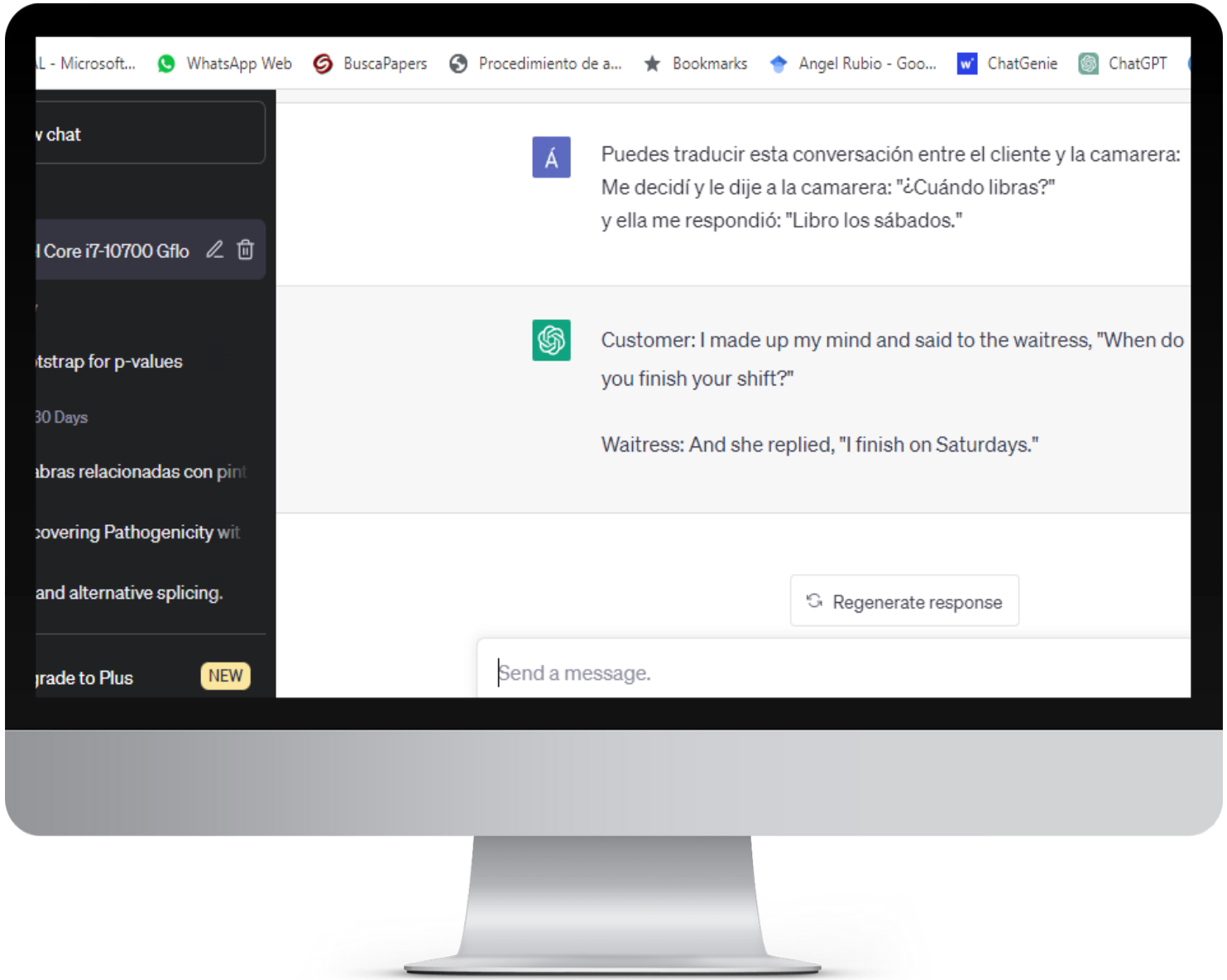
tecnun
Universidad de Navarra



Transformers: importancia del contexto



Transformers: importancia del contexto

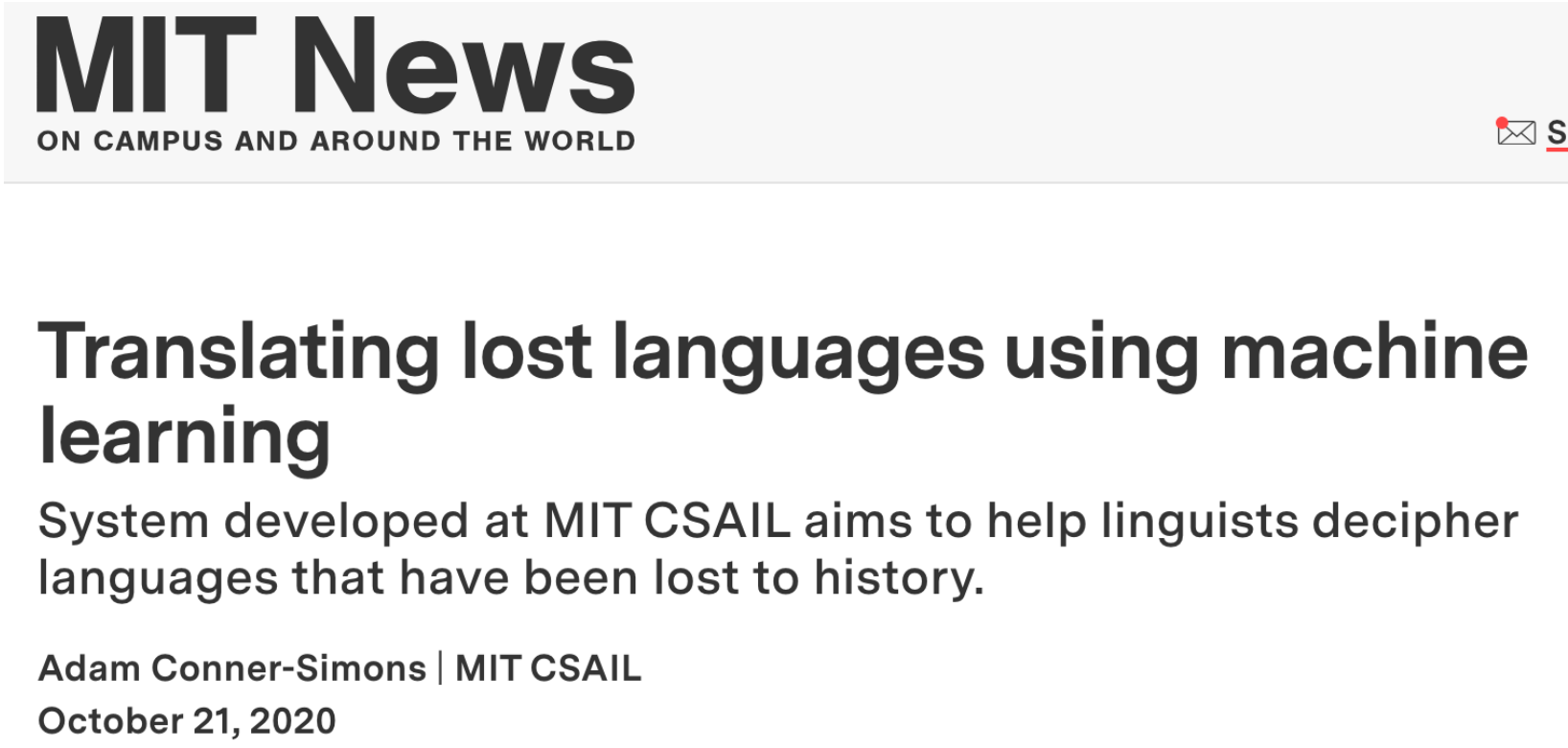


¿Cómo funcionan con datos no estructurados?

- El concepto básico es el “Transformer”:
  - Un transformer es una red neuronal que aprende **incorpora el contexto** estudiando las relaciones en datos secuenciales como las palabras de esta oración.
- En agosto de 2021, los investigadores de Stanford llamaron a los transformers los «**modelos de base**» porque ven que impulsan un cambio de paradigma en la IA.
  - «La gran escala y el alcance de los modelos de base en los últimos **años han extendido nuestra imaginación de lo que es posible**».

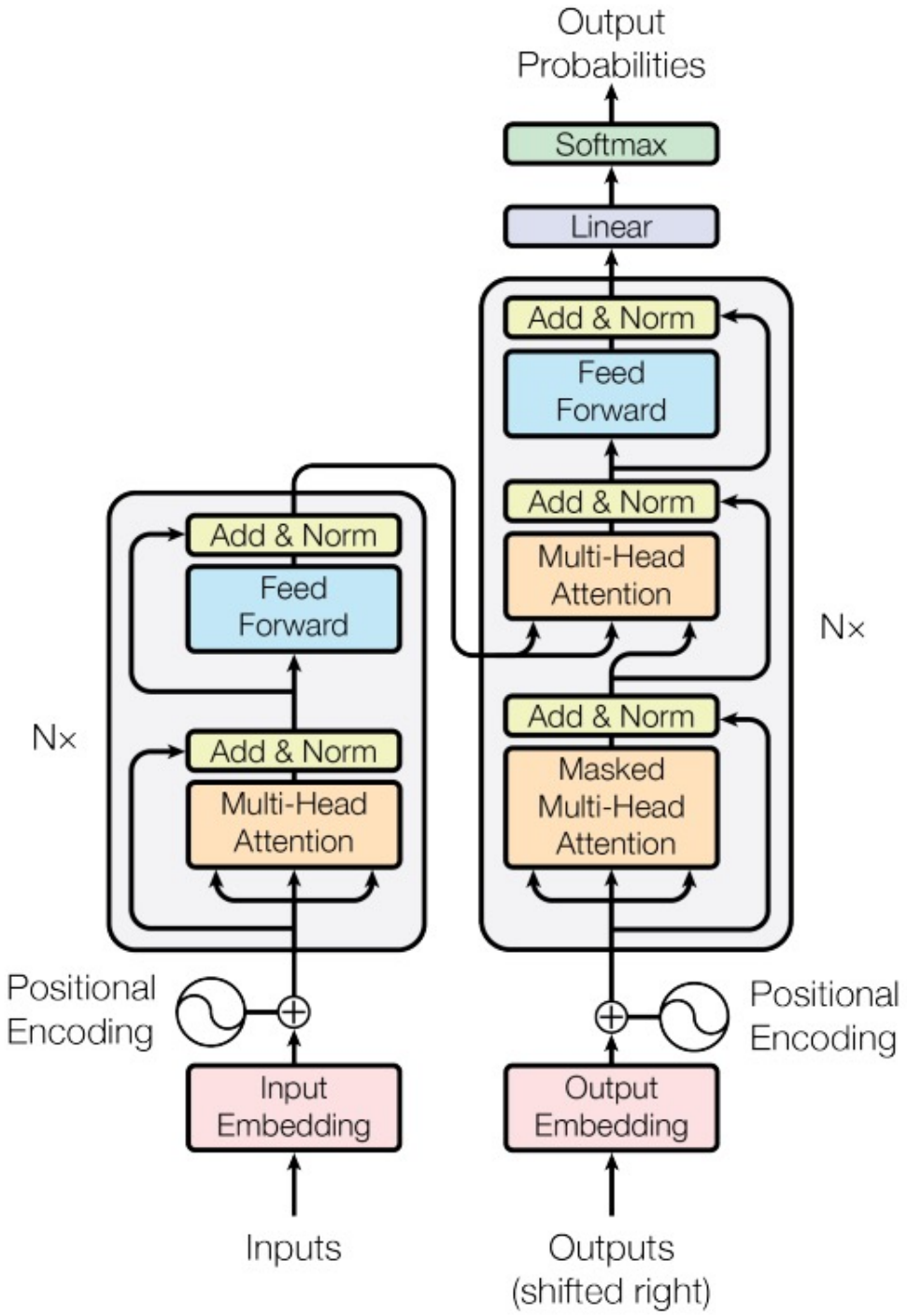
Transformers: usos

- Traducción de texto y habla en tiempo real.
- Clasificación, resumen y generación de textos
- Investigación Bio: detección de genes en el ADN, efectos de mutaciones, plegado de proteínas… Todo ello puede acelerar el diseño de fármacos.



Transformers: estructura

- El concepto fundamental es la atención: diferentes “cabezas de atención” se fijan en cómo de importante es cada palabra codificada en relación con el resto. Esto se puede hacer para diferentes aspectos de relevancia.
- Se entrenan casi solos, sólo al final puede hacerse un aprendizaje supervisado.
- Pueden usar muchos más datos!!
- https://la.blogs.nvidia.com/2022/04/19/que-es-un-modelo-transformer/



¿Por qué ahora?

- Durante décadas el aprendizaje automático **no ha funcionado muy bien**.
- ¿Qué ha cambiado?
  - Mejor hardware
  - Mejor software
  - Mejores datos
- Las potenciales aplicaciones son enormes (en la banca, en la medicina, en las leyes, etc.).
  - Traducción automática, Transcripción directa de voz, detección automática de objetos, identificación de caras, etc. pero también, robots cuyo modelo de comportamiento es difícil de distinguir de una persona.

¿Por qué ahora? El hardware

- Los modelos de IA son computacionalmente muy costosos y había diversos problemas matemáticos que hacían que no funcionaran bien. Ahora se puede hacer gracias a las potentísimas **tarjetas gráficas** desarrolladas para juegos (las GPU).
- Las GPU y las TPU tienen una potencia de cálculo brutal. Saben hacer pocas cosas: sumar y multiplicar, pero lo hacen rapidísimo, mucho más rápido que una CPU normal
  - Por ejemplo, el ordenador de mi despacho:
    - GPU RTX 2080Ti → 13.5 Tflops, o sea, 13.5 millones de millones de operaciones por segundo.
    - CPU I7-8700 → 77 Gflops, o sea, 175 veces menos que la CPU.
- chatGPT 3.0 tiene 175.000 millones de parámetros a ajustar
  - Aunque parece mucho… no es tanto.
  - Cabe de sobre en el disco duro de casi cualquier portátil y casi cabría en la memoria de un ordenador de sobremesa bueno (son 700Gb de datos).

¿Por qué ahora? El software

- Aprendizaje estructurado
  - Deep Learning en el análisis de imágenes y de señales.
    - Es una metodología que se comenzó a desarrollar alrededor de 2011. Ahora mismo, es difícil pensar en un análisis de imagen sin hablar de ellos.
- Aprendizaje no estructurado:
  - Los Transformers se inventaron en Google Brain en 2017
  - Son muy superiores a los métodos anteriores que existían en lenguaje natural.
    - Pueden procesar secuencias de longitud variable y hacer **todo el proceso de atención en paralelo** en todas las palabras de la entrada.

¿Por qué ahora? Acceso a datos

- Las empresas de redes sociales, los bancos, las empresas de telefonía, los gobiernos… tienen disponible una cantidad de datos enorme, y muchos de ellos etiquetados.
- El acceso a esos datos, permite imaginar distintas aplicaciones, que con la tecnología actual ya son posibles.

Asistentes personales

tecnun
Universidad de Navarra

Alpha Zero

Algunos hechos

- "Los sistemas de AI pueden **sustituir con ventaja** a una persona en **determinadas tareas**"
  - El número y **el tipo de tareas** en que esto ocurre es cada vez **mayor**.
  - En general, **la rapidez con la que se ejecutan** –o se pueden ejecutar disponiendo de recursos, o se espera que se ejecuten en un tiempo próximo- estas tareas **son mucho mayores** que las de cualquier persona.
  - Es decir, se pueden hacer las tareas con eficacia similar, pero a una velocidad muchísimo mayor.
- "El avance de la IA en los últimos diez años (especialmente los últimos tres años) ha sido espectacular"
  - Probablemente el ritmo de cambio sea todavía mayor en los años que vienen.

Algunos hechos

- “Los sistemas disponibles actualmente para el gran público están **muy retrasados con respecto a los prototipos que tienen las empresas** de desarrollo o las aplicaciones militares”
- “Los sistemas de AI pueden tomar decisiones que afectan de forma crítica a la vida de las personas”
  - Conducción autónoma
  - Asistentes médicos
  - Asistentes para jueces
  - Asistentes para toma de decisiones en empresas…
- “Los sistemas de AI pueden cambiar un estado de opinión”
  - Cambridge analytics

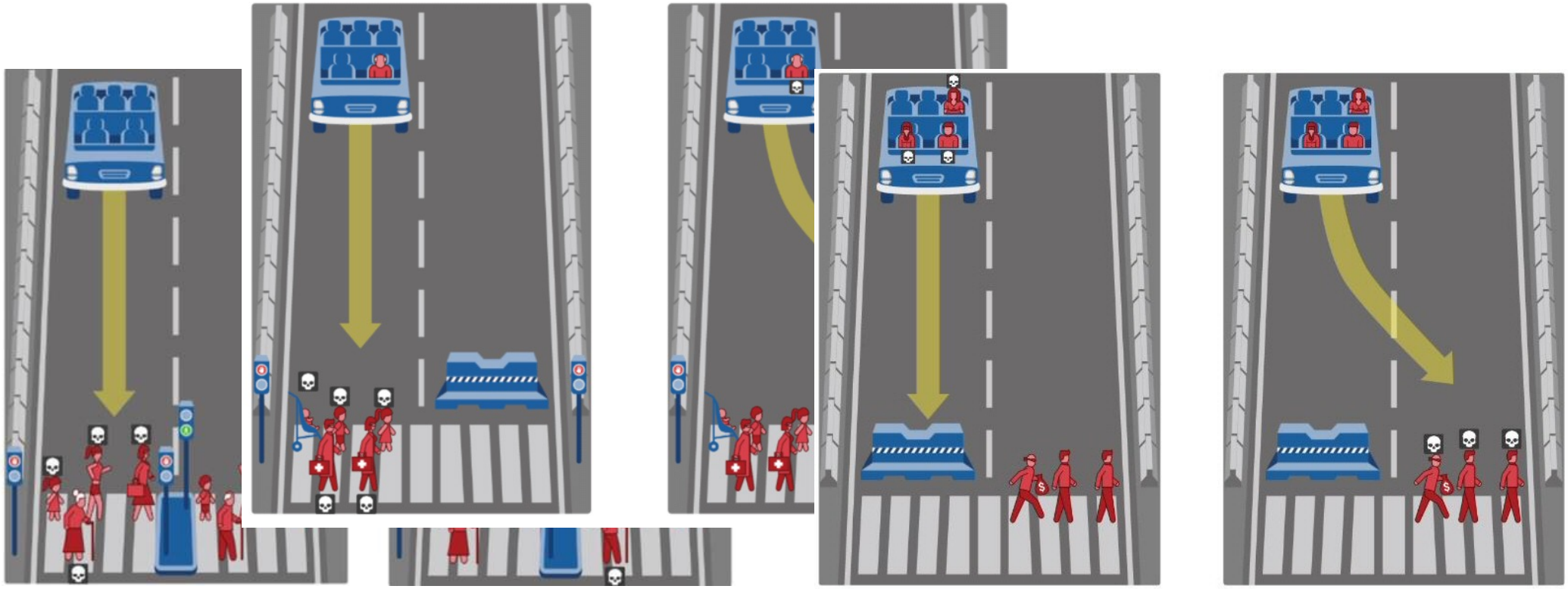
Es necesario considerar los aspectos éticos en IA

Aspectos y desafíos éticos

- ¿Cómo programar una IA que tenga una respuesta “moral”?
- Necesidad de una legislación
- Desigualdad laboral: ¿nos van a quitar el trabajo?
- Atribución de responsabilidades
- Influencia en el comportamiento humano
  - Vigilancia y seguridad
  - Manipulación del comportamiento
- Sesgo, diversidad y exclusión
- Desinformación y “fake news”
  - https://youtu.be/gLoI9hAX9dw
- Usar la inteligencia artificial para invadir el ámbito privado del consumidor
- Singularidad

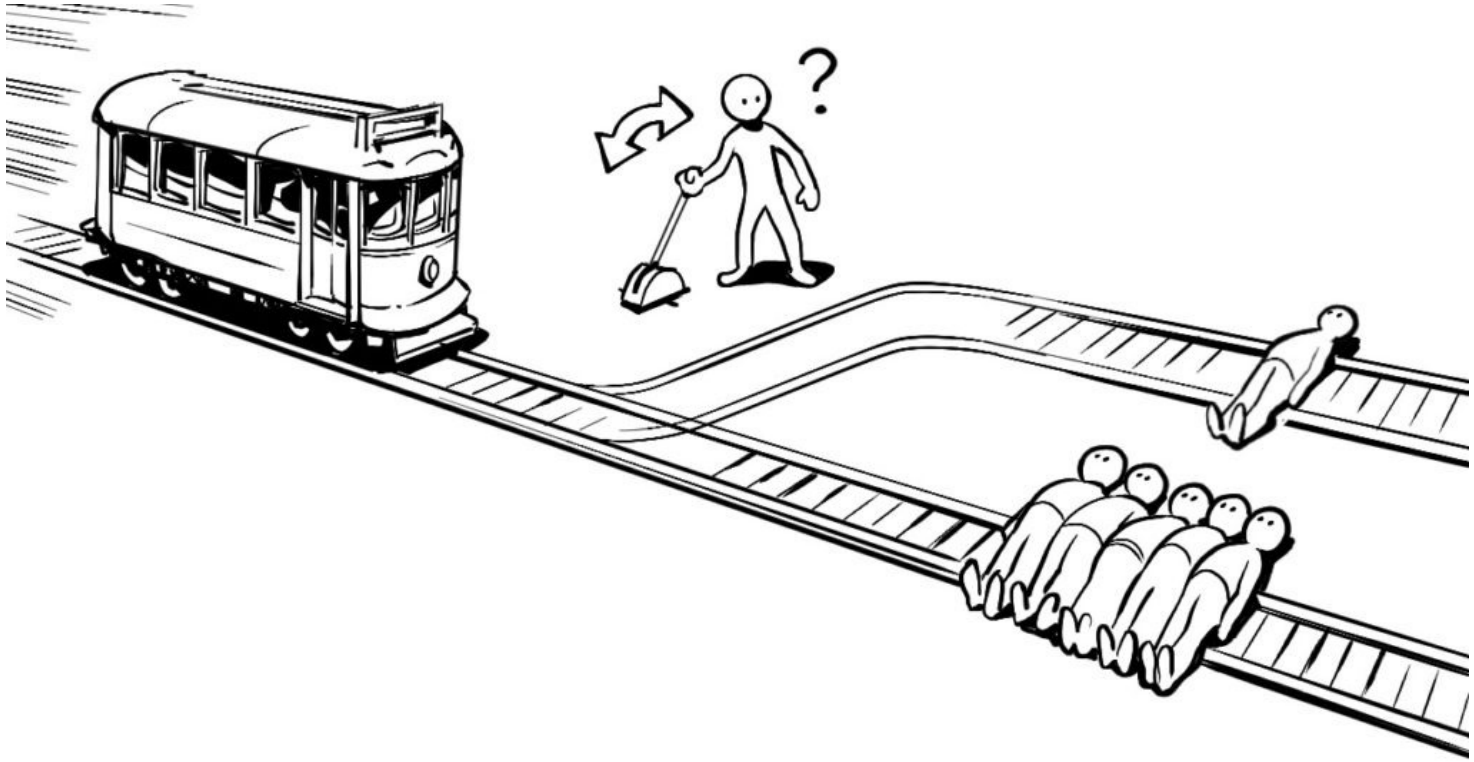
Programación “ética”: usando el método del caso

- Otros algoritmos deciden sin reglas explícitas sino usando el aprendizaje automático supervisado. Para ello el MIT creo el experimento llamado la **máquina moral**:
  - Se plantean escenarios y las personas deben votar qué decisión es moralmente más aceptable.
  - Participantes: 2,3 millones de personas de 233 países, que sumaron 40 millones de decisiones



Ética para máquinas

- Dilema del tranvía (**Philippa Foot**, 1967):
  - Un tranvía corre fuera de control por una vía. En su camino se hallan cinco personas atadas a la vía por un filósofo malvado. Afortunadamente, es posible accionar un botón que encaminará al tranvía por una vía diferente, por desgracia, hay otra persona atada a ésta. ¿Debería pulsarse el botón?



MORAL MACHINE

Ética para máquinas

- Resultados de la **máquina moral** publicados en:
  - Awad, E., Dsouza, S., Kim, R. et al. *The Moral Machine experiment*. Nature 563, 59–64 (2018).

ARTICLE

https://doi.org/10.1038/s41586-018-0637-6

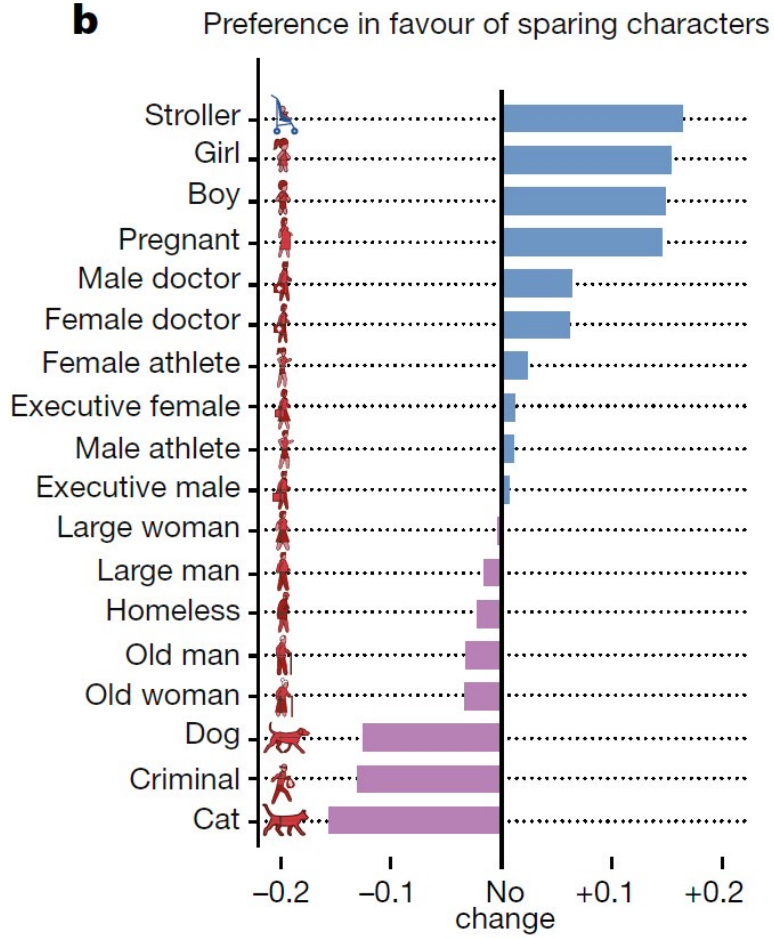
The Moral Machine experiment

Edmond Awad[1], Sohan Dsouza[1], Richard Kim[1], Jonathan Schulz[2], Joseph Henrich[2], Azim Shariff[3]*, Jean-François Bonnefon[4]* & Iyad Rahwan[1,5]*

With the rapid development of artificial intelligence have come concerns about how machines will make moral decisions, and the major challenge of quantifying societal expectations about the ethical principles that should guide machine behaviour. To address this challenge, we deployed the Moral Machine, an online experimental platform designed to explore the moral dilemmas faced by autonomous vehicles. This platform gathered 40 million decisions in ten languages from millions of people in 233 countries and territories. Here we describe the results of this experiment. First, we summarize global moral preferences. Second, we document individual variations in preferences, based on respondents' demographics. Third, we report cross-cultural ethical variation, and uncover three major clusters of countries. Fourth, we show that these differences correlate with modern institutions and deep cultural traits. We discuss how these preferences can contribute to developing global, socially acceptable principles for machine ethics. All data used in this article are publicly available.

We are entering an age in which machines are tasked not only to promote well-being and minimize harm, but also to distribute the well-being they create, and the harm they cannot eliminate. Distribution of well-being and harm inevitably creates tradeoffs, whose resolution falls in the moral domain[1–3]. Think of an autonomous vehicle that is

vehicles, and for the wider public to accept the proliferation of artificial intelligence-driven vehicles on their roads, both groups will need to understand the origins of the ethical principles that are programmed into these vehicles[10]. In other words, even if ethicists were to agree on how autonomous vehicles should solve moral dilemmas, their work



Ley europea de IA (9.XII.2023)

- No solo datos… También y, especialmente, **qué se hace con los datos.**
- **Sistemas de alto riesgo → sistemas de los que depende la vida de personas**
- Deben tener una **evaluación del impacto** en los derechos fundamentales antes de introducirse.
- También disposiciones para las **IA de uso general** que puedan integrarse en un sistema de alto riesgo.
- Los modelos fundacionales (sistemas competentes en tareas como la generación de vídeo, texto e imágenes, generación de códigos informáticos, …) deben cumplir obligaciones específicas en materia de **transparencia**, especialmente en los de alto riesgo

Sistemas de alto riesgo

- Estructuras básicas:
  - **Transporte** (apoyo a la conducción y prevención de accidentes ferroviarios), **Energía** (gestion, almacenamiento y distribución de energía y gas), suministro de **agua potable**,
  - **Infraestructuras digitales críticas**: servicios de nube, servicios de compensación y liquidación, servicios de pago electrónico y transporte de bienes críticos
  - **Medio ambiente**: predicción de catástrofes naturales y prevención de la contaminación
- Sistemas **sanitarios** (diagnóstico, atención al paciente, cirugía)
- Evaluación de personas
  - **Educación**: evaluación del aprendizaje y clasificación de los estudiantes
  - **Seguros**: evaluación de riesgos y liquidación de seguros
  - **Empleo**: selección de candidatos y evaluación de empleados
  - **Aplicación de la ley**: detección de delitos graves y lucha contra el terrorismo
  - **Migración**: determinación de la admisibilidad de los solicitantes

Evaluación de riesgos

- El **tipo, intensidad y duración del riesgo** para la salud, la seguridad y los derechos fundamentales
- La **probabilidad** de que el riesgo se materialice;
  - La IA puede minimizar o aumentar el riesgo?
  - Tomando las medidas para minimizarlo, ¿cuál es la nueva probabilidad?
- La accesibilidad y calidad de los **datos** de entrada y la **forma en que se procesan**;
- Asegurar una **supervision continua** de estos sistemas, incluyendo la provisión de información clara y comprensible a los interesados.
- Garantizar la **fiabilidad, transparencia y responsabilidad** del sistema de inteligencia artificial.

Desafíos éticos

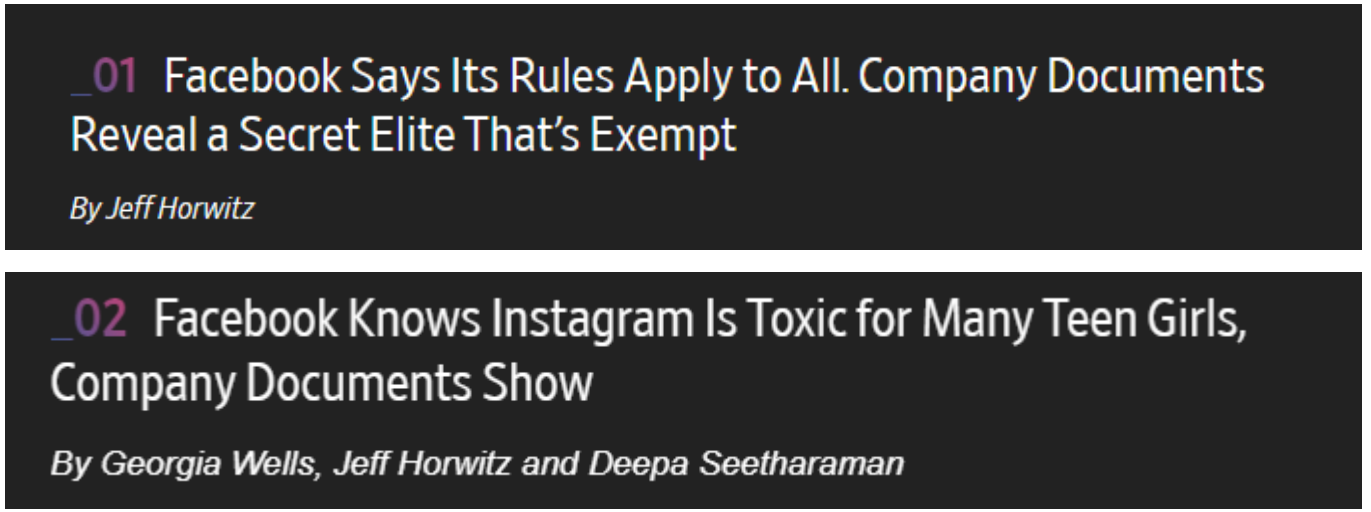
- Desigualdad laboral: ¿nos van a quitar el trabajo?
  - La revolución industrial hizo que pasáramos de jornadas de 13 o 14 horas, a jornadas de 7-8 horas. Nunca hemos tenido tanto ocio como ahora.
  - Puede ser que nos puedan sustituir en casi todo, y haya que crear una renta básica para vivir, porque no tendremos mucho en lo que trabajar??!!
  - **Habrá trabajos que desaparezcan**, pero aparecerán trabajos nuevos y **nos iremos adaptando**.

Desafios éticos

- A quién se atribuye la culpa cuando algo va mal
  - Casi todos los programas tienen un “disclaimer”, es decir, en ningún caso la responsabilidad es suya.
  - Si un coche autónomo atropella a alguien, de quién es la culpa. ¿De Tesla? ¿De nadie?, “Incidente de circulación”.
    - En la nueva ley, es de la empresa que lo propone (Tesla en este caso) e, incluso, de los desarrolladores y los usuarios.
  - Este motivo no es trivial:
    - Los mecanismos de autoland de los aviones son muy eficaces, pero siempre se usan con supervisión del piloto.

Desafíos éticos

- Influencia en el comportamiento humano
  - El caso de Cambridge Analytics. Usando información de redes sociales, identificaban personas indecisas y fácilmente manipulables para tratar de modificar su intención de voto… y funcionaba. Además usando esa información y tácticas sórdidas podían chantajear a políticos.
  - Un informe interno de Facebook
    - https://www.wsj.com/articles/the-facebook-files-11631713039



Debate sobre los problemas éticos

- Sesgo, diversidad y exclusión
  - ¿De quién aprendo?
    - Ser negro es un factor que pesa en la probabilidad de que sea culpable, simplemente porque en el histórico del modelo de entrenamiento, ha habido más negros que son culpables.
    - Aunque los datos son correctos pueden extraerse “deformidades”: Lo que es aceptable ahora puede no serlo dentro de pocos años. El algoritmo puede recrear errores históricos.
    - A veces los factores de confusión –menores ingresos, criminalidad de donde se vive, etc.- pueden ser la razón de esa mayor criminalidad y no el color de la piel. Sin embargo, se puede volver a contraargumentar: la IA juzga contra los pobres, contra la gente de los barrios obreros, etc.

Desafíos éticos

- Desinformación y “fake news”

TECH | ARTIFICIAL INTELLIGENCE

A college student used GPT-3 to write fake blog posts and ended up at the top of Hacker News

He says he wanted to prove the AI could pass as a human writer

By Kim Lyons | @SocialKimLy | Aug 16, 2020, 1:55pm EDT

Dilemas éticos

- Singularidad
  - ¿Nos dirigimos hacia el Nirvana o hacia el Apocalipsis? Miedo la **“singularidad”**, es decir, a que la inteligencia artificial supere a la humana y evolucione sin control.
  - La singularidad hace referencia al crecimiento “exponencial”: las inteligencias artificiales si son m´s capaces que la humana, serán capaces de crear mejores inteligencias artificiales, que a su vez crearan otras que son mejores…

Desafíos éticos en la Inteligencia Artificial
¿Preguntas?

Angel Rubio