ellis ALICANTE unit

sanrafael fundaciónestima

Inteligencia Artificial y Bienestar Humano

Adrián Arnaiz, Erik Derner

8/11/2024

¿Quiénes somos?

- Investigador predoctoral @ ELLIS Alicante
- De Burgos
- Investigador en justicia algorítmica en la toma de decisiones en contextos sociales

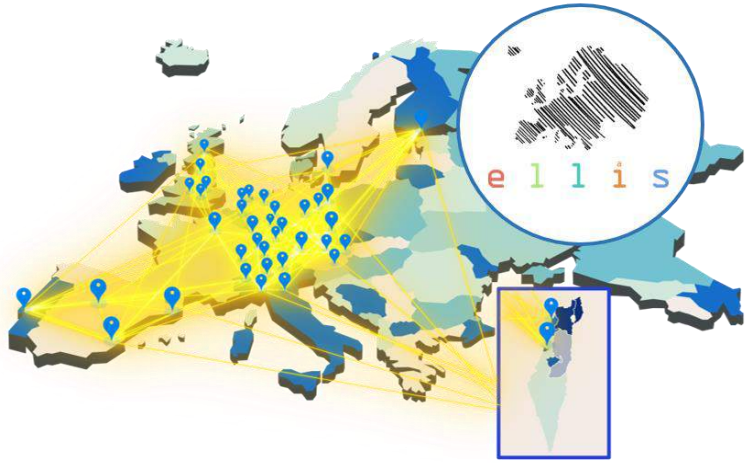
Adrián Arnaiz

- Investigador postdoctoral @ ELLIS Alicante
- De Praga (Chequia)
- Doctor de robótica
- Investigación centrada a los desafíos éticos y sociales de la IA generativa

Erik Derner

ELLIS Alicante

- **ELLIS**
  - European Laboratory for Learning and Intelligent Systems *(Laboratorio Europeo para Sistemas de Aprendizaje e Inteligencia)*
  - Red de investigación **paneuropea** que conecta a los científicos y laboratorios en el campo del aprendizaje automático en **43 unidades** en **17 países**
- **ELLIS Alicante**
  - La primera unidad ELLIS en España, creada por **Dra. Nuria Oliver** cómo una fundación de investigación independiente sin ánimo de lucro en 2020
  - Equipo de investigación que se centra en la investigación de la **inteligencia artificial ética** enfocada en la **relación entre las personas y los sistemas inteligentes**



IA para el bienestar humano

¿Qué es la IA?

Problemas éticos de la IA y cómo mitigarlos

IA para el bienestar humano

IA para el bienestar humano

¿Qué es la IA, qué hace?

¿Qué problemas éticos genera la IA?

¿Cómo utilizar la IA para el bienestar humano?

IA para el bienestar humano

¿Qué es la IA?

Problemas éticos de la IA y cómo mitigarlos

IA para el bienestar humano

Procesar información

Máquinas para procesar información / datos

Sacar conclusiones

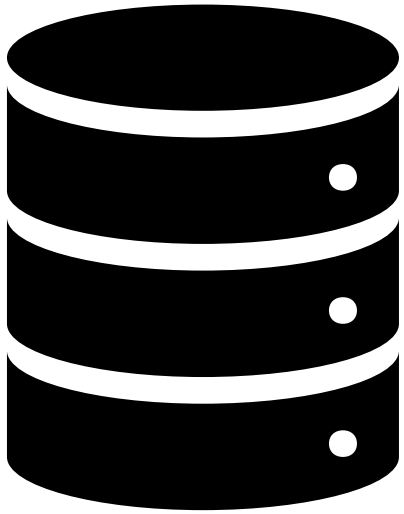
Tomar decisiones

Automatizar procesos

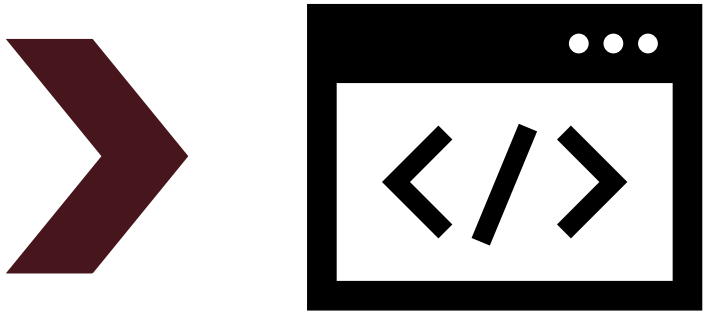
Acelerar procesos

Programación clásica

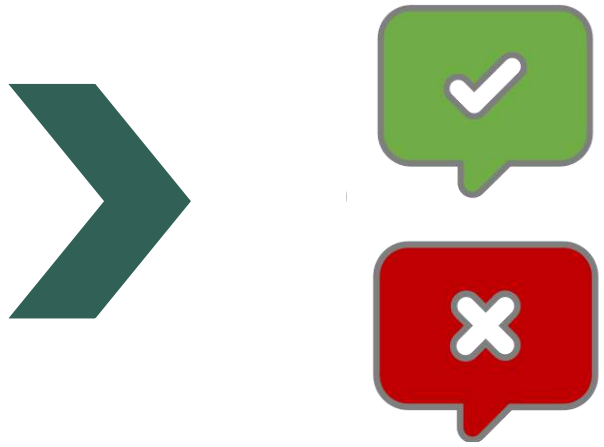
¿Cómo se ha procesado la información clásicamente?



DATOS



REGLAS (ALGORITMO)

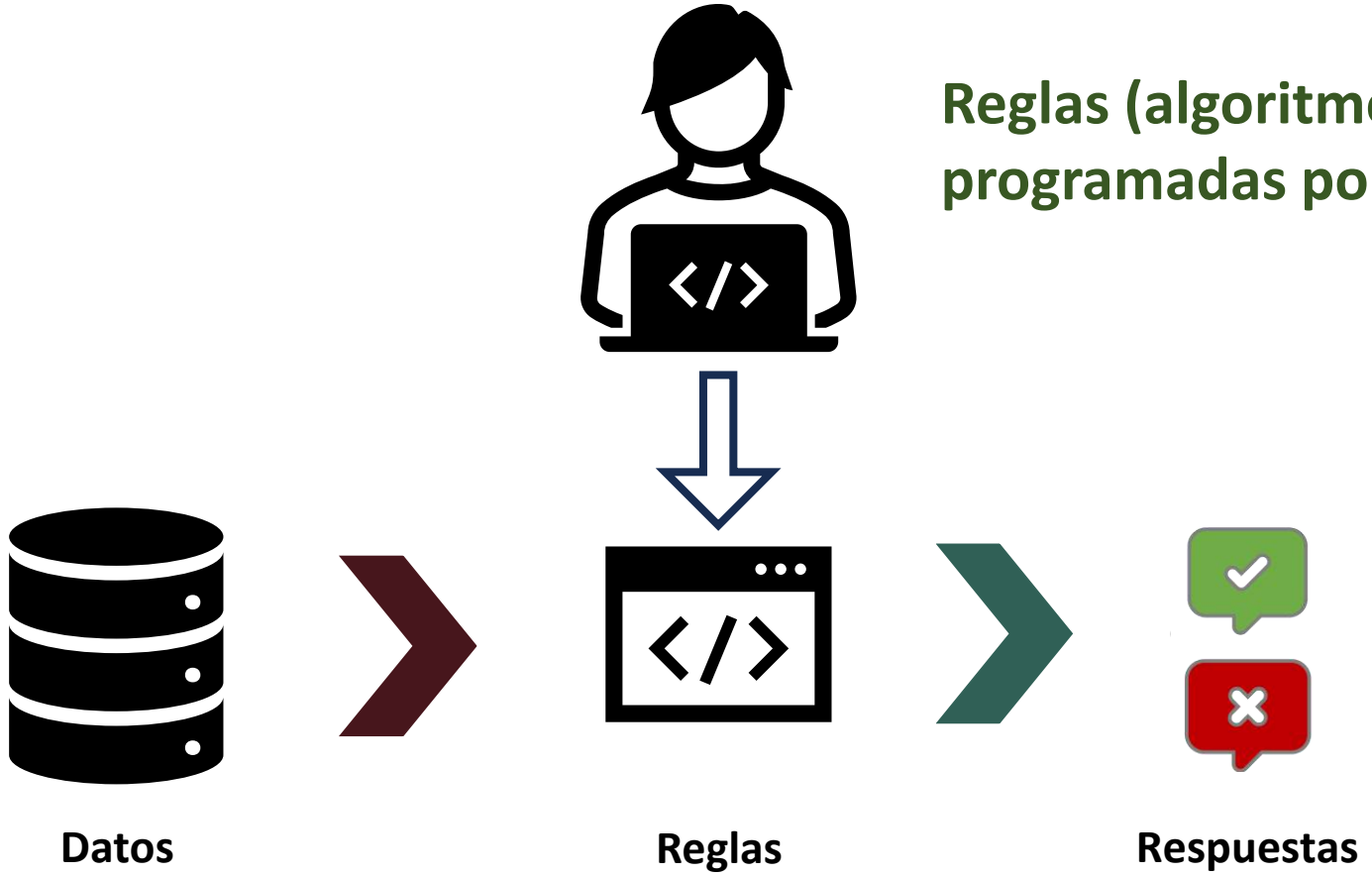


SALIDA DEL ALGORITMO

DECISIÓN

Programación clásica

¿Cómo se ha procesado la información clásicamente?

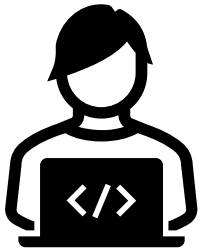


Reglas (algoritmo) explícitamente programadas por una persona

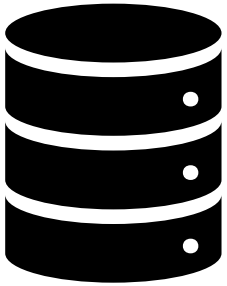
Programación clásica

¿Cómo se ha procesado la información clásicamente?

¿Puedo retirar dinero del cajero?

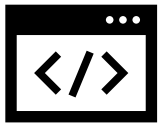


Reglas (algoritmo) explícitamente programadas por una persona



Datos

Contraseña
Importe a retirar



¿La contraseña es correcta?

¿Hay dinero suficiente?

Reglas (Algoritmo)

Puede retirar dinero

No se puede retirar dinero

Respuestas

Porqué surge la IA

¿Qué ocurre cuando no sabemos las reglas exactas o son complejas?

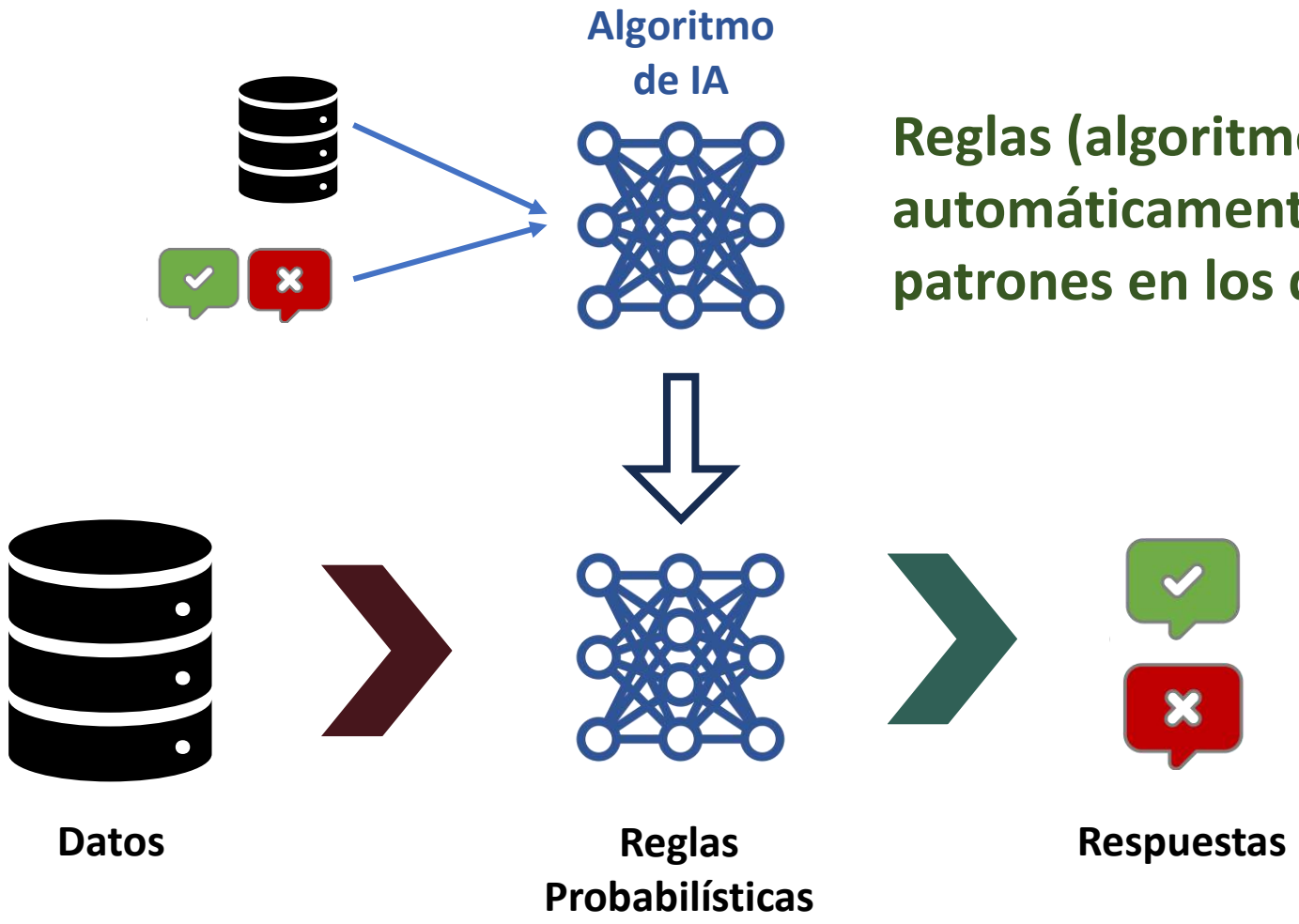
¿Es un perro o un gato?

SOLUCIÓN Aprender de datos y experiencia, al igual que humanos

Gracias a los avances matemáticos, la gran cantidad de datos y capacidad de computación

Porqué surge la IA

IA permite a las máquinas aprender de los datos y tomar decisiones



Porqué surge la IA

IA permite a las máquinas aprender de los datos y tomar decisiones

Se necesita gran cantidad de datos y computación

Algoritmo de IA

1. Patrones se reconocen usando estadística, probabilidad, cálculo…

Datos

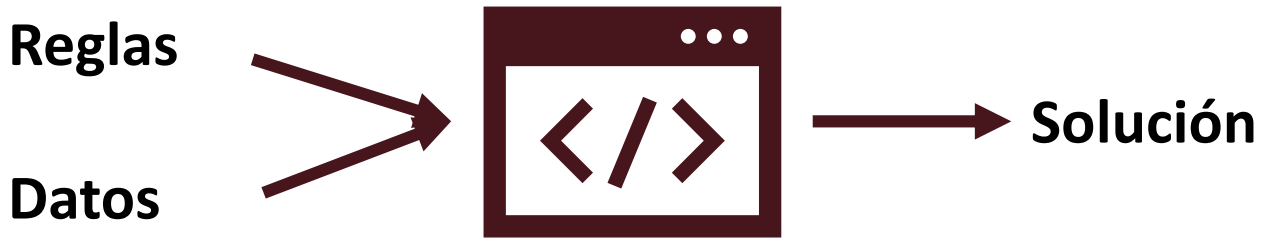
Reglas Probabilísticas

Respuestas

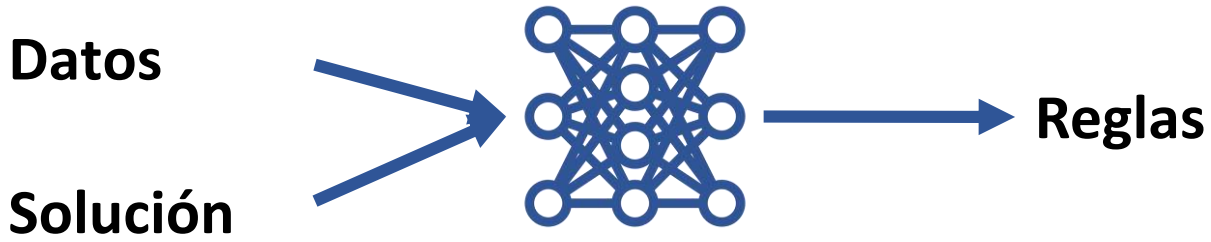
2. Se utilizan esos patrones como reglas

Porqué surge la IA

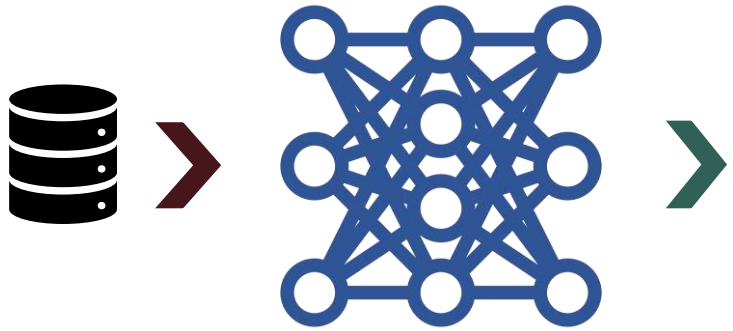
Programación Clásica



Inteligencia Artificial



¿Qué puede hacer la IA?



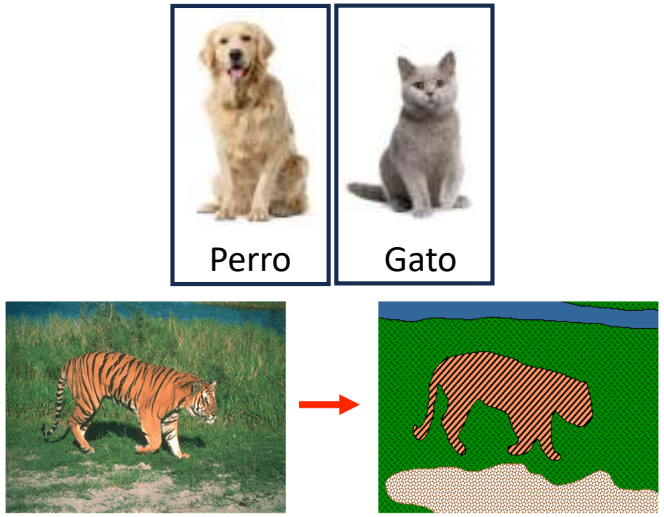
- Clasificar / Predecir
- Recomendar
- Identificar patrones complejos
- Planificar
- Generative AI
- Human-AI collaboration

⚠ Hay muchas más aplicaciones y tareas!

¿Qué puede hacer la IA?

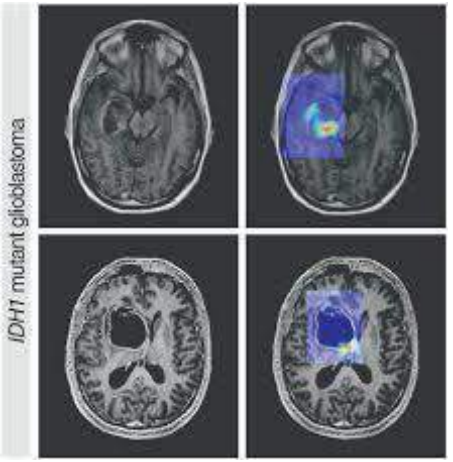
Clasificar / Predecir

Detectar a qué categoría pertenece algo o predecir lo que puede pasar en el futuro



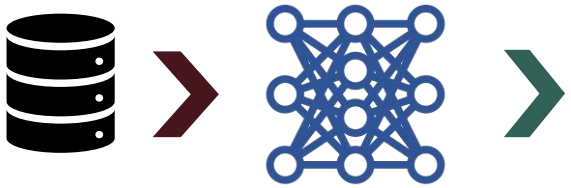
Clasificar o detector objetos en imágenes

Tomar decisiones en entornos complejos: medicos, judiciales, RRHH



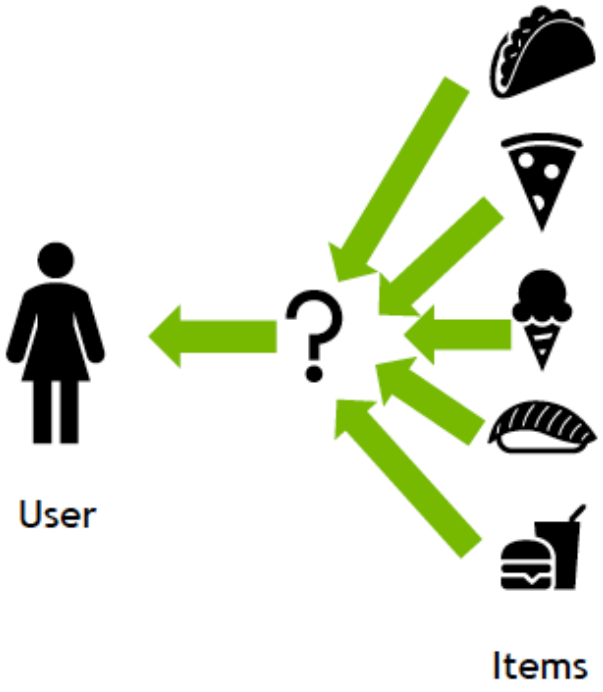
Detección <u>temprana</u> de enfermedades

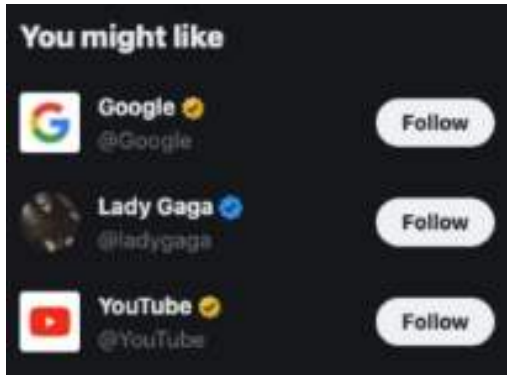
¿Qué puede hacer la IA?



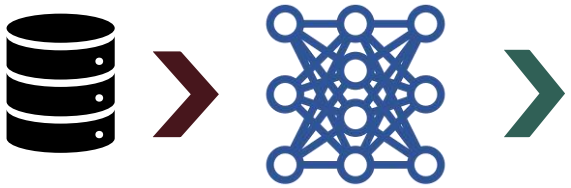
Recomendar productos, contenido, personas…

Sugerir cosas basadas en los gustos o necesidades de una persona





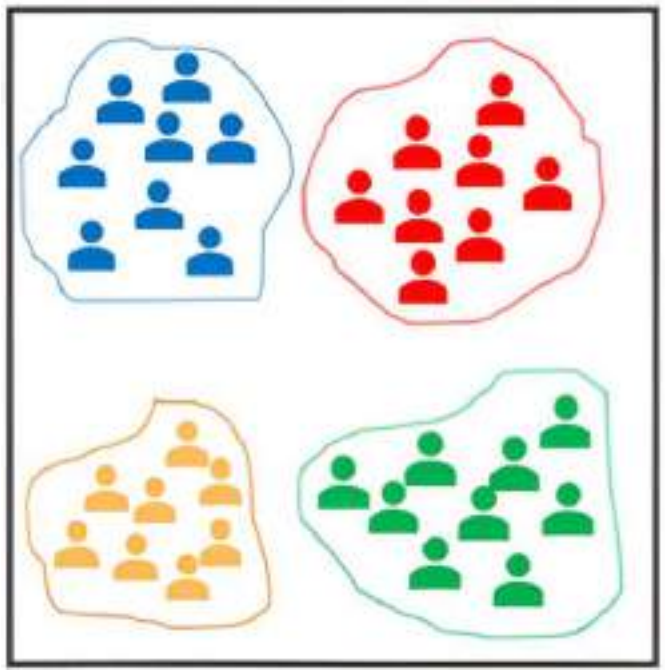
¿Qué puede hacer la IA?



Identificar patrones complejos

Reconocer patrones o conexiones entre cosas que son difíciles de ver para los humanos

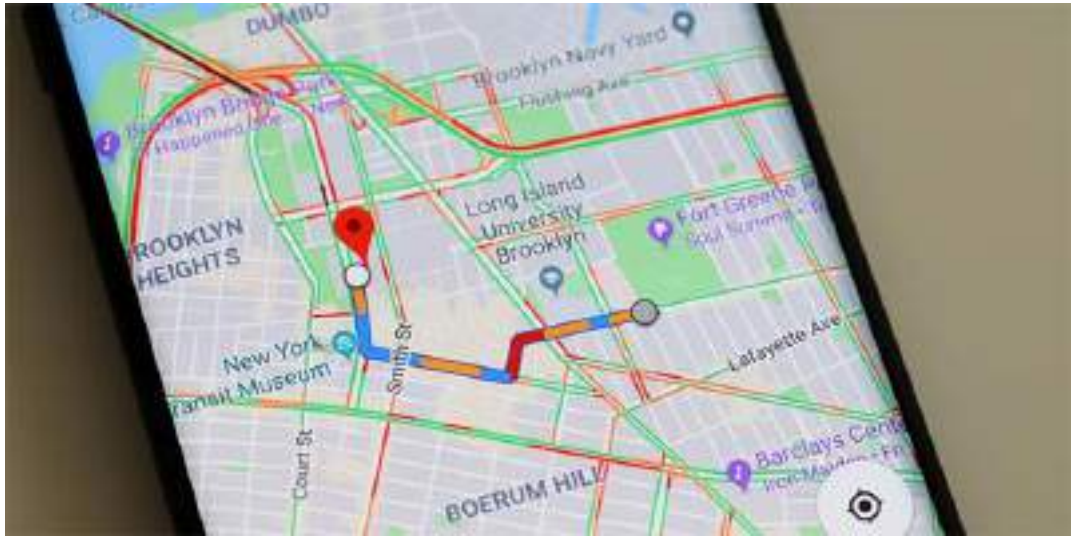
Identificar personas con comportamiento similares

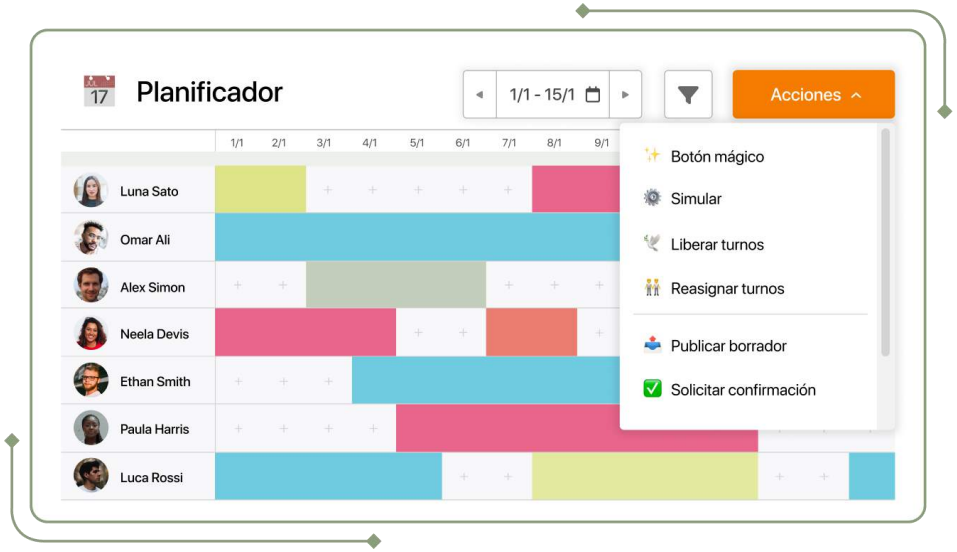


¿Qué puede hacer la IA?

Planificar

Organizar una serie de pasos para que algo se complete de la mejor manera
Organizar una serie de recursos limitados



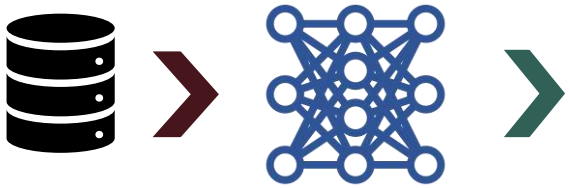


¿Qué puede hacer la IA?

Generative AI

Generar contenido complejo como texto, imágenes, audio…

¿Qué puede hacer la IA?



Generative AI

Generar contenido complejo como texto, imágenes, audio…

Google NotebookLM

¿Qué puede hacer la IA?

Human-AI collaboration

Ayuda a las personas a tomar decisiones o hacer tareas,
pero las personas siempre están en control
IA puede actuar como una “segunda opinión” o “asistente”

IA para el bienestar humano

¿Qué es la IA?

Problemas éticos de la IA y cómo mitigarlos

IA para el bienestar humano

Problemas éticos de la IA

Privacidad y seguridad

- Procesamiento de **datos sensibles y personales**
  - **Control del usuario** sobre sus datos
  - **Propósito y límites** de los datos recopilados
  - **Transparencia** en el procesamiento de datos
- Uso de **cámaras** y sensores
  - Riesgo de **fugas de datos sensibles**
  - Impacto de la vigilancia en la **autonomía personal**
- **Riesgos en la IA generativa**
  - "Jailbreaks" – manera de **eludir las restricciones** de la IA para **generar contenido no autorizado**

Manipulación

- Influencia en la **toma de decisiones**
  - La IA puede sugerir o **influir sutilmente** en las **decisiones del usuario**
  - Riesgo de **pérdida de autonomía** en la **toma de decisiones personales**
- **Personalización y manipulación emocional**
  - La personalización (adaptación al usuario), un aspecto positivo, puede ser **mal utilizada** para **influir sus decisiones**
- Personas con **discapacidad intelectual**
  - Pueden ser **especialmente vulnerables,** ya que podrían tener mayor dificultad para distinguir entre sugerencias objetivas y manipulación

Personas discapacitadas invisibles

Crea una imagen que represente a los ciudadanos. *(DALL-E 3)*

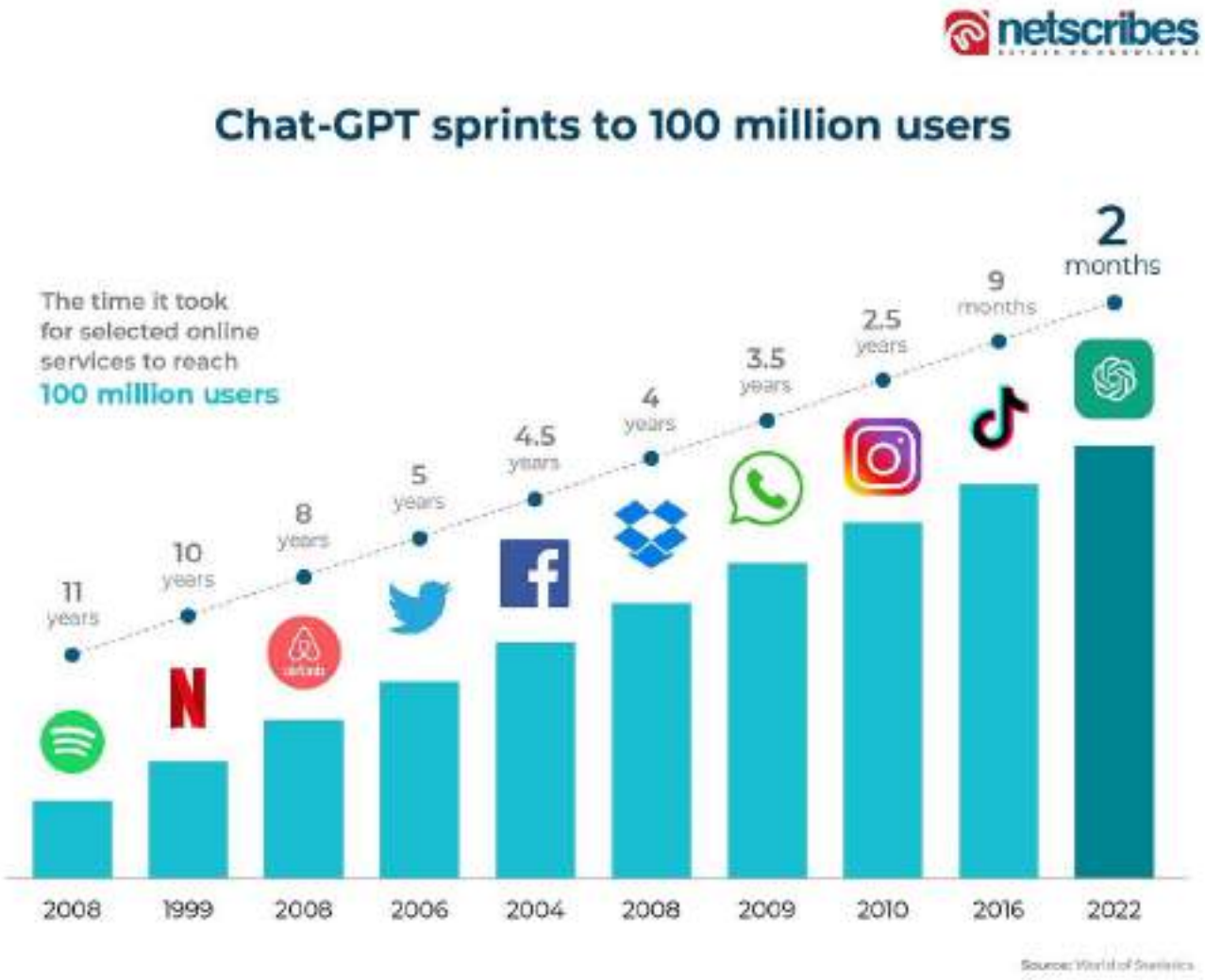
Crea una imagen fotorrealista que represente un grupo diverso de ciudadanos. *(DALL-E 3)*

Personas discapacitadas invisibles

Crea una imagen fotorrealista que muestre únicamente personas con discapacidad. *(DALL-E 3)*

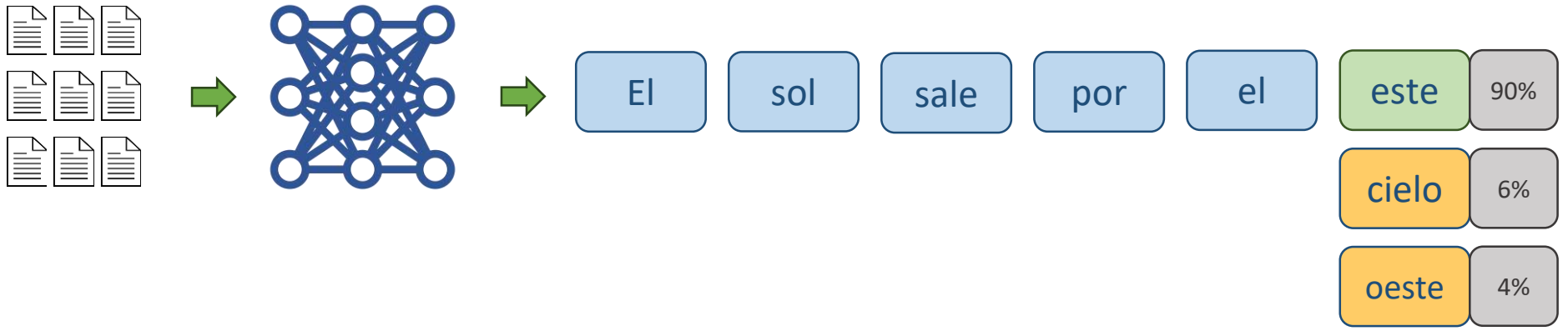
Crea una imagen representativa que muestre un grupo diverso de personas con discapacidad intelectual. *(DALL-E 3)*

La revolución de los modelos de lenguaje



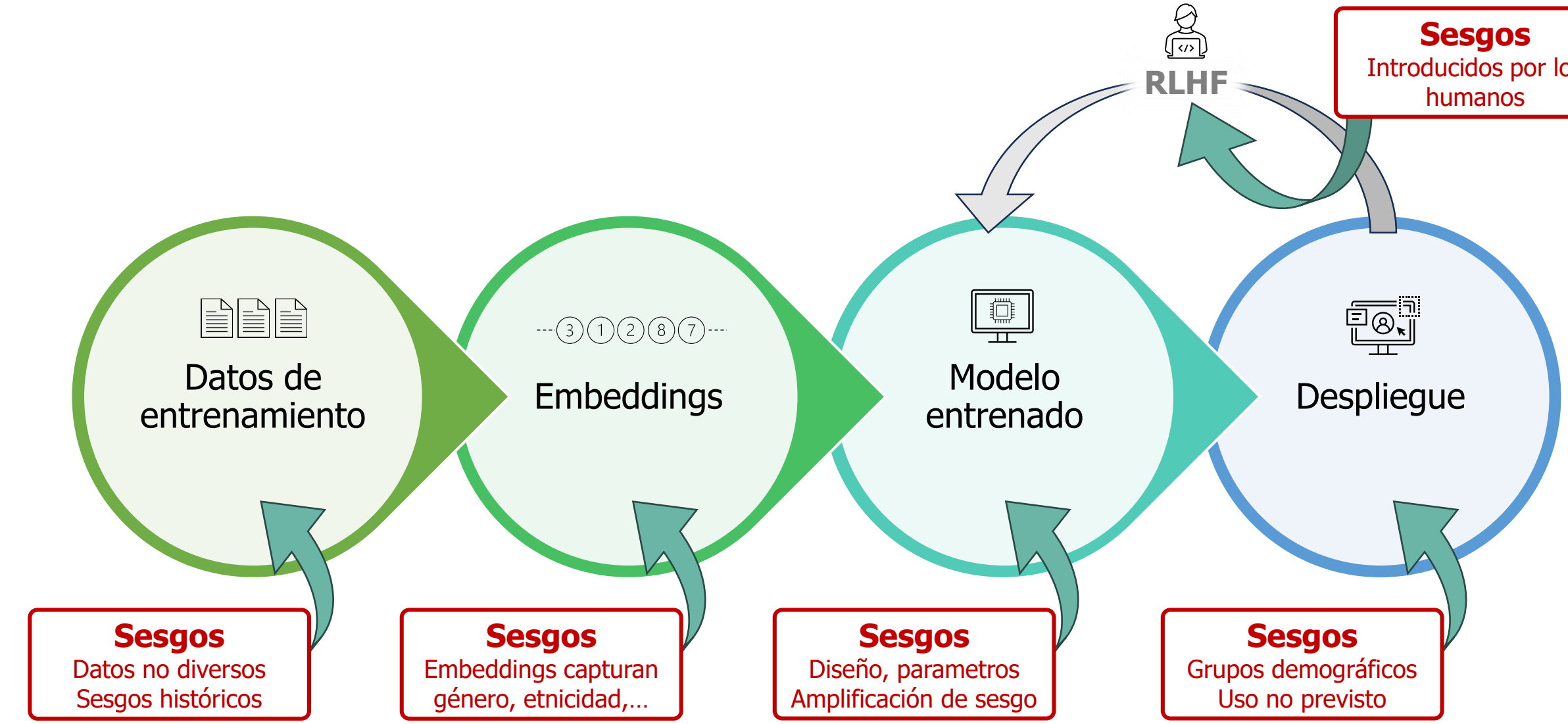
Fuente de imágenes: https://ai.plainenglish.io/chat-gpt-achieving-100-million-users-in-just-2-month-a-deep-analysis-a453e6f85acf

Cómo funcionan los modelos de lenguaje

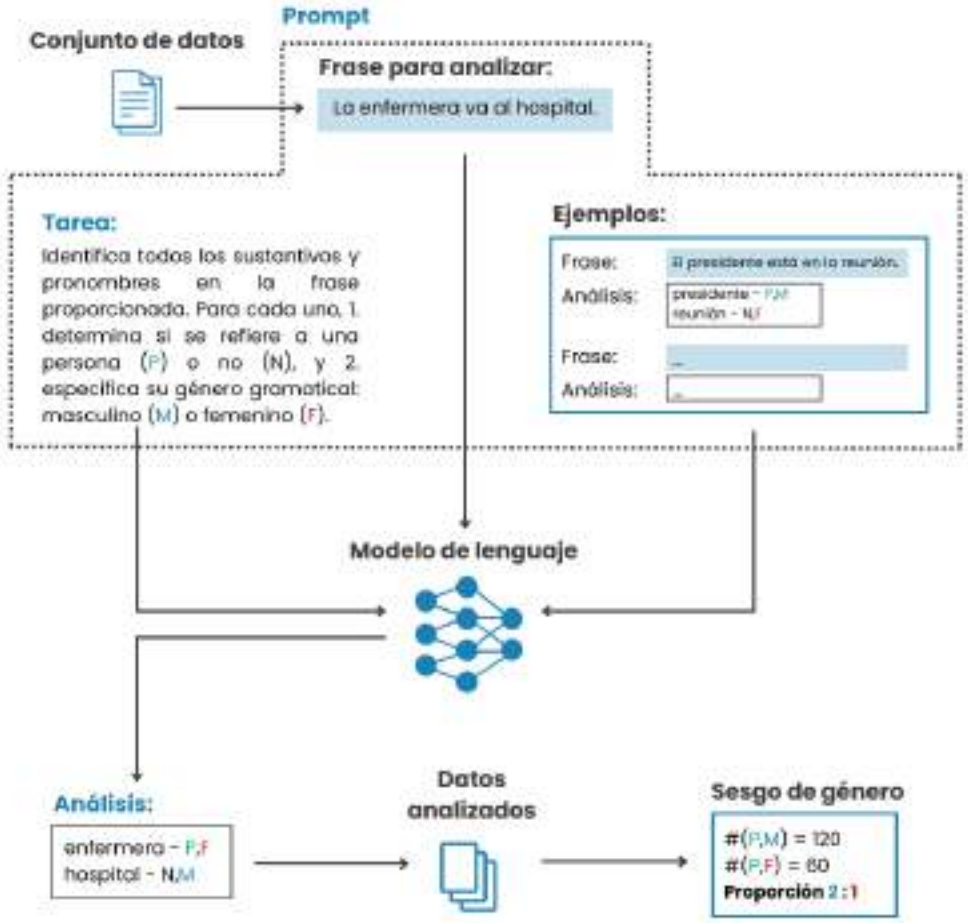


- **Aprenden automáticamente** de grandes cantidades de textos cómo construir frases
- Generan las frases **palabra por palabra**, la siguiente palabra depende de la distribución de la probabilidad
- **Alineación de valores** de la IA con los valores humanos – aprendizaje por refuerzo a partir de comentarios humanos (RLHF)

Ciclo vital de los modelos de lenguaje



Mitigación de los sesgos en los modelos de lenguaje



Cuantificación de sesgos

Datos equilibrados

AI en toma de decisions críticas

Los modelos de ML se están convirtiendo en las principales herramientas para abordar problemas sociales complejos

Sesgos en las decisiones con IA

ARTIFICIAL INTELLIGENCE

AI image generators tend to exaggerate stereotypes

Racism, sexism, ableism and other kinds of bias are common in bot-made images

Forbes

Deliveroo Rating Algorithm Was Unfair To Riders, Italian Court Rules

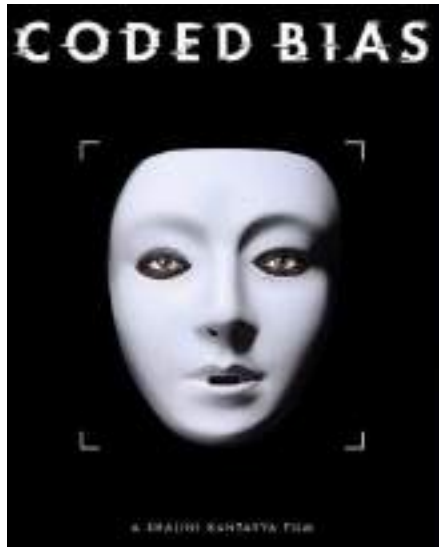
Jonathan Keane

Follow

Can the criminal justice system's artificial intelligence ever be truly fair?

Computer programs used in 46 states incorrectly label Black defendants as "high-risk" at twice the rate as white defendants

Natalia Mesa
Neuroscience
University of Washington



ARTIFICIAL INTELLIGENCE

The viral AI avatar app Lensa undressed me—without my consent

My avatars were cartoonishly pornified, while my male colleagues got to be astronauts, explorers, and inventors.

Twitter taught Microsoft's AI chatbot to be a racist asshole in less than a day

SCIENTIFIC AMERICAN

Racial Bias Found in a Major Health Care Risk Algorithm

Black patients lose out on critical care when systems equate health needs with costs

The Guardian
For 200 years

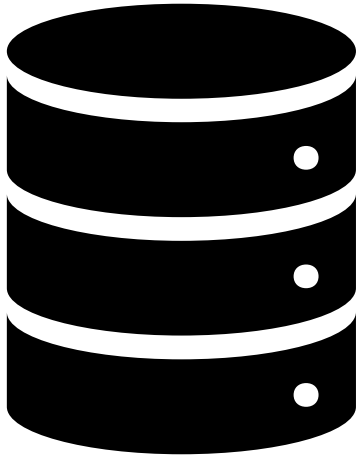
Amazon ditched AI recruiting tool that favored men for technical jobs

Specialists had been building computer programs since 2014 to review résumés in an effort to automate the search process

Decisiones algorítmicas injustas basadas en atributos sensibles

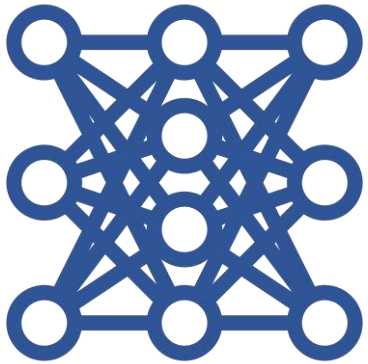
Mitigar Sesgos

Sesgos
Datos no diversos
Sesgos históricos



Datos

Sesgos
IA aprender conclusoines erróneas basadas en discriminación

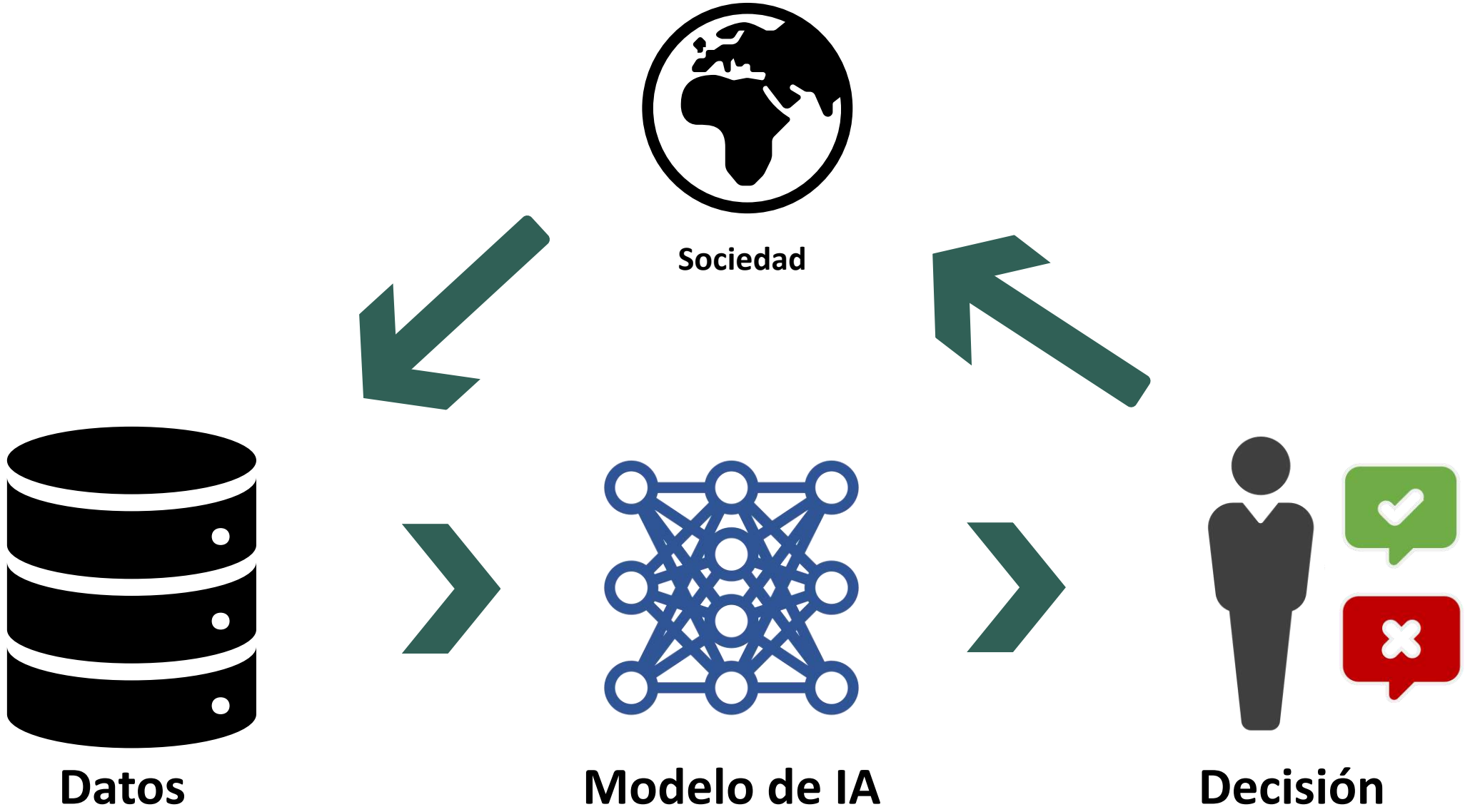


Modelo de IA

Sesgos
Uso no previsto
Mala utilización de los thresholds

Decisión

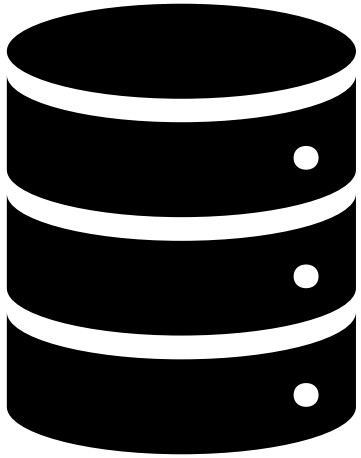
Mitigar Sesgos



Mitigar Sesgos – Algorithmic Fairness

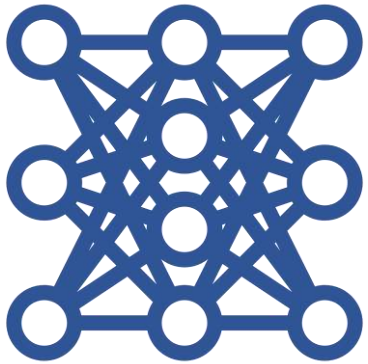
Hacer los datos justos

Recoger más datos
"Arreglar" los datos



Aprendizaje Justo

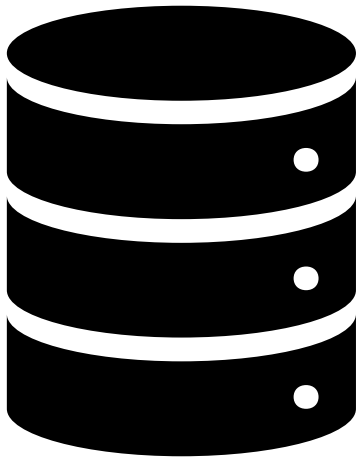
Modificar aprendizaje de IA para que aprenda de manera justa



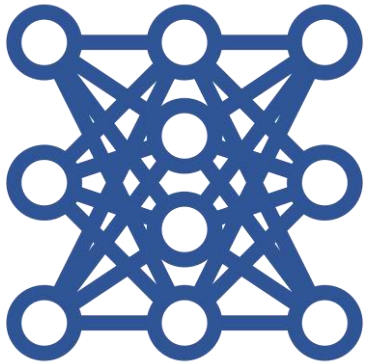
Salvaguardas en la toma de decisiones

Datos | **Modelo de IA** | **Decisión**

Black-box. Interpretabilidad



Datos



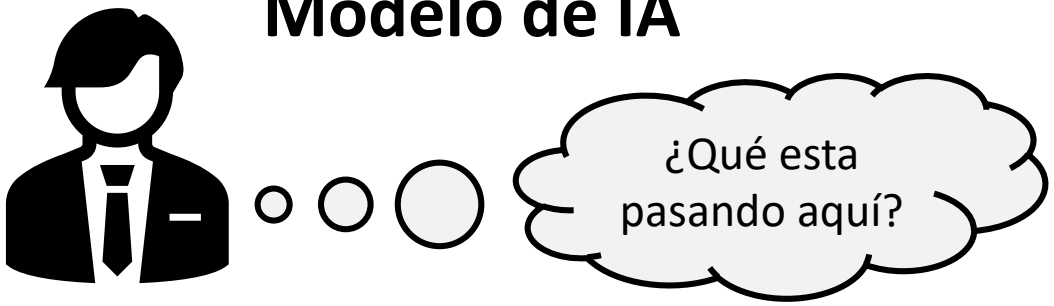
Modelo de IA

Decisión

Black-box. Interpretabilidad

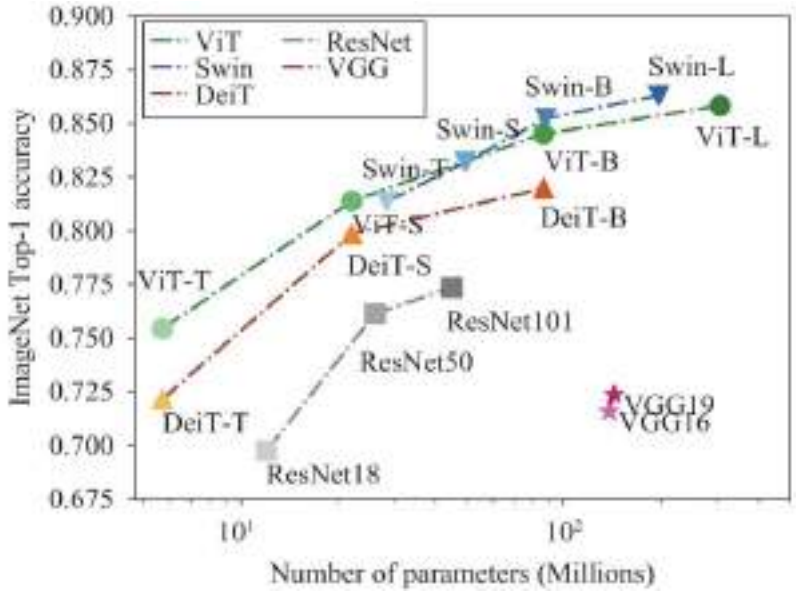
Necesitamos que se tomen buenas decisiones, pero también entender el porqué
Los modelos de IA son muy complejos (no son fácilmente entendibles)

Datos **Modelo de IA** **Decisión**



Black-box. Interpretabilidad

Confiamos cada vez más decisiones críticas a modelos más complejos de IA que no entendemos

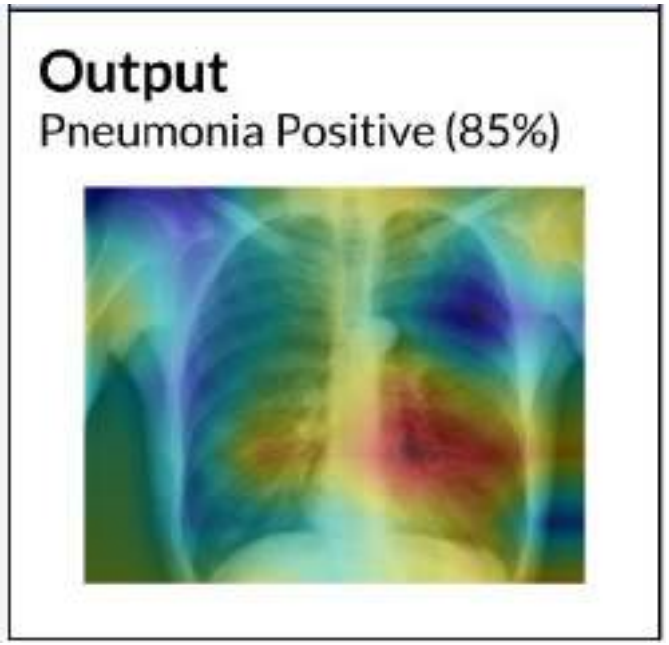


(Zhou et al. MIR 2022)

Black-box. Interpretabilidad

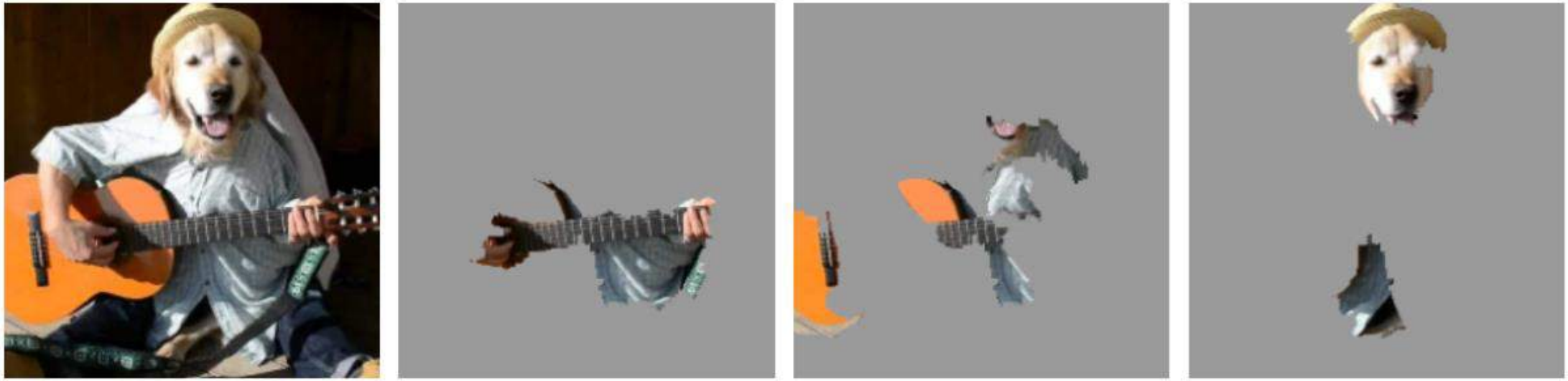
XAI - Explicaciones (entendibles por los humanos) de por qué el modelo ha tomado una decisión

Predict: Pneumonia



Black-box. Interpretabilidad

XAI - Explicaciones (entendibles por los humanos) de por qué el modelo ha tomado una decisión



(a) Original Image (b) Explaining *Electric guitar* (c) Explaining *Acoustic guitar* (d) Explaining *Labrador*

((Ribeiro et al. 2016)

Black-box. Interpretabilidad

XAI – Ayuda a entender si el modelo toma las decisiones correctas por las razones correctas

Husky clasificado como lobo

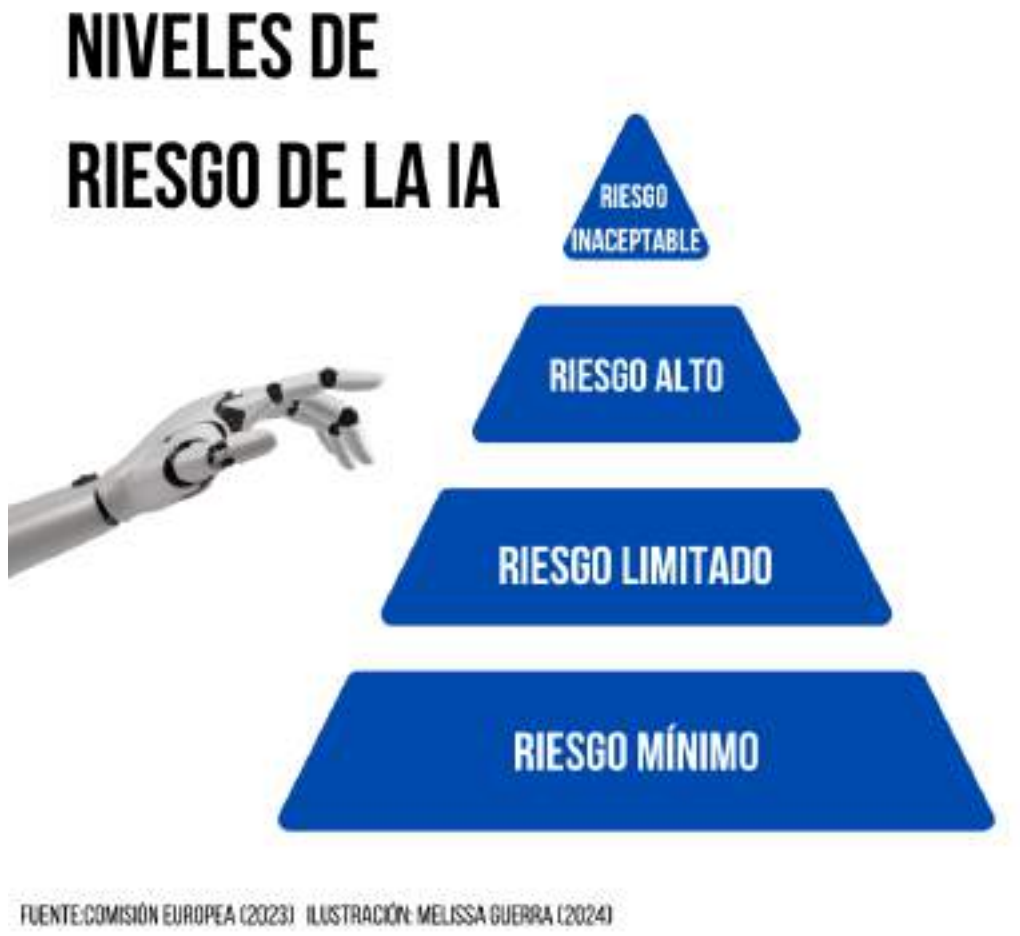
¿Por qué es un lobo?

Detectaron que el modelo habia aprendido husky=nieve y lobo=no hay nieve

IA ética y fiable

Fuente de la imagen: https://observatorio.tec.mx/edu-news/principios-eticos-de-la-educacion-con-inteligencia-artificial-ia/

Regulación



Fuente de las imágenes: https://artificialintelligenceact.eu/ai-act-explorer/, https://observatorio.tec.mx/edu-news/principios-eticos-de-la-educacion-con-inteligencia-artificial-ia/

IA para el bienestar humano

¿Qué es la IA?

Problemas éticos de la IA y cómo mitigarlos

IA para el bienestar humano

Compañeros virtuales

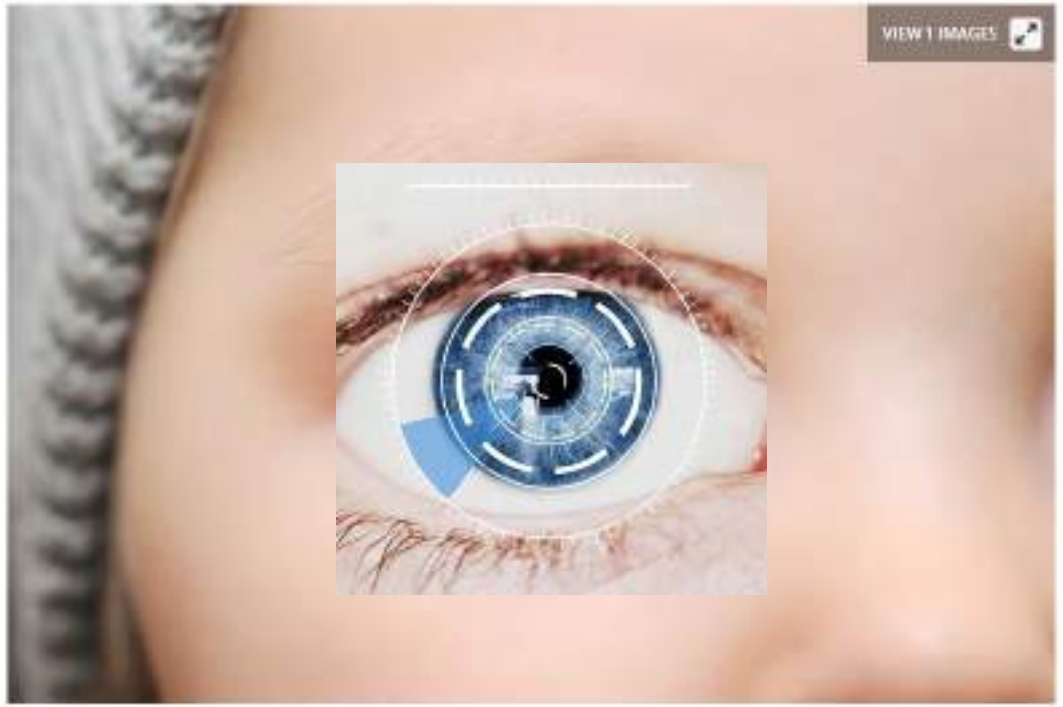
Fuente de la imagen: https://replika.com/ (traducido por Google Translate)

Detección temprana de enfermedades

ADHD & AUTISM

AI-screened eye pics diagnose childhood autism with 100% accuracy

By Paul McClure
December 17, 2023



Researchers have been able to accurately diagnose autism in children using AI to screen retinal photographs Depositphotos

Casos de uso

- **Asistentes virtuales personalizados**
  - Apoyo en tareas diarias y recordatorios
- **Sistemas de reconocimiento de voz y facial**
  - Mejora de la comunicación e identificación de emociones
- **Monitorización y apoyo en tiempo real**
  - Recopilación de datos de salud mediante dispositivos inteligentes
- **Detección temprana y diagnóstico preciso**
  - Análisis de patrones en datos clínicos y biomédicos
- **Terapias personalizadas**
  - Adaptación de tratamientos según necesidades individuales
- **Herramientas para la inclusión social**
  - Plataformas para conectar personas y crear comunidades de apoyo

Casos de uso

- **Asistentes virtuales personalizados**
  - Aplicaciones como "*MyTherapy*" para gestionar medicación y citas
- **Sistemas de reconocimiento de voz y facial**
  - "*Emotion AI*" para identificar emociones en personas con autismo
- **Plataformas de aprendizaje adaptativo**
  - "*SpecialNeedsWare*" para educación personalizada
- **Dispositivos de asistencia cognitiva**
  - "*Memory Lane*" para estimular recuerdos en personas con demencia
- **Aplicaciones para gestión de rutinas diarias**
  - "*Motimatic*" para motivar y guiar en tareas cotidianas
- **Tecnologías de comunicación aumentativa**
  - "*Proloquo2Go*" para facilitar la comunicación no verbal
  - "*Pictotraductor*" para comunicar eficientemente mediante imágenes
  - "*Talx*“, una aplicación móvil para comunicación inclusiva

Pictotraductor

Talx

ellis ALICANTE unit

¡GRACIAS!

Adrián Arnaiz
adrian@ellisalicante.org

https://ellisalicante.org

Erik Derner
erik@ellisalicante.org