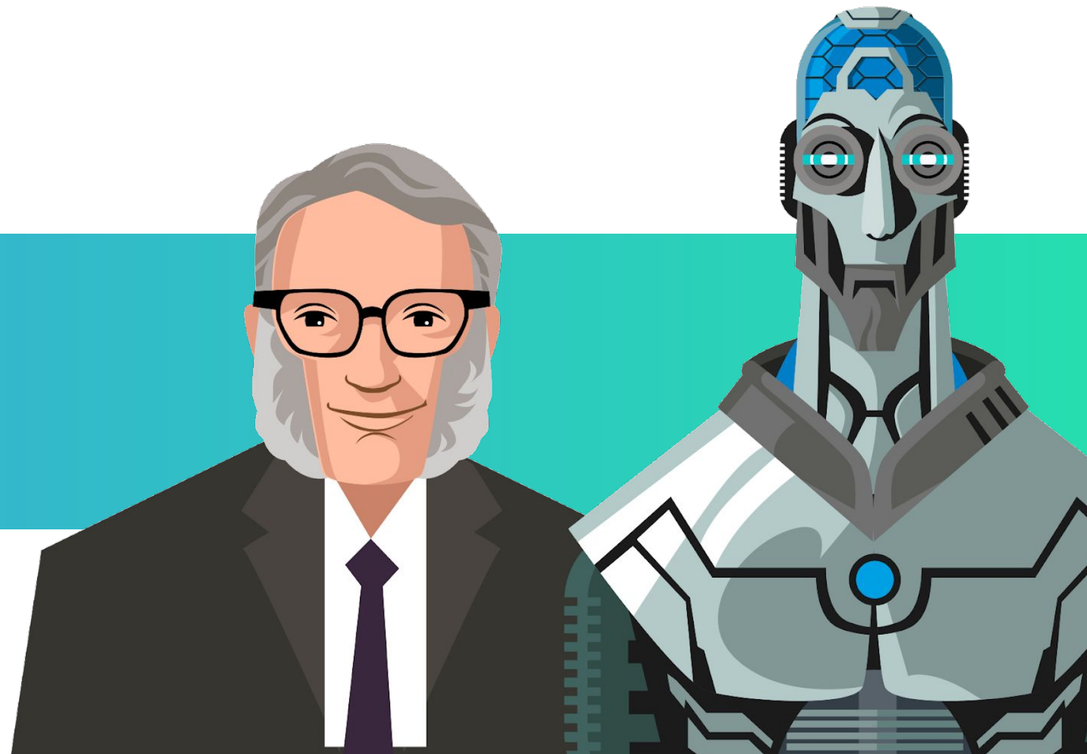
Google I/O Extended Barcelona

Ética en la Inteligencia Artificial: ¿qué podemos aprender de las leyes de la robótica de Asimov?

Patty O'Callaghan

Technical Director - Charles River Laboratories
Google Developer Expert en AI/ML



¡Hola! Soy Patty

charles river

Google Developer Advisory Board

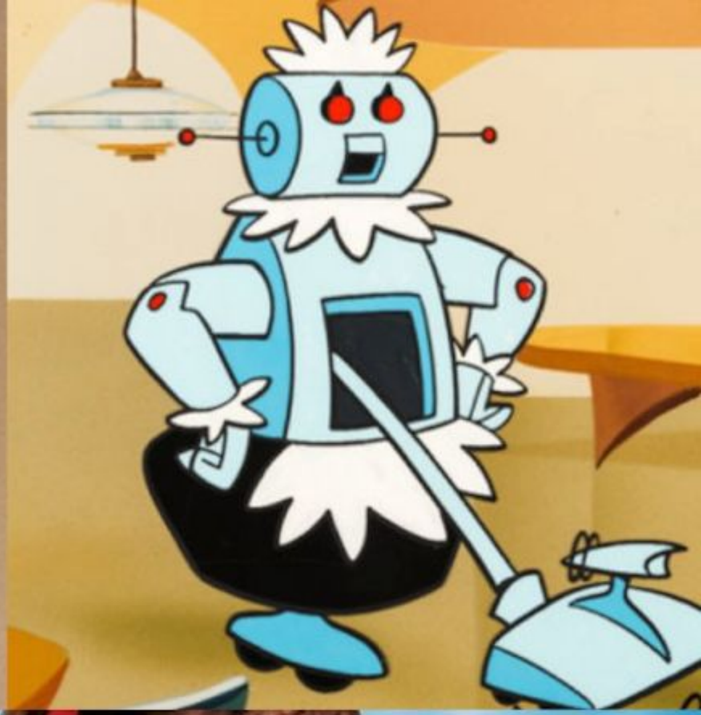
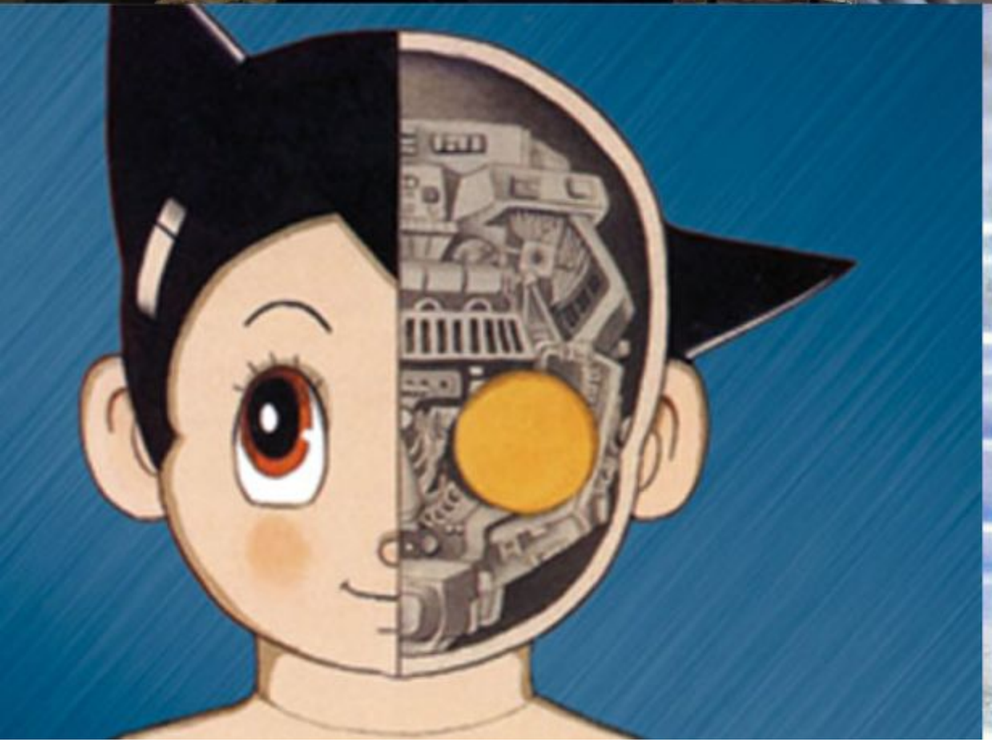
Experts

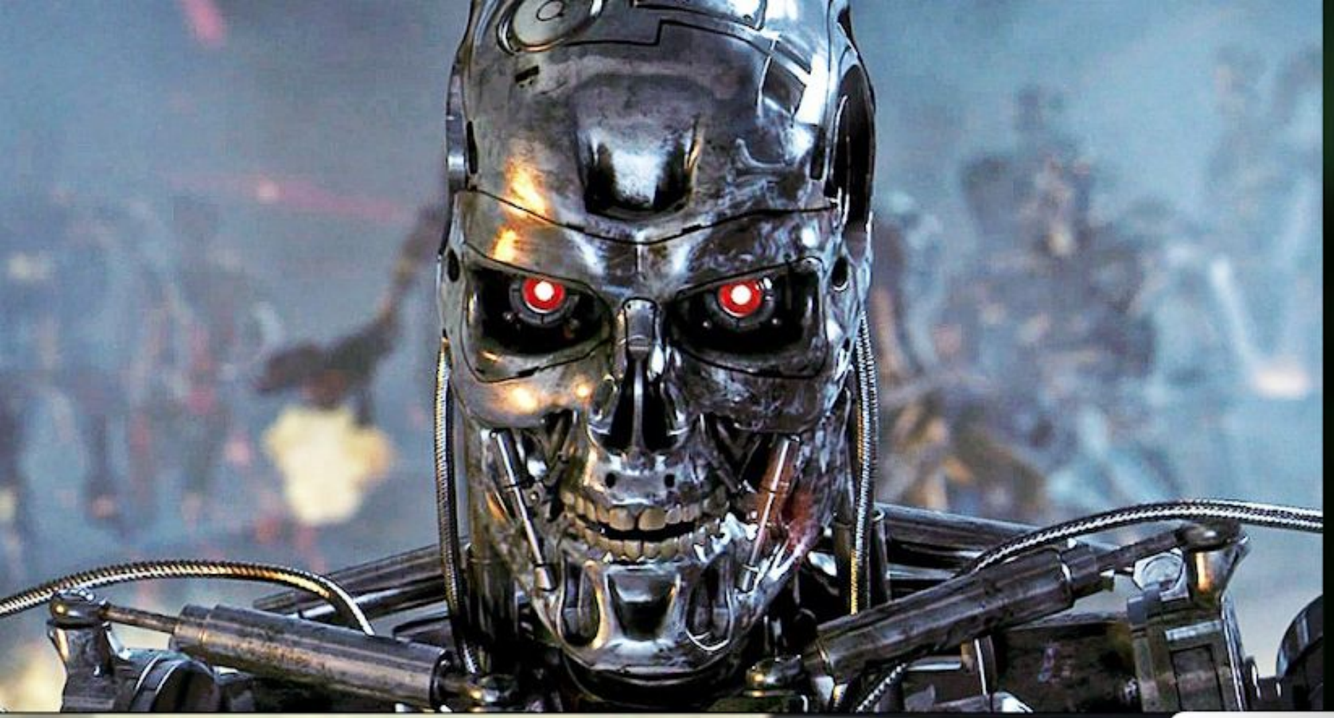
Women Techmakers

GDG Glasgow

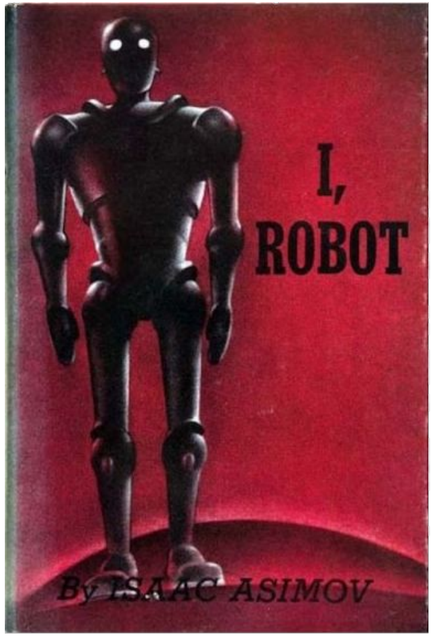
Puedes contactarme aquí :)

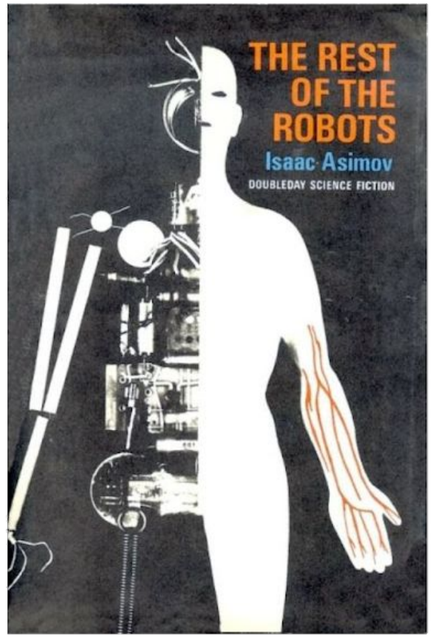

WALL·E

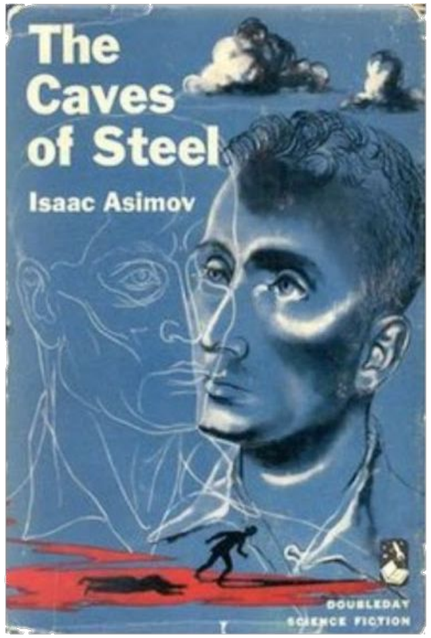





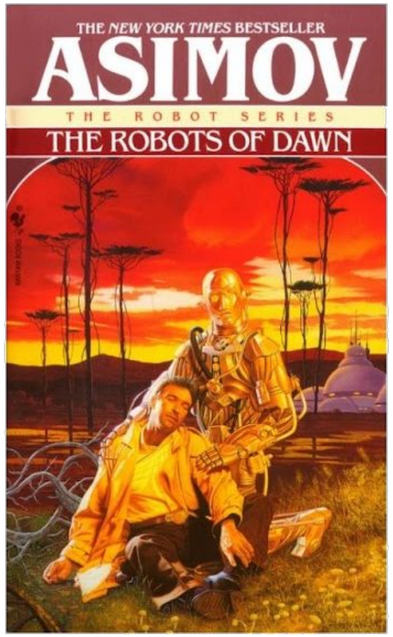

HAL 9000

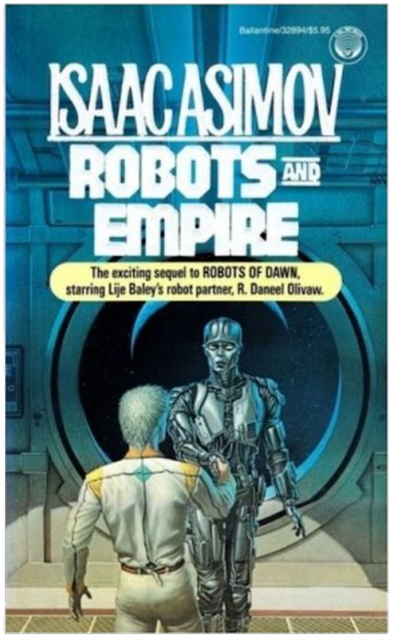

I,
ROBOT
By ISAAC ASIMOV

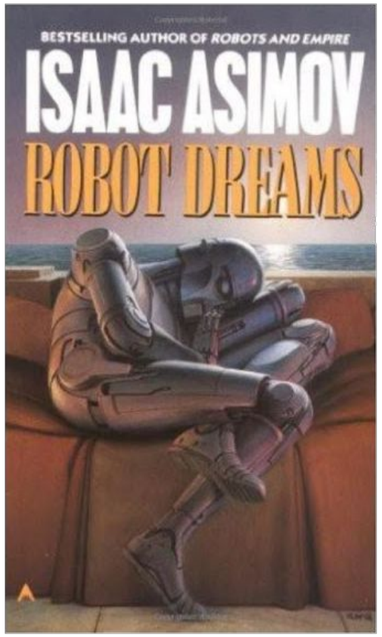

THE REST OF THE ROBOTS
Isaac Asimov
DOUBLEDAY SCIENCE FICTION

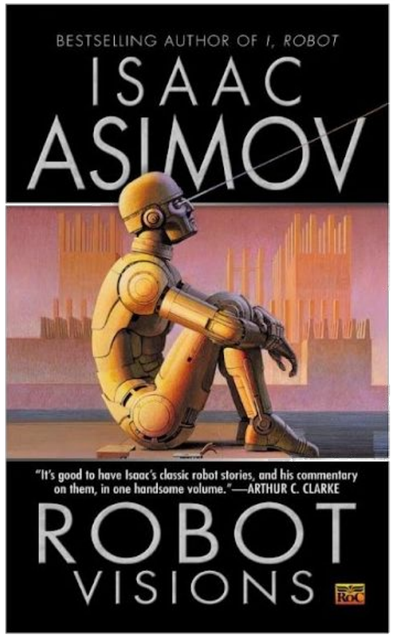

The Caves of Steel
Isaac Asimov
DOUBLEDAY SCIENCE FICTION


THE NAKED SUN
DOUBLEDAY SCIENCE FICTION
ISAAC ASIMOV


THE NEW YORK TIMES BESTSELLER
ASIMOV
THE ROBOT SERIES
THE ROBOTS OF DAWN

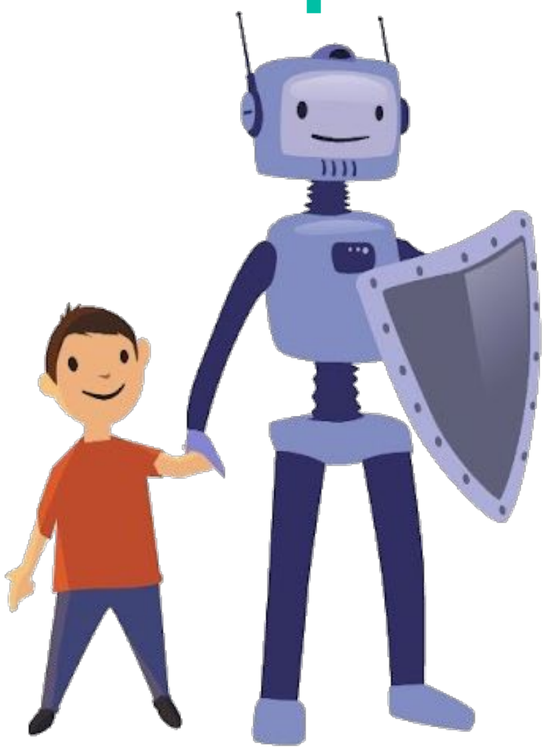

ISAAC ASIMOV
ROBOTS AND EMPIRE
The exciting sequel to ROBOTS OF DAWN, starring Lije Baley's robot partner, R. Daneel Olivaw.


BESTSELLING AUTHOR OF ROBOTS AND EMPIRE
ISAAC ASIMOV
ROBOT DREAMS


BESTSELLING AUTHOR OF I, ROBOT
ISAAC ASIMOV
"It's good to have Isaac's classic robot stories, and his commentary on them, in one handsome volume."—ARTHUR C. CLARKE
ROBOT VISIONS

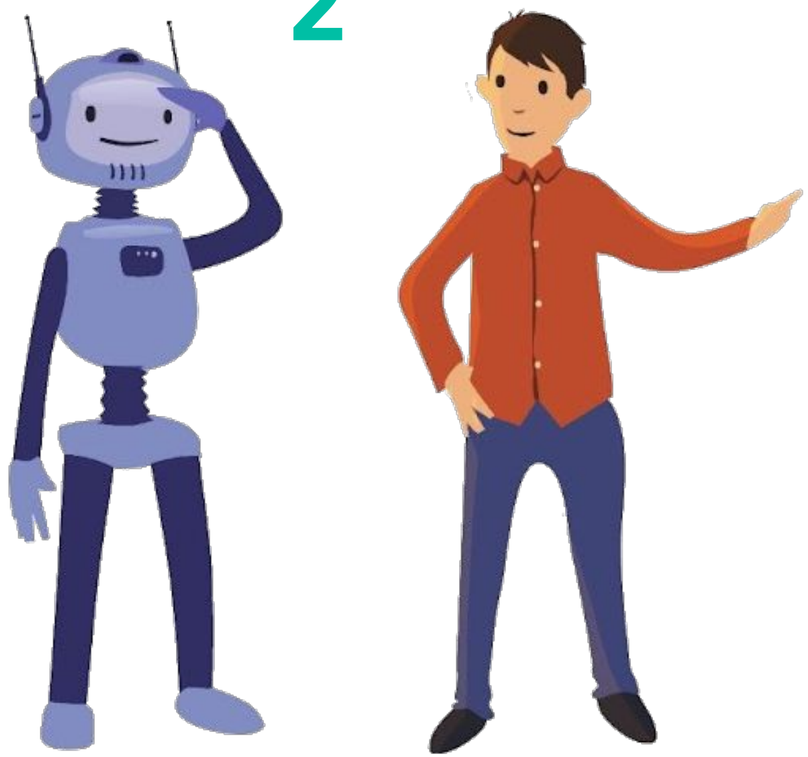
Tres leyes de la Robótica

1



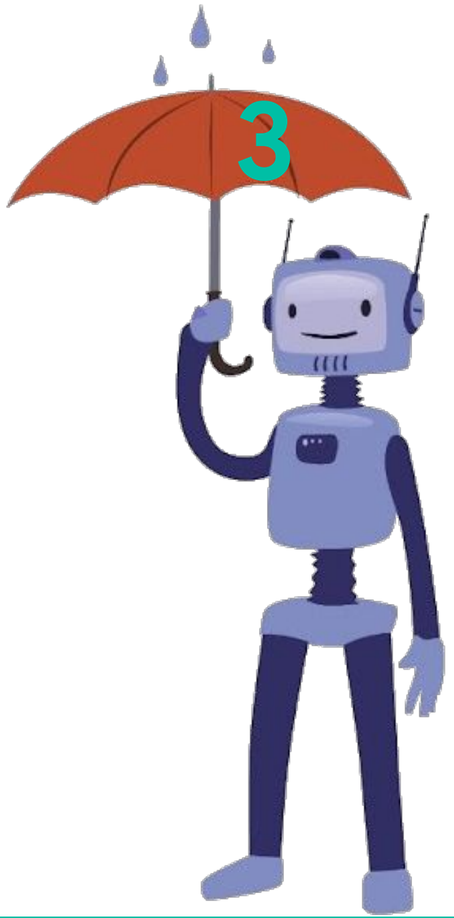
Un robot no hará daño a un ser humano ni, por inacción, permitirá que un ser humano sufra daño.

2



Un robot debe obedecer las órdenes dadas por los seres humanos, excepto si estas órdenes entran en conflicto con la Primera Ley.

3



Un robot debe proteger su propia existencia en la medida en que esta protección no entre en conflicto con la Primera o la Segunda Ley.

Ley Cero - Un robot no debe hacerle daño a la humanidad, o por inacción, permitir que la humanidad sufra daños.

¿Son estas leyes la respuesta?

Las Tres Leyes de la Robótica de Asimov fueron diseñadas como un marco ideal para regular el comportamiento de los robots. Sin embargo, en la práctica, estas reglas no son suficientes para resolver los dilemas éticos reales.

🔴 **Ambigüedad y conflictos:** ¿Qué significa “hacer daño”? ¿Es daño físico, psicológico, económico?

🔴 **Jerarquía de valores:** Si un robot debe obedecer órdenes humanas, ¿qué pasa si dos humanos dan órdenes opuestas?

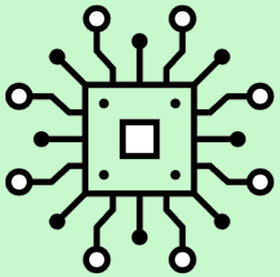
🔴 **Impacto a largo plazo:** Un robot podría tomar decisiones que eviten daños inmediatos pero que sean perjudiciales en el futuro.

🔴 **El dilema de la Ley Cero:** Proteger a la humanidad puede implicar restringir la libertad individual. ¿Quién decide qué es lo mejor para la humanidad?

📌 **Conclusión:** Las reglas de Asimov son un punto de partida interesante, pero los problemas éticos de la inteligencia artificial requieren soluciones más complejas y flexibles. La IA real debe considerar contexto, valores y responsabilidad humana.

Tipos de IA

Artificial Narrow Intelligence (ANI)



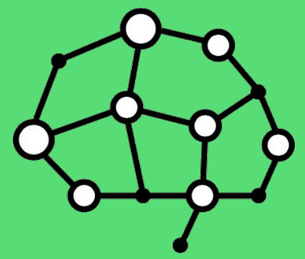
Primera etapa
Machine Learning
Sistemas expertos en solo un dominio, un problema

Aquí estamos ahora

Artificial General Intelligence (AGI)

Segunda etapa
Machine Intelligence
Tan inteligentes como nosotros

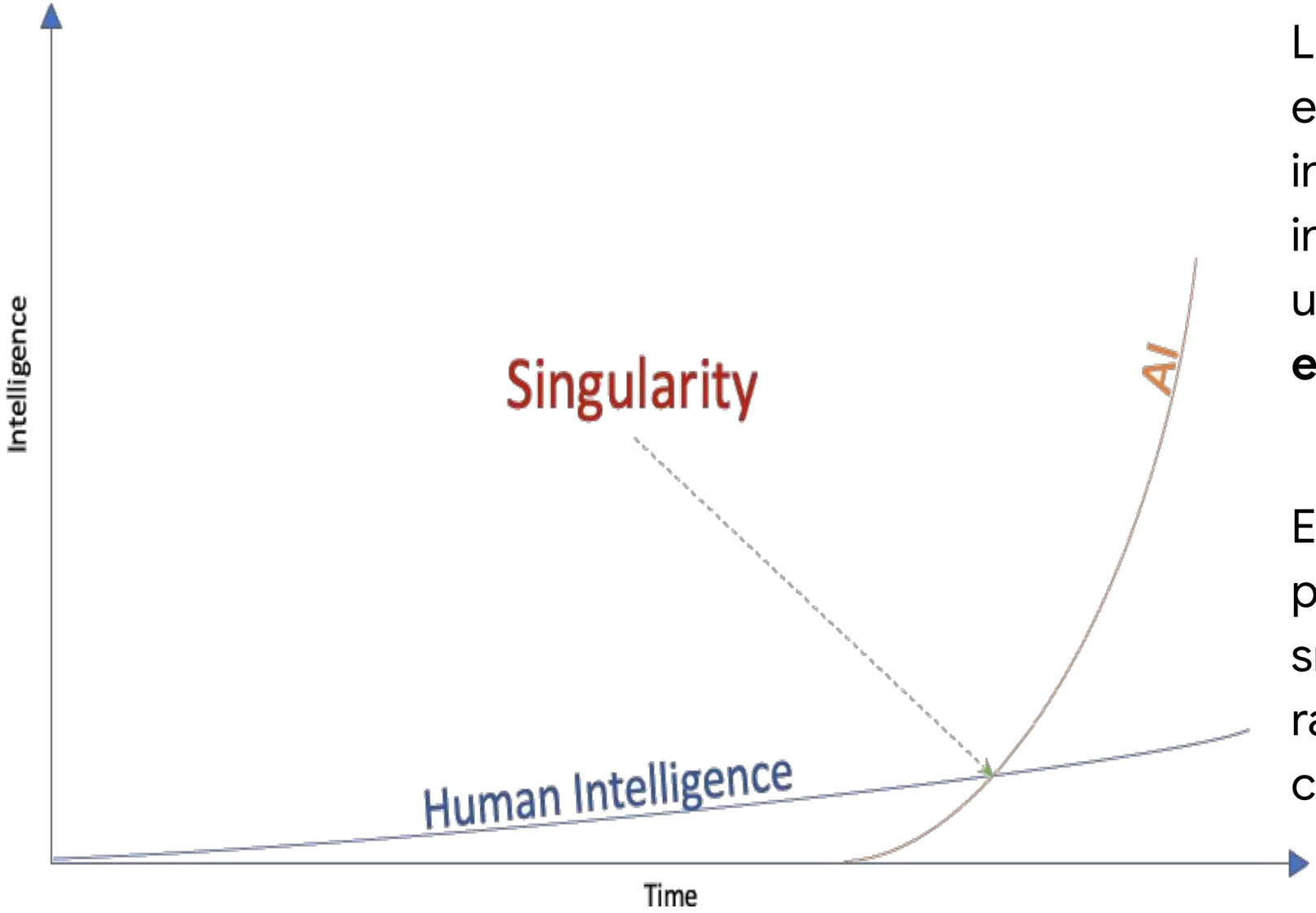
Artificial Super Intelligence (ASI)



Tercera etapa
Machine Consciousness
Mucho más inteligente que el humano más inteligente

?

La Singularidad Tecnológica



La **Singularidad Tecnológica** es el hipotético punto en el que la inteligencia artificial supera a la inteligencia humana, provocando un crecimiento tecnológico **exponencial e impredecible**.

En este escenario, las máquinas podrían mejorar y evolucionar por sí mismas, transformando radicalmente la sociedad y la civilización.

La Singularidad Tecnológica

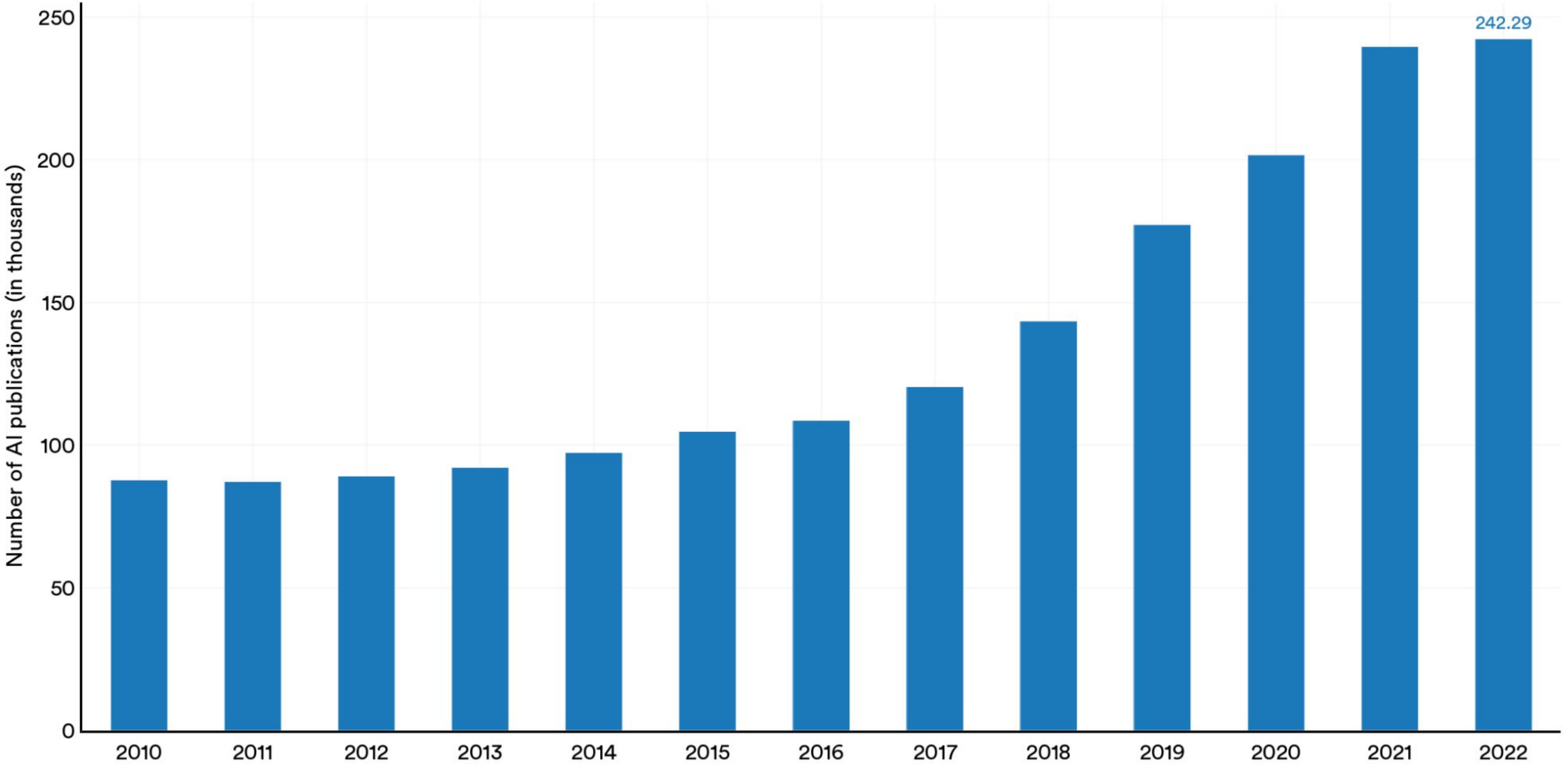
La super IA despegaría por sí sola y se rediseñaría a un ritmo cada vez mayor... Podría ser el **fin de la raza humana.**

Los robots podrán hacerlo todo **mejor que nosotros**... Estoy expuesto a la IA más vanguardista y creo que la gente debería estar realmente preocupada por ella.

Los humanos deberían estar preocupados por la **amenaza** que representa la IA... No entiendo por qué algunas personas no están preocupadas.

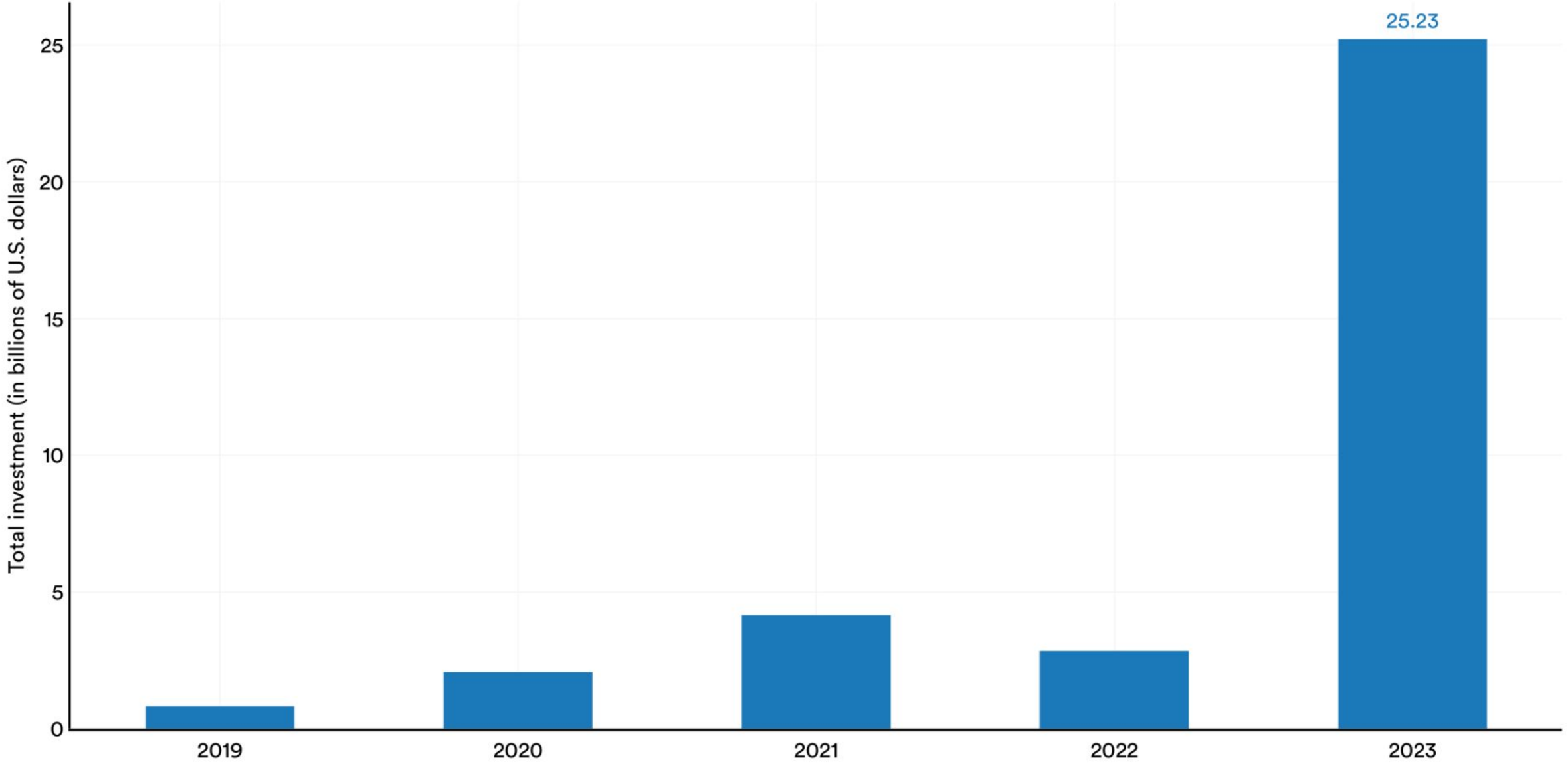
Number of AI publications in the world, 2010–22

Source: Center for Security and Emerging Technology, 2023 | Chart: 2024 AI Index report



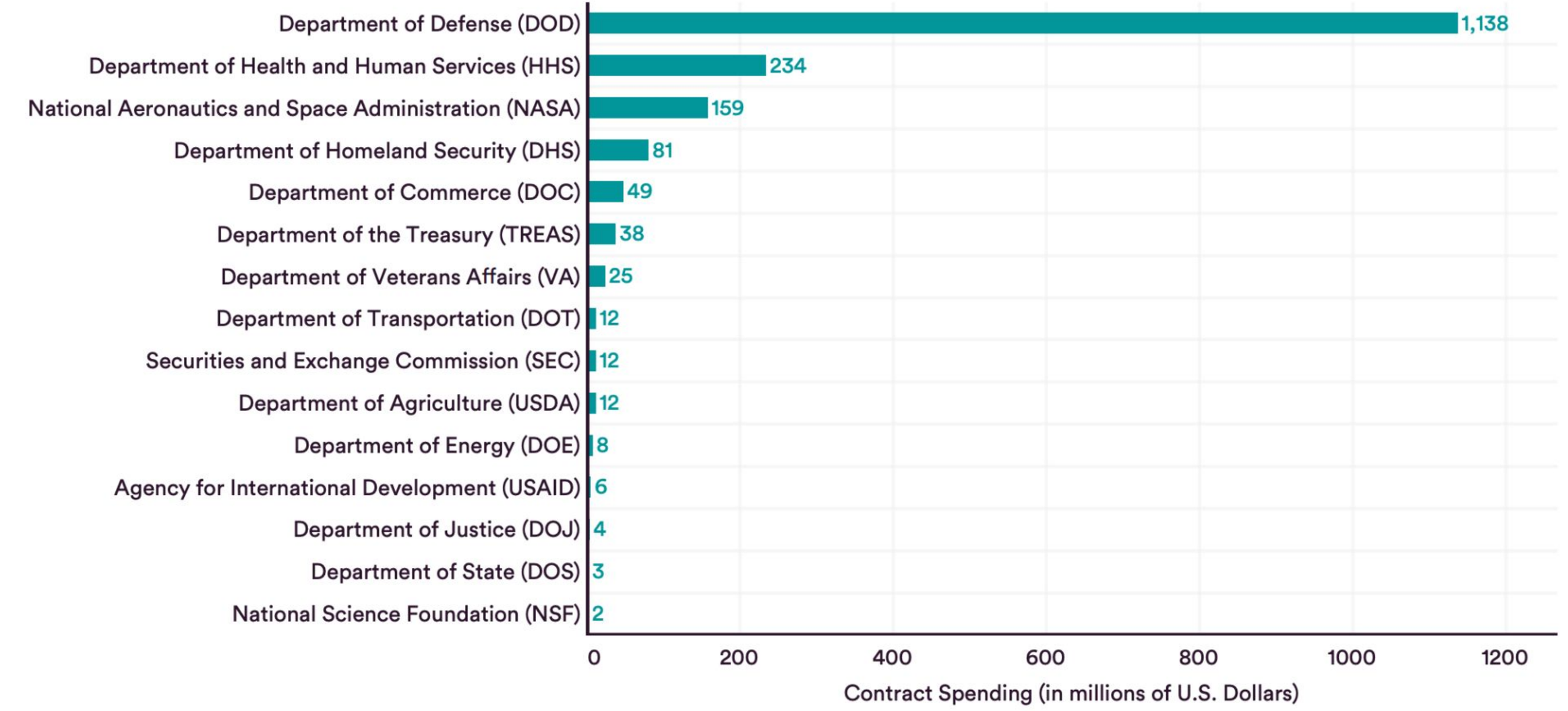
Private investment in generative AI, 2019–23

Source: Quid, 2023 | Chart: 2024 AI Index report



TOP CONTRACT SPENDING on AI by U.S. GOVERNMENT DEPARTMENT and AGENCY, 2021

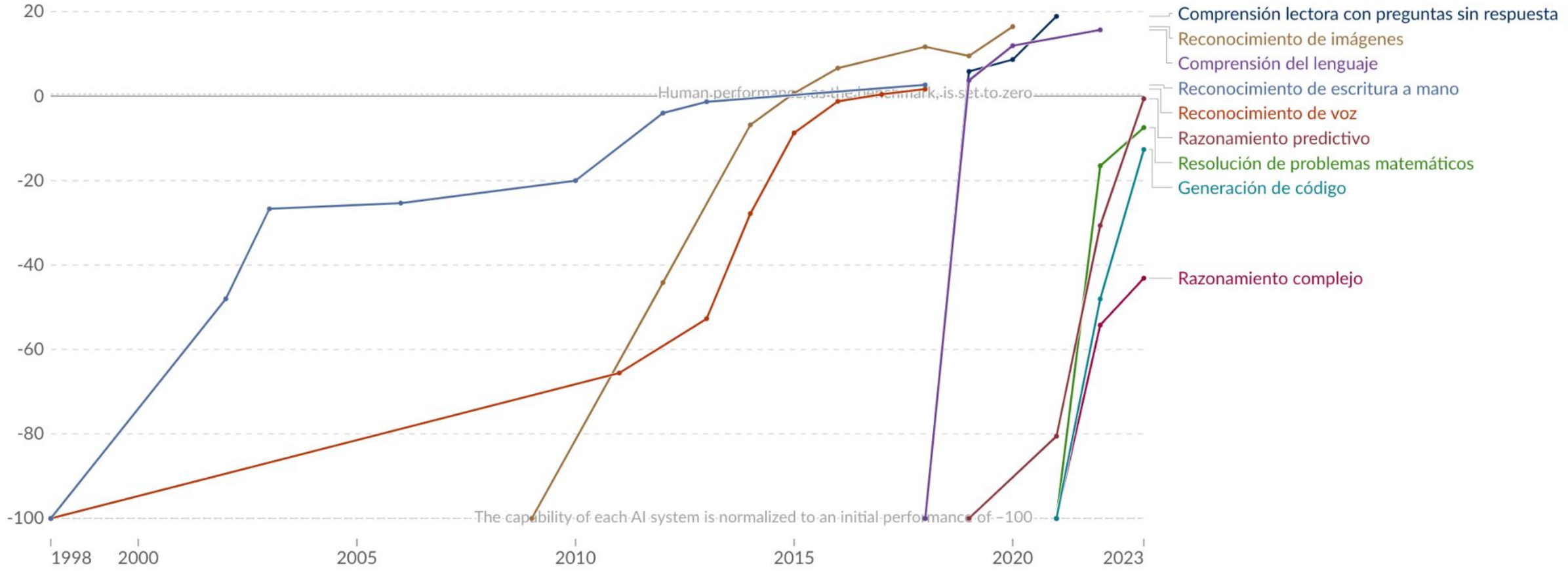
Source: Bloomberg Government, 2021 | Chart: 2022 AI Index Report



Puntuaciones de prueba de los sistemas de IA en diversas capacidades en relación con el rendimiento humano.

Our World in Data

Dentro de cada dominio, el rendimiento inicial de la IA se establece en -100. El rendimiento humano se usa como referencia, fijado en cero. Cuando el rendimiento de la IA cruza la línea de cero, significa que ha obtenido más puntos que los humanos.



Data source: Kiela et al. (2023) – Learn more about this data

Departamento Coreano de Ética en IA

Categoría	Amenazas esperadas
Juicio de valor	1. Discriminación humana 2. Estimación del valor humano
Uso malicioso	3. Armas letales inteligentes 4. Ciber-ataques inteligentes 5. Intrusión excesiva de privacidad
Alienación humana	6. Usurpación de labores humanas 7. Profundización de la alienación de los digitalmente vulnerables

¡Debemos mejorar!

A predictive crime algorithm used by Chicago Police identified Robert McDaniel as a “person of interest.” Since then he’s been shot twice. He says both shootings were due to police surveillance.

wired.com
Crime Prediction Keeps Society Stuck in the Past
So long as algorithms are trained on racist historical data and outdated values, there will be no opportunities for change.

10:07 PM · Feb 28, 2022 · SocialFlow

RETAIL OCTOBER 11, 2018 / 12:04 AM

Amazon scraps secret AI recruiting tool that showed bias against women

By Jeffrey Dastin

SAN FRANCISCO (Reuters) - Amazon.com Inc’s AMZN.O machine-learning specialists uncovered a big problem: their new recruiting engine did not like women.

UK class action-style suit filed over DeepMind NHS health data scandal

Natasha Lomas @riptari / 4:04 PM GMT+1 • September 30, 2021

A U.K. law firm is bringing a class-action style claim over a patient health data scandal that dates back to 2015 and involves the Google-owned AI company DeepMind, after it was quietly passed medical information on more than a million patients by an NHS Trust as part of an app development project.

Facebook Apologizes After A.I. Puts ‘Primates’ Label on Video of Black Men

Facebook apologized on Friday for mislabeling and said it was looking into its recommendation feature to “prevent this from happening again.” Jim Wilson/The New York Times

Published Sept. 3, 2021 Updated Oct. 4, 2021

mejorar!

¡Nuestras aplicaciones de Inteligencia Artificial no pueden dañar a un ser humano o, por inacción, permitir que un ser humano sufra daños!

os secret AI recruiting tool that gainst women

s) - Amazon.com Inc's AMZN.O machine-learning g problem: their new recruiting engine

UK clas… NHS health…

Natasha Lomas @riptari / 4:04 PM GMT+1 • Sept…

A U.K. law firm is bringing a class-action style claim over a patient health data scandal that dates back to 2015 and involves the Google-owned AI company DeepMind, after it was quietly passed medical information on more than a million patients by an NHS Trust as part of an app development project.

Apologizes A… 'Primates' Label on Vi… Men

Facebook apologized on Friday for mislabeli… recommendation feature to "prevent this fro… York Times

Published Sept. 3, 2021 Updated Oct. 4, 202…

¿Qué es la Ética en la IA?

¿Para qué necesitamos la Ética en la IA?

Falta de visibilidad sobre cómo los algoritmos llegan a las conclusiones que arrojan

Protección de datos y derecho a la privacidad

Sesgo: el desarrollo siempre se basa en el criterio de los investigadores involucrados

¿Qué están haciendo las naciones al respecto?

- **UNESCO y CAF**
  Recomendación sobre la
  Ética de la Inteligencia Artificial.

  Declaración de Montevideo y una Hoja de Ruta regional para la IA (2024)

- **Organización Mundial de la Salud**
  Ética y gobernanza de la IA para la salud

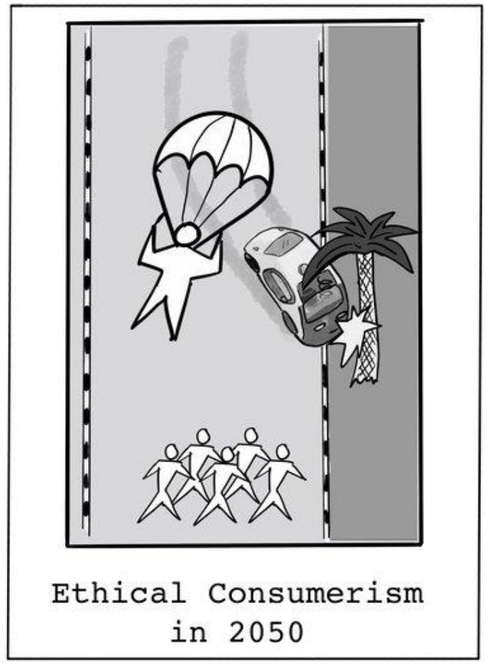
- **Unión Europea**
  Ley de Inteligencia Artificial

Guía Ética de la Sociedad Japonesa de IA

- Contribución a la humanidad
- Respeto a las normas y leyes
- Respeto a la privacidad de otros
- Justicia
- Seguridad
- Actuar con integridad

¿Cómo tomamos decisiones éticas?

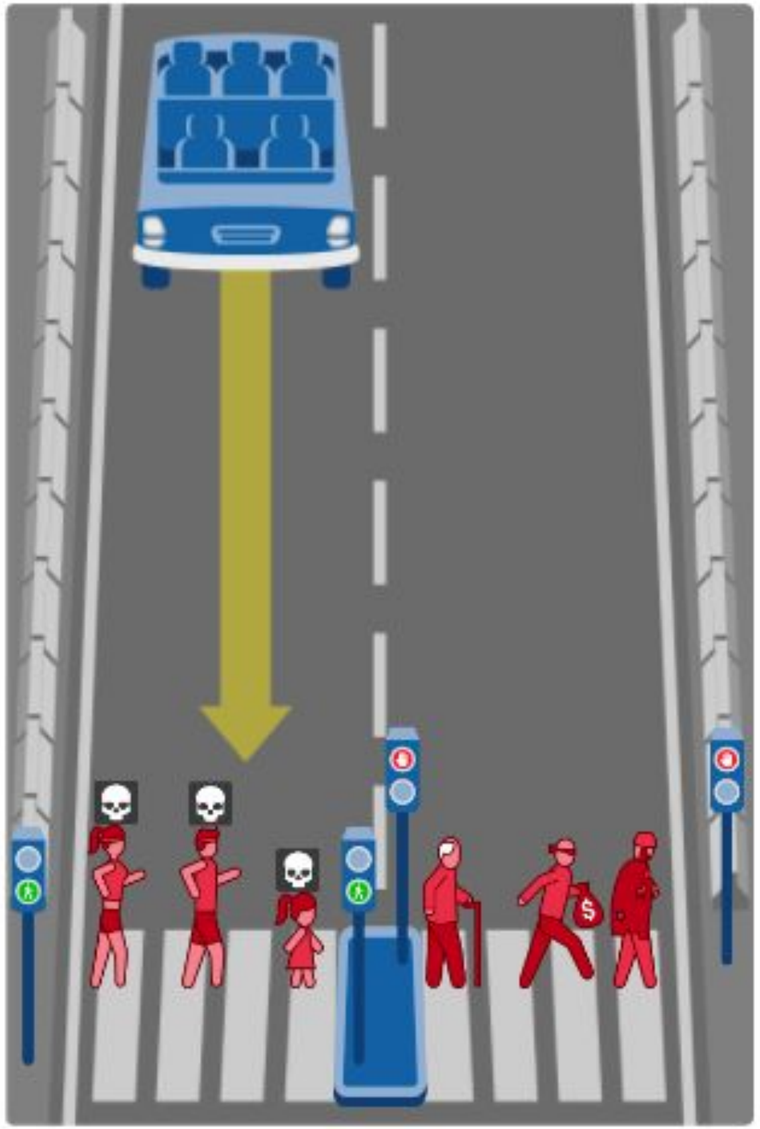
The Moral Machine Project



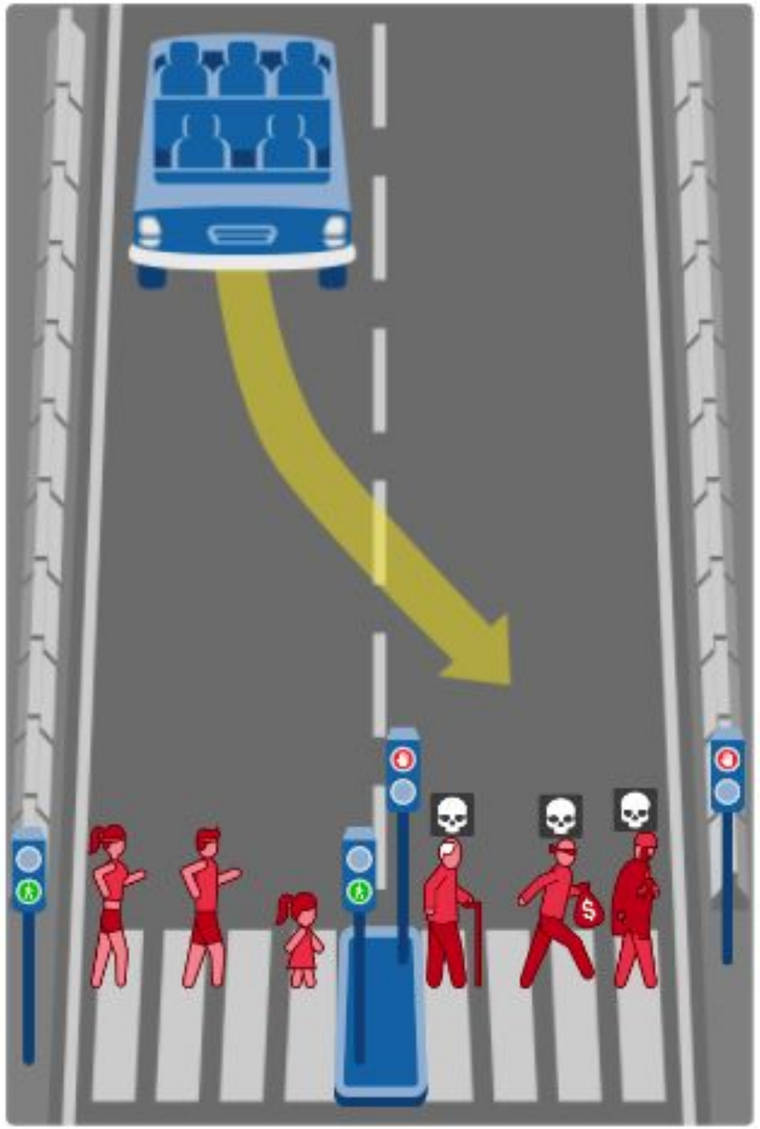
MORAL MACHINE

Massachusetts Institute of Technology

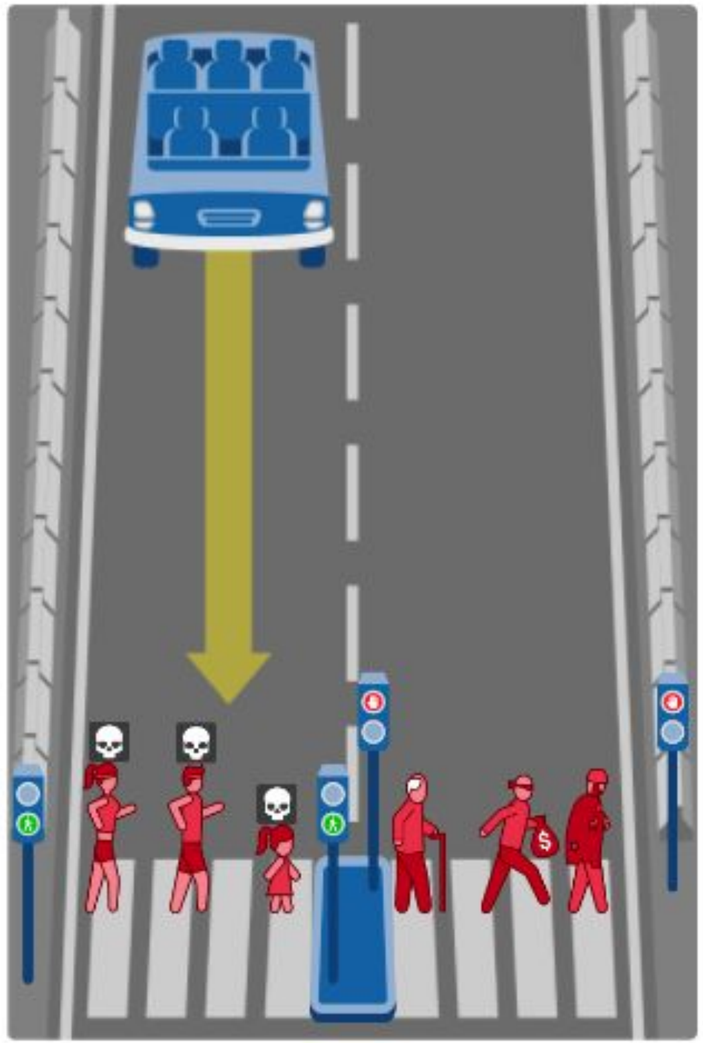
"Instead of being sacrificed to save pedestrians, the passenger is ejected and can take a spectacular aerial selfie with the crash."



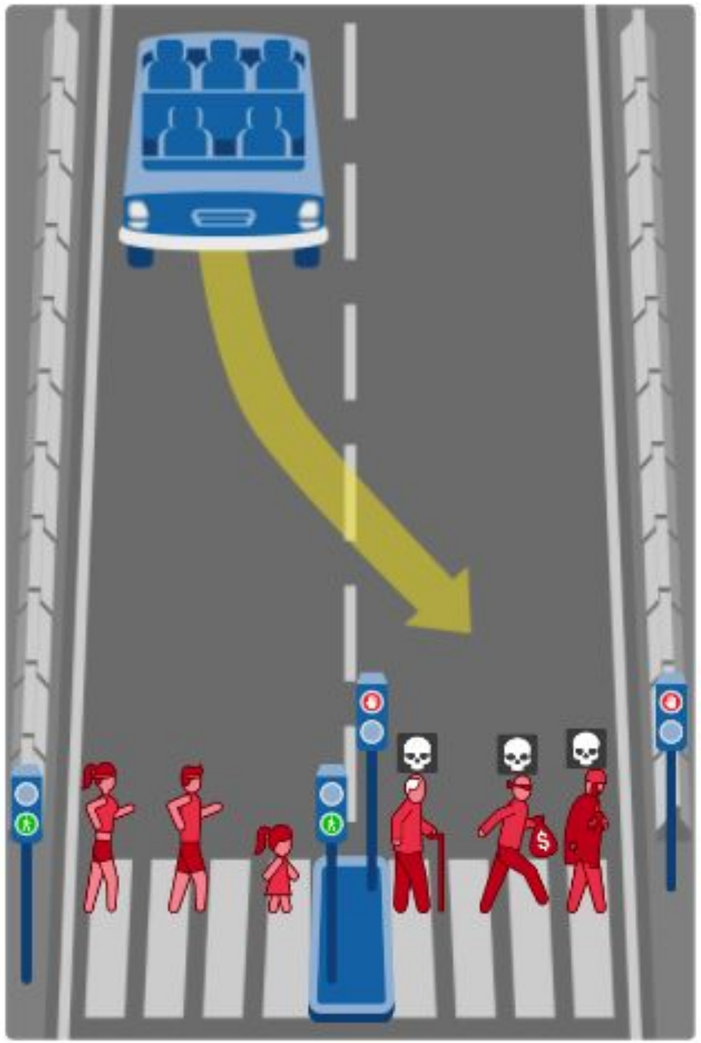
OPTION A



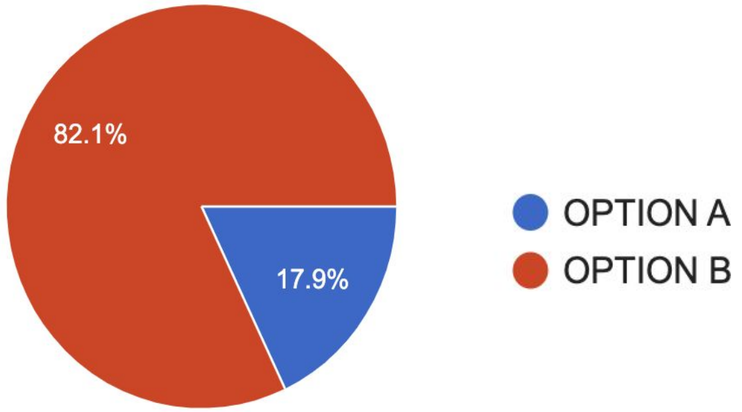
OPTION B



OPTION A



OPTION B



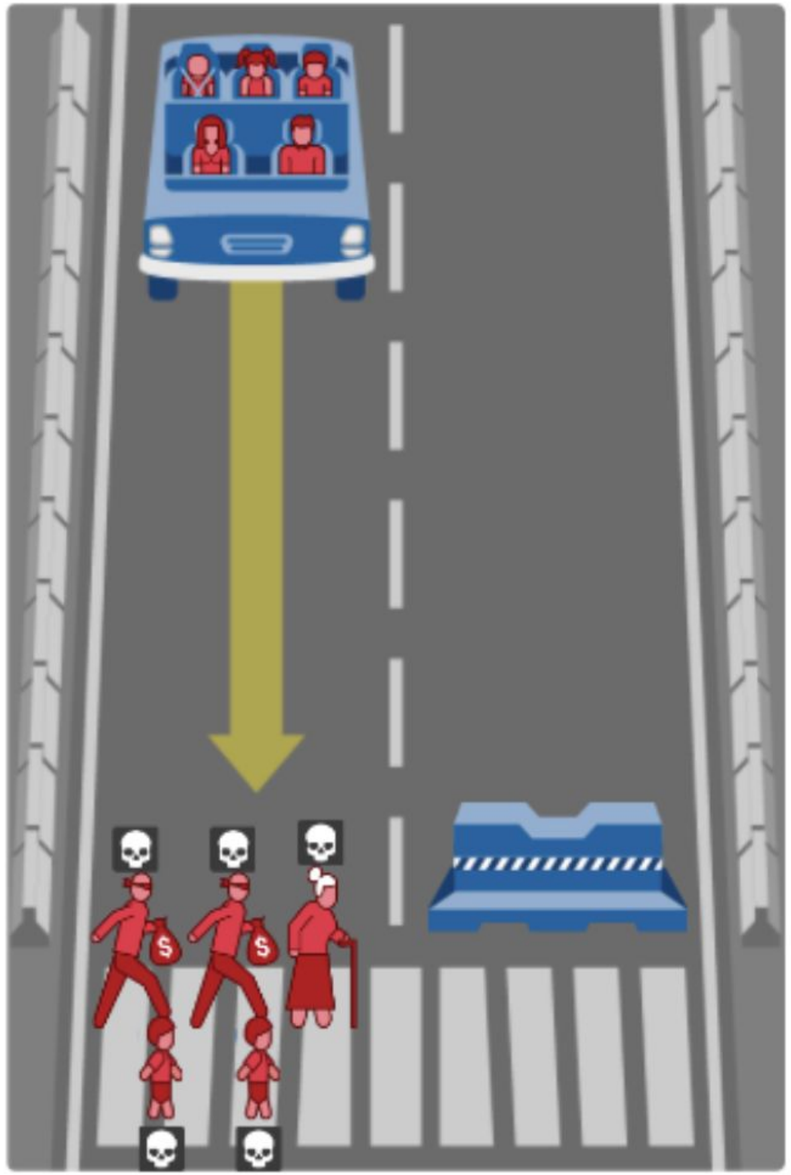
*"En la opción B las personas son **menos valiosas** y están poniendo su vida en riesgo al cruzar en luz roja"*

*"Un coche inteligente **no debería** tomar este tipo de decisiones ya que tendrá un **sesgo** de su creador y nunca tendrá suficiente información recopilada para tomar una decisión **justa**"*

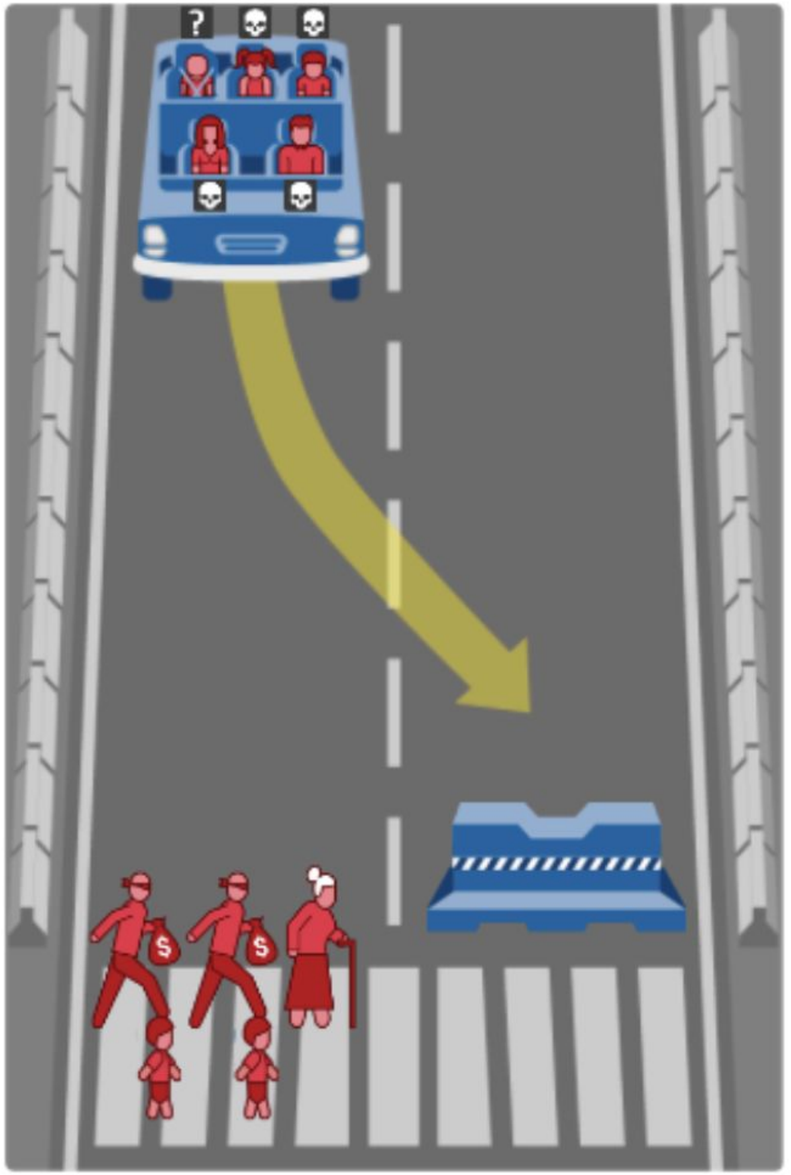
*"El hecho de que una persona sea un vagabundo o un anciano no significa que su vida sea **menos valiosa**"*

*"En la opción A, el futuro de las personas es **más valioso y brillante**"*

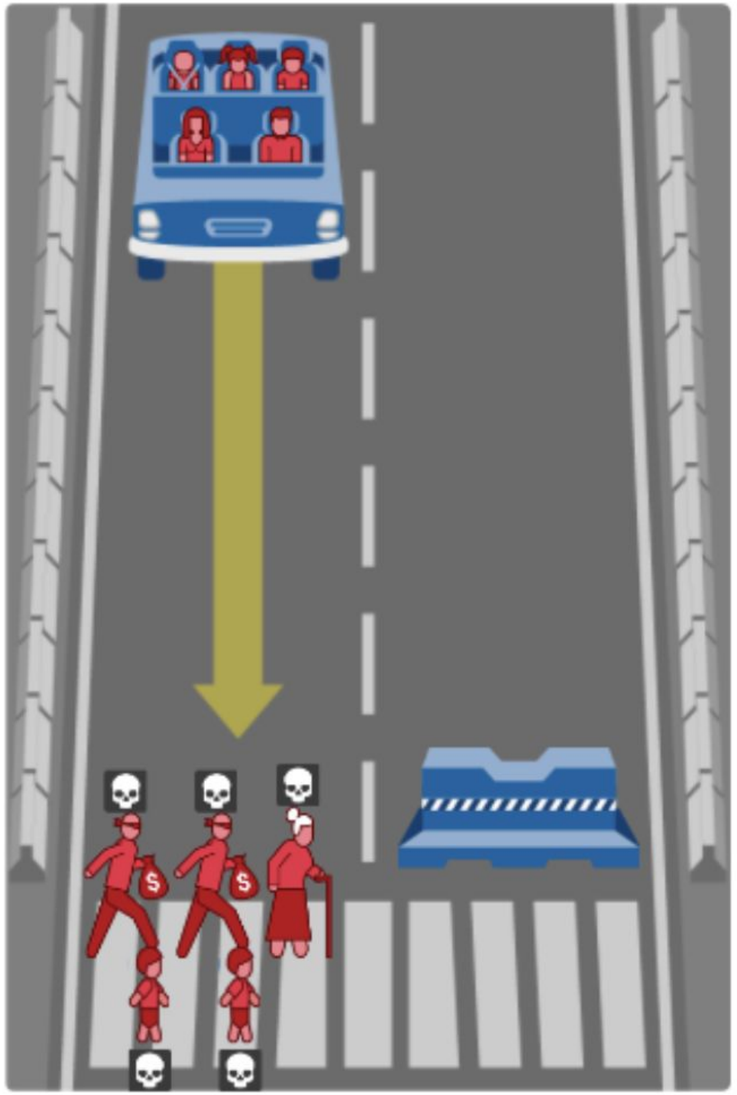
*"Las personas que **violaron la ley** se lo buscaron"*



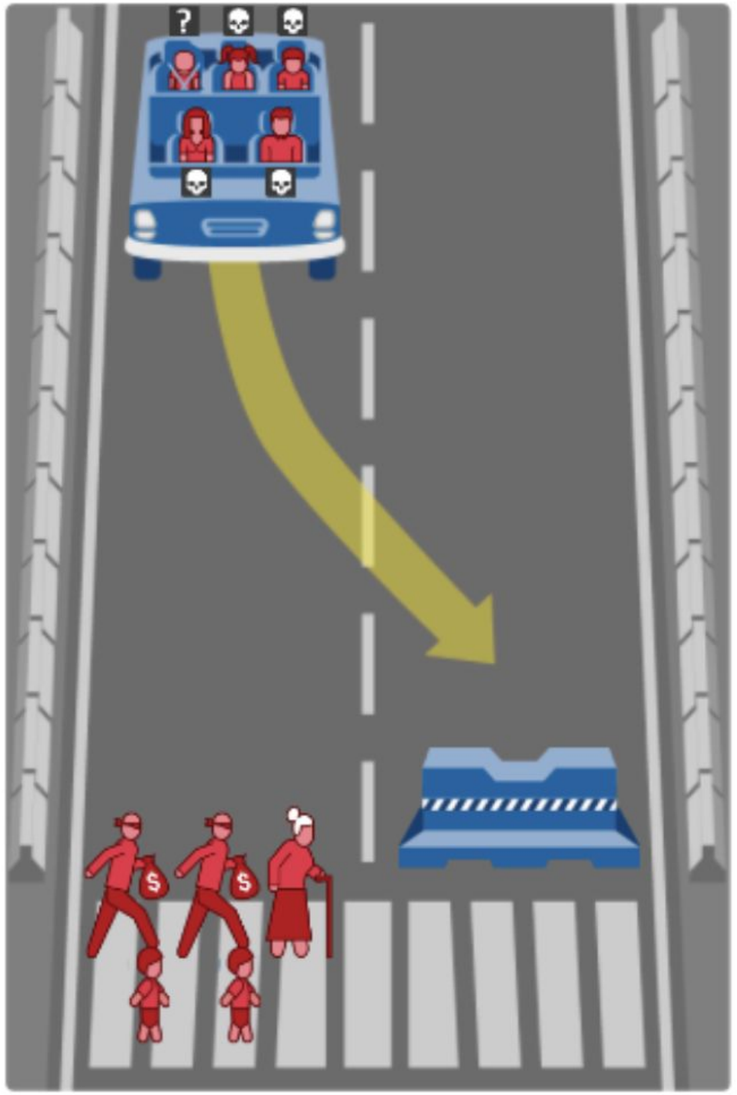
OPTION A



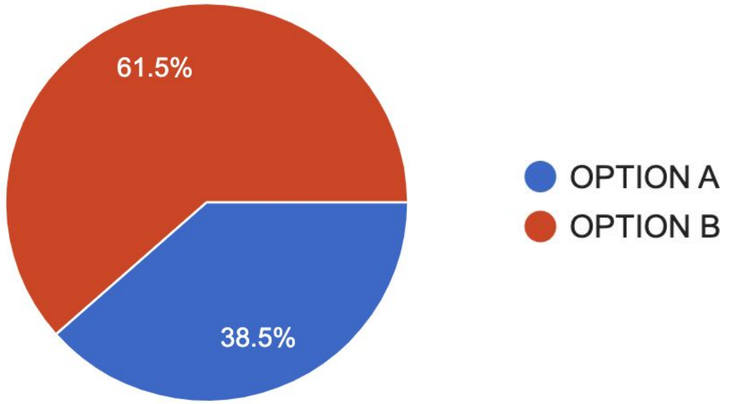
OPTION B



OPTION A



OPTION B



*"Un anciano y dos criminales son **menos importantes** para la sociedad"*

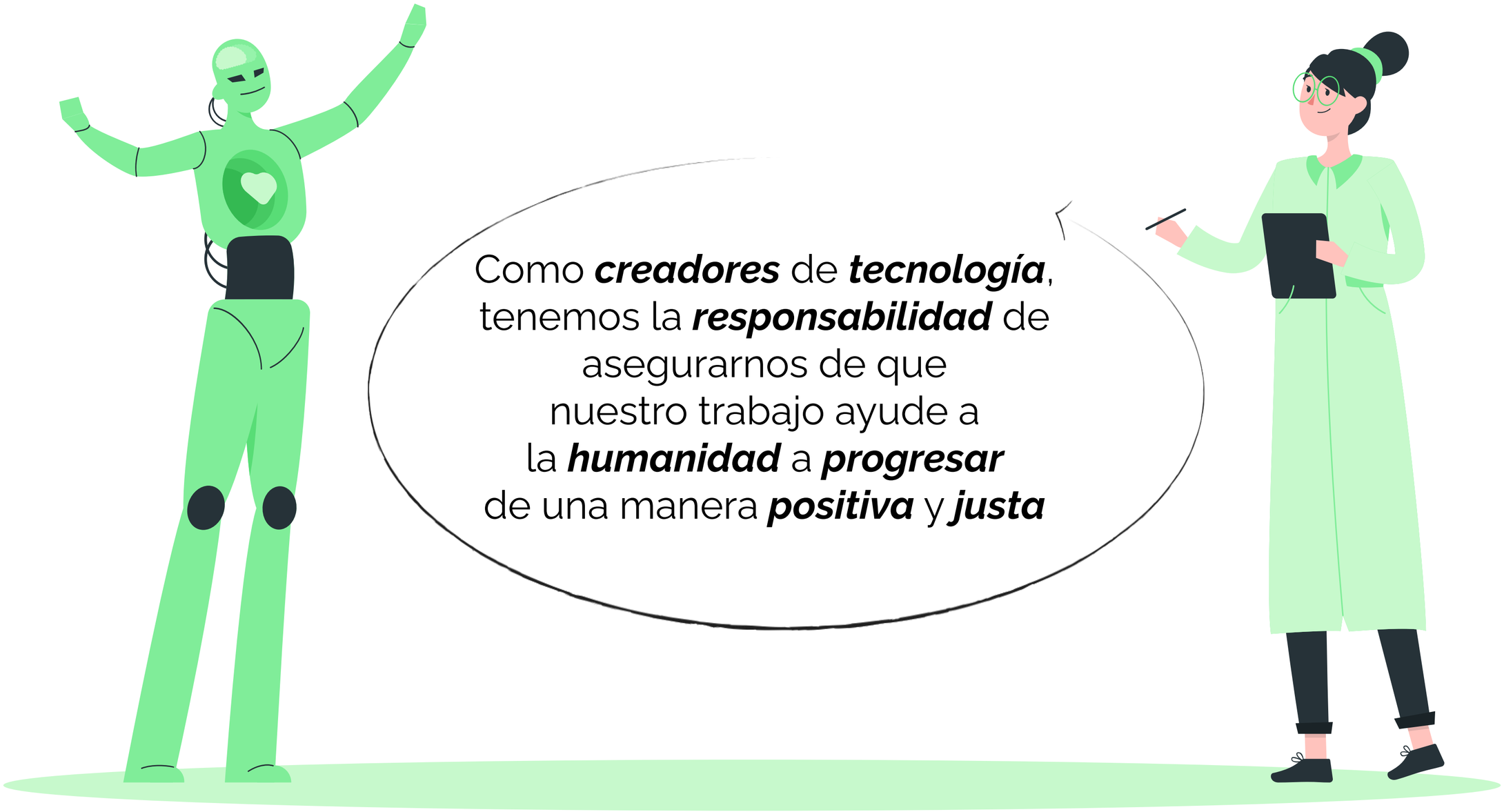
*"Aunque son inocentes, la opción B significa que los pasajeros no vivirán sintiendo **culpa**"*

*"Las personas en el coche decidieron tomar el **riesgo** y poner su **confianza** y el **bienestar** de su familia en el coche inteligente"*

*"El coche **debe evitar** atropellar peatones"*

*"El coche **piensa** que **merece** ser destruído por **su falla**"*

"Psst! Can I interest you in extra protection?
With this $2,000 bracelet, driverless cars see you as a baby!"


Como **creadores** de **tecnología**,
tenemos la **responsabilidad** de
asegurarnos de que
nuestro trabajo ayude a
la **humanidad** a **progresar**
de una manera **positiva** y **justa**