Consideraciones Éticas de la Inteligencia Artificial

Joaquín López Herraiz
jlopezhe@ucm.es

Profesor Titular de Universidad
Grupo de Física Nuclear e IPARCOS. Universidad Complutense de Madrid
Instituto de Investigación Sanitaria del Hospital Clínico San Carlos (IdISSC), Madrid

ÍNDICE:

1) ÉTICA: I+D+i RESPONSABLE & ANALISIS DAFO
2) RETOS DE LA IA
3) ¿QUÉ SE PUEDE HACER?
4) REGULACIÓN: Ley de IA Europea

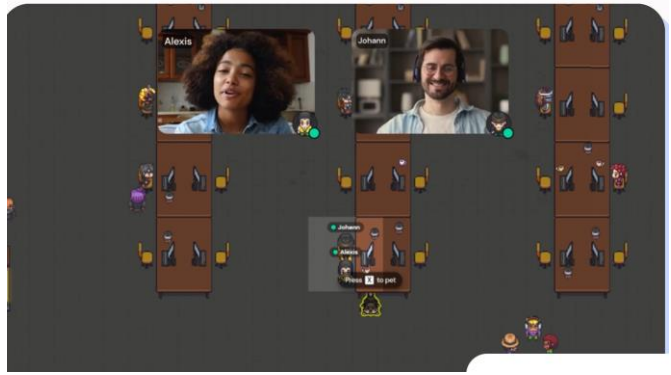
BLOQUE 1: Toda I+D+i tiene consecuencias
No debemos trabajar sin ser conscientes de ello

- Eso es especialmente importante con tecnologías punteras como la I.A.
- ¿Qué consecuencias (positivas y negativas) tendrá aquello que estamos desarrollando?
  - ¿Cómo podemos potenciar sus efectos positivos?
  - ¿Cómo podemos mitigar/adaptarse a sus efectos negativos?
  - Hay que anticiparse para realizar una I+D+I responsible.

Ejemplo: Clases Online por Videoconferencia

- **EFECTOS POSITIVOS (y cómo potenciarlos)**
  - Flexibilidad (Aprender desde múltiples entornos) → Permitir acceso desde PC, Tablet, Movil
  - Llegar a un público más amplio → Cursos Masivos (Coursera..)
  - Democratizar aprendizaje → Titulos online (EdX, MITx..)
- **EFECTOS NEGATIVOS (y cómo mitigarlos / adaptarse)**
  - Atención se ve disminuida → Adaptar contenidos al nuevo entorno
  - No hay conexión entre estudiantes → https://www.gather.town/
  - No todos los estudiantes tienen acceso → Programas para universalizar acceso
  - Impacto en el empleo → Reconvertir

En todas las acciones para fomentar efectos positivos y mitigar/adaptar a efectos negativos de surgen oportunidades de negocio.



Las cuestiones éticas se están planteando en todos los ámbitos

https://www.telefonica.com/es/sala-comunicacion/los-retos-sociales-y-eticos-del-metaverso/

Desafíos Directos

- **Sesgos y discriminación no deseada.**
- **Explicabilidad y algoritmos de caja negra.**
- **Intervención humana y autonomía adecuada del sistema**
- **Privacidad**
- **Seguridad y protección**
- **Huella de carbono de los algoritmos**
- **Noticias falsas y deep fakes**
- **La relación entre las personas y las máquinas –** ¿Una I.A. Premio Nobel?
- **Derechos de autor**
- **La concentración de datos y riqueza y la creciente desigualdad**
- **IA para la defensa y la guerra**

Desafios Indirectos

- **Aumento de la polarización en las sociedades** (Sesgo de refuerzo positivo)
- **Adicción a la tecnología, a los juegos o a las aplicaciones de las redes sociales**
- **El aumento de la anorexia y la vigorexia en los adolescentes por la presión social**

Desde la aparición de las redes sociales mujeres de entre 15 y 19 años tuvieron un aumento del 70% de casos de suicidio. Entre 10-14 años fue del 151%

Centers for Disease Control and Prevention, Increase in Suicide Mortality in the United States, 1999-2018, April 2020.

Análisis DAFO de la Inteligencia Artificial

PRESENTE

FORTALEZAS

¡Está de moda!
Cada vez es más fácil de usar
Ya ha logrado importantes éxitos

Debilidades

Es aún demasiado nuevo
No se comprende del todo (heurística)
Cambia demasiado rápido

FUTURO

Oportunidades

Solucionar problemas que siguen abiertos
Puede abrir nuevos campos

Amenazas

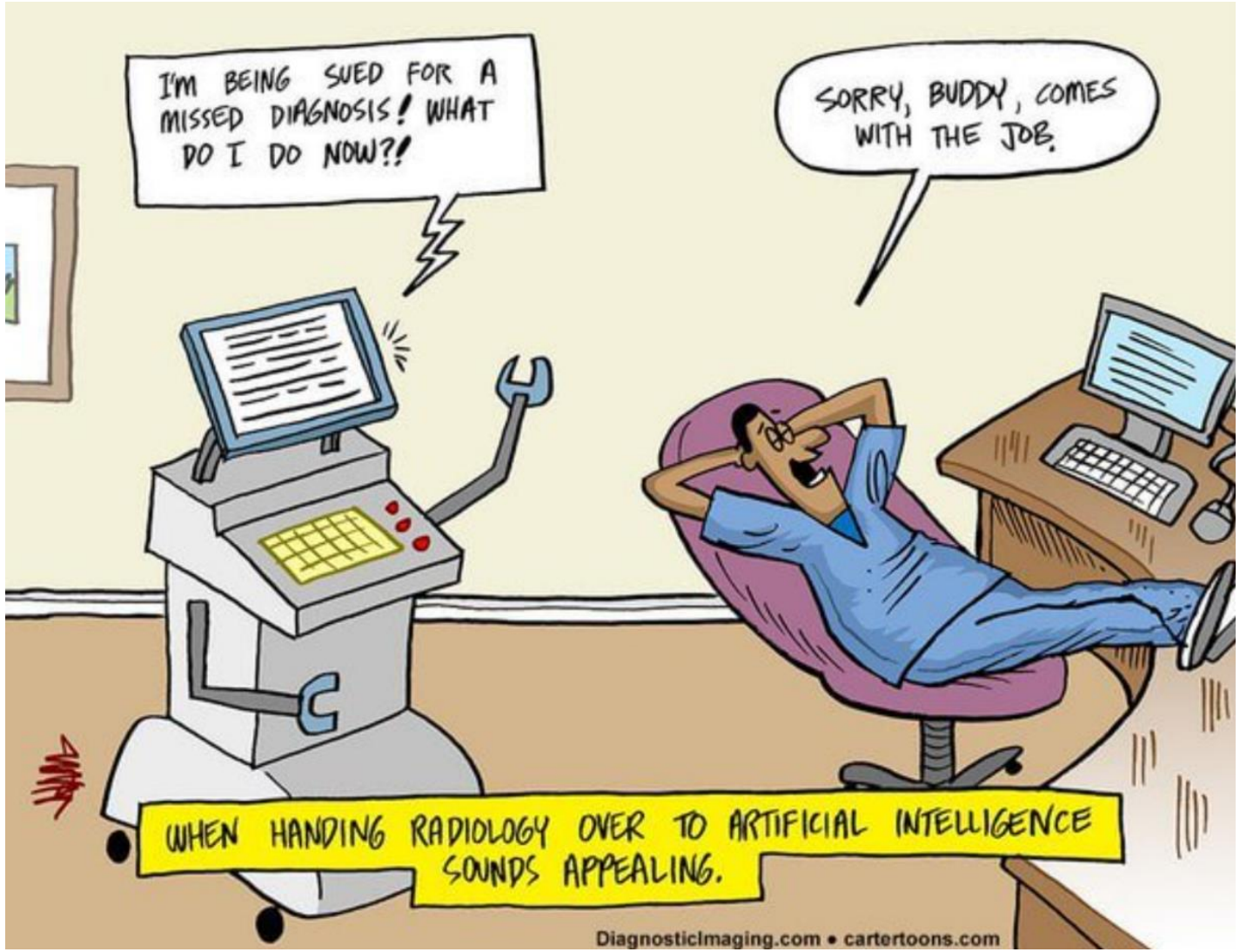
Esta charla

Puede usarse sin expertos: RETOS
(riegos de errores y sesgos)
Numerosos Aspectos éticos
Se puede usar con malos fines

BLOQUE 2: Retos de la Inteligencia Artificial

1) Regulación
2) Comprender y Explicar los Resultados
3) Robustez y Fiabilidad
4) Datos y Etiquetado
5) Armonización
6) Multidisciplinar
7) Correlación≠Causa
8) Otras cuestiones éticas

Reto 1) Regulación



https://www.fda.gov/medical-devices/software-medical-device-samd/artificial-intelligence-and-machine-learning-software-medical-device

Reto 1) Regulación

La I.A. ya se está usando en Medicina de manera regulada

Type to filter data...

Name of device or algorithm	Name of parent company	Short description	FDA approval number	Type of FDA approval	Mention of A.I. in announcement	If no mention of A.I. in FDA announcement	Date	Medical specialty	Secondary medical specialty
Arterys Cardio DL	Arterys Inc	software analyzing cardiovascular images from MR	K163253	510(k) premarket notification	deep learning		2016 11	Radiology	Cardiology
EnsoSleep	EnsoData, Inc	diagnosis of sleep disorders	K162627	510(k) premarket notification	automated algorithm		2017 03	Neurology	
Arterys Oncology DL	Arterys Inc	medical diagnostic application	K173542	510(k) premarket notification	deep learning		2017 11	Radiology	Oncology
Idx	IDx LLC	detection of diabetic retinopathy	DEN180001	de novo pathway	A.I.		2018 01	Ophthalmology	
Koios DS for Breast	Koios Medical, Inc	diagnostic software for lesions suspicious for cancer	K190442	510(k) premarket notification	machine learning		2019 06		
ContaCT	Viz.AI	stroke detection on CT	DEN170073	de novo pathway	A.I.		2018 02	Radiology	Neurology
OsteoDetect	Imagen Technologies, Inc.	X-ray wrist fracture diagnosis	DEN180005	de novo pathway	deep learning		2018 02	Radiology	Emergency medicine
Guardian Connect System	Medtronic	predicting blood glucose changes	P160007	PMA	A.I.		2018 03	Endocrinology	
EchoMD Automated Ejection Fraction Software	Bay Labs, Inc.	echocardiogram analysis	K173780	510(k) premarket notification	machine learning		2018 05	Radiology	Cardiology
DreaMed	DreaMed Diabetes, Ltd	managing Type 1 diabetes.	DEN170043	de novo pathway	A.I.		2018 06	Endocrinology	
LungQ	Thirona Corporation	Quantitative analysis of chest CT scans	K173821	PMA	Not available	https://bit.ly/2EkHTCM	2018 06	Radiology	Pulmonology
BriefCase	Aidoc Medical, Ltd.	triage and diagnosis of time sensitive patients	K180647	510(k) premarket notification	deep learning		2018 07	Radiology	Emergency medicine
ProFound™ AI Software V2.1	iCAD, Inc	breast density via mammogprahy	K191994	510(k) premarket notification	deep learning		2018 07	Radiology	Oncology

https://medicalfuturist.com/fda-approved-ai-based-algorithms/

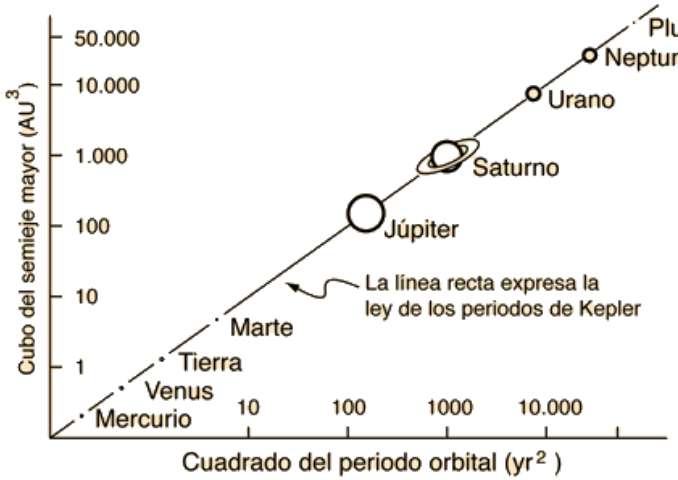
RETO 2) Comprender y Explicar Resultados de una IA

Tycho Brahe y Johannes Kepler

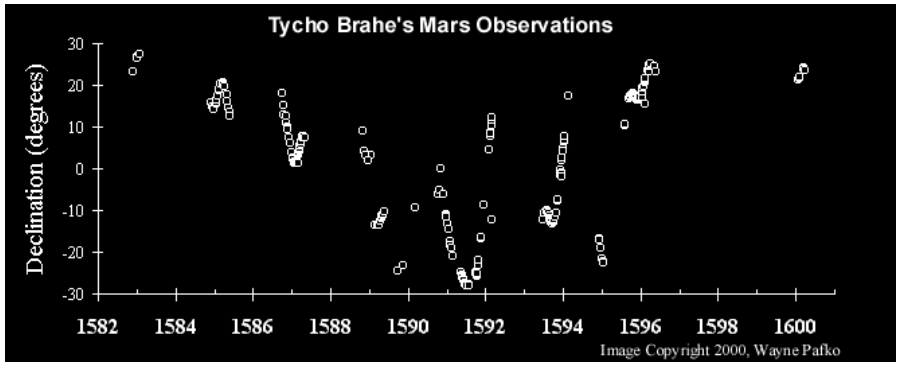
Método Científico Clásico

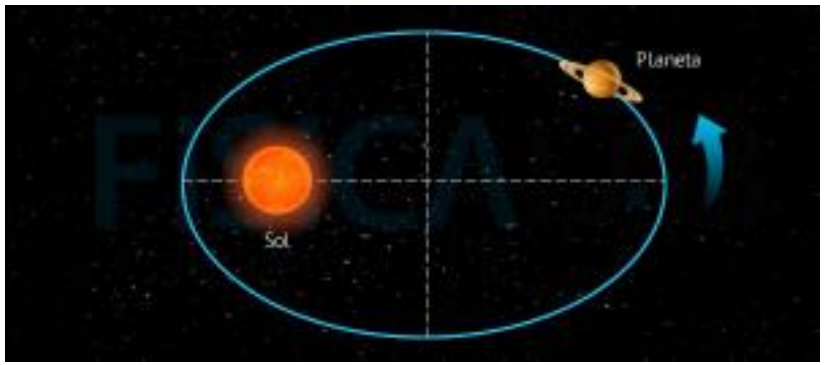
$$T^2 = ka^3$$

Modelo Propuesto con Parámetro k



Datos vs Modelo “Simple”





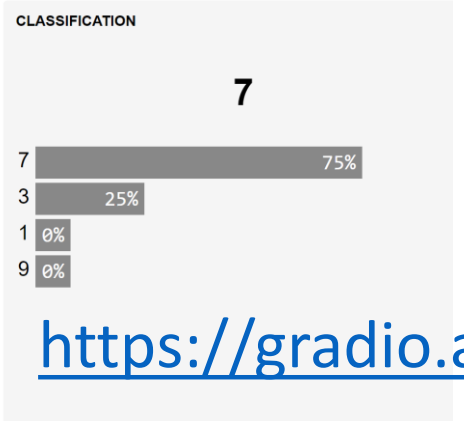
Modelo Ajustado = Conocimiento …pero es limitado

RETO 2) Comprender y Explicar Resultados de una IA

¿Perro o Gato?

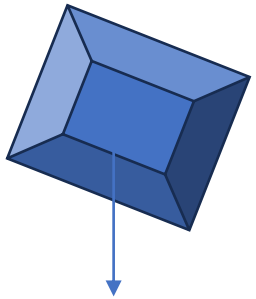
¿Qué es?

¿Por qué?



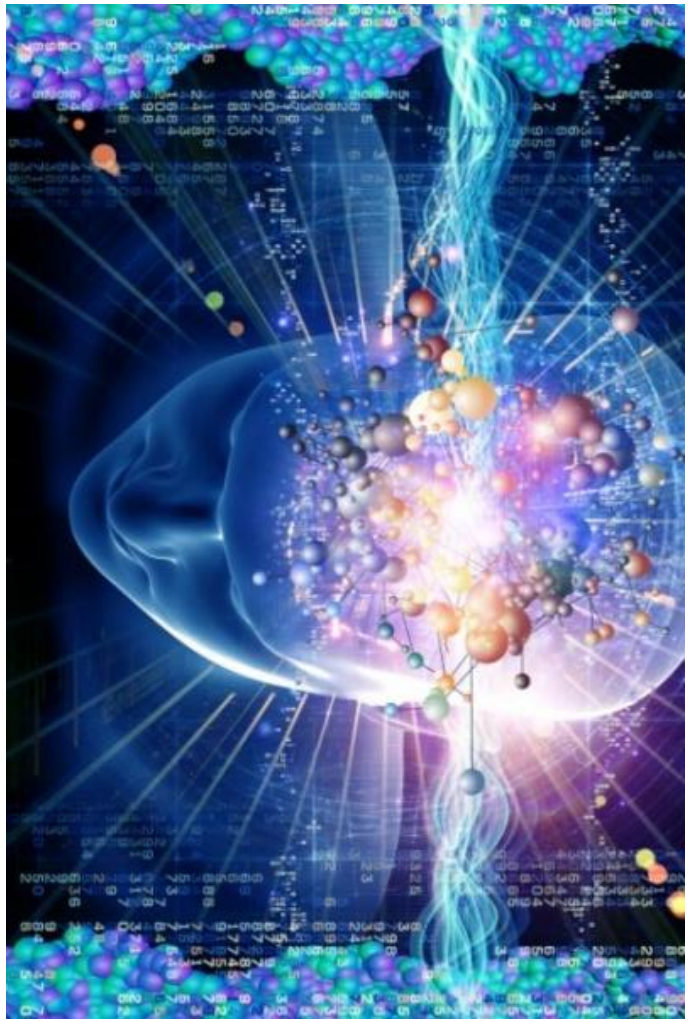
https://gradio.app/

¿Cuánto tarda exactamente un objeto ligero en caer?



Demasiadas variables (forma, distribución, densidad del aire)… para un modelo simple

Inteligencia Artificial vs Método Científico Clásico



APRENDIZAJE AUTOMÁTICO (MACHINE LEARNING)

1) MEDIDAS (DATOS)
2) ~~PROPONER UN MODELO (CON PARAMETROS)~~
3) AJUSTAR LOS DATOS A **CUALQUIER** MODELO
4) CREAR PREDICCIONES con el MEJOR MODELO
5) VALIDAR MODELO

LIMITACIONES:

1) DIFÍCIL INTERPRETACIÓN
2) RIESGO DE SOBREAJUSTES, SESGOS..
3) USO FUERA DE SU DOMINIO DE APLICACIÓN

Inteligencia Artificial Vs Métodos Clásicos

Alpha Zero vs **Stockfish**

(28 Victorias, 72 Tablas, 0 Derrotas)

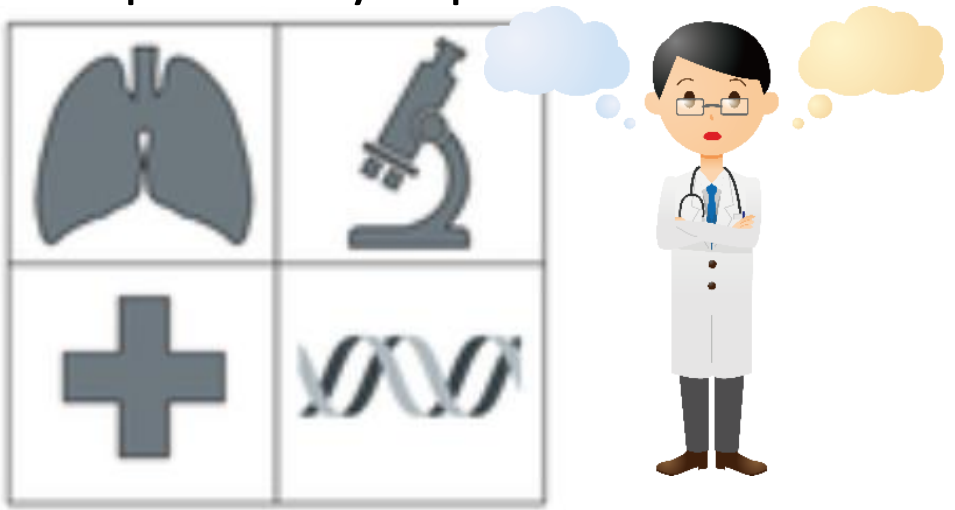
Alpha Zero jugando contra sí mismo durante sólo 4 horas encontró “modelo” de manera correcta de jugar.

ALPHAZERO VS STOCKFISH
https://www.youtube.com/watch?v=9nkLJreG21c

ALPHAZERO VS ALPHAZERO
https://www.youtube.com/watch?v=mOqmLYlFdBo

RETO 2) Comprender y Explicar Resultados de una IA

Resultados de una IA basados en Analíticas, Genética, Radiómica, Variables ambientales.. Serán muy potentes pero difícil de comprender y explicar…



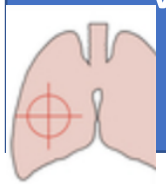
RETO 2) Comprender y Explicar Resultados de una IA

EJEMPLO DE SOLUCIÓN: #1 – IA EXPLICABLE

DATOS

MODELO DE IA

IDENTIFICAR CONEXIÓN CON ALGUNAS VARIABLES



EXPERTO LOGRA EXPLICAR CON ELLO EL RESULTADO

Ejemplo:Heat-maps

Localizar región más correlacionada con el resultado

RETO 2) Comprender y Explicar Resultados de una IA

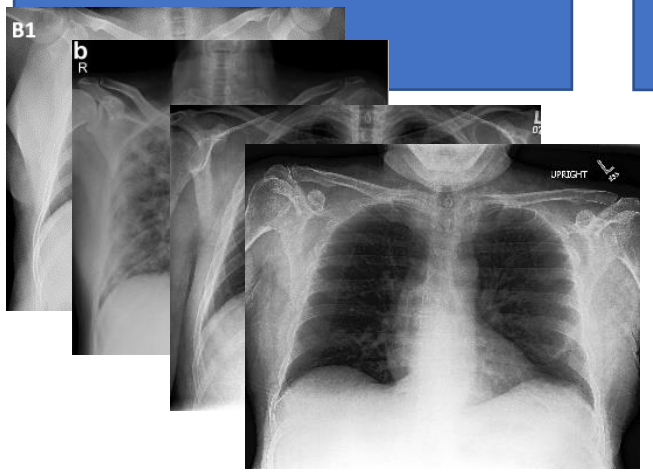
EJEMPLO DE SOLUCION #2 – ENCONTRAR CASOS SIMILARES

DATOS

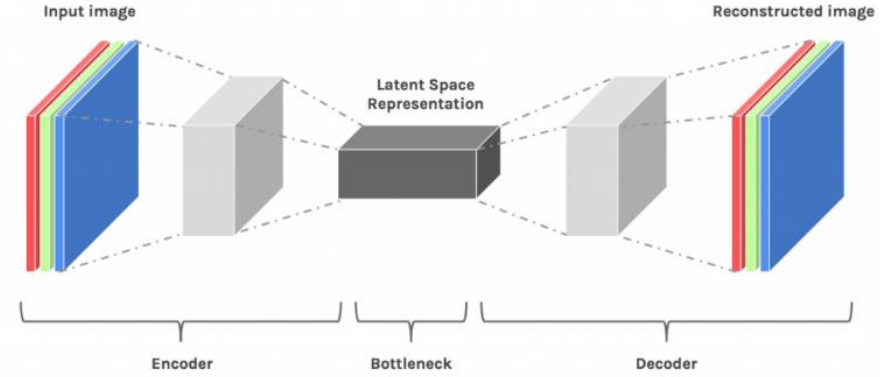
MODELO IA

ENCONTRAR CASOS SIMILARES:

DIAGNOSTICO
TRATAMIENTO
RESULTADOS



MEDICINA PERSONALIZADA BASADA EN LA EVIDENCIA



"Searching for pneumothorax in x-ray images using autoencoded deep features". A. Sze-To et

RETO 3) ROBUSTEZ Y FIABILIDAD

RIESGO: Usar la IA fuera de su rango de aplicación

Las herramientas de IA deberían admitir cuando no están lo suficientemente entrenadas para una determinada tarea (Como hacemos los humanos)

SOLUCION: Indicar el nivel de similitud del problema tratado con otros previos con los que ha sido entrenado. “Out-of-Distribution” Si es un caso completamente nuevo, la fiabilidad de la respuesta debería ser reducida.

¿Qué sucede con las minorías?

Worst-case Scenario: Considerar si es suficiente con desarrollar herramientas que “en promedio” funcionan bien, o si es necesario ver qué pasa con el “peor escenario posible”.

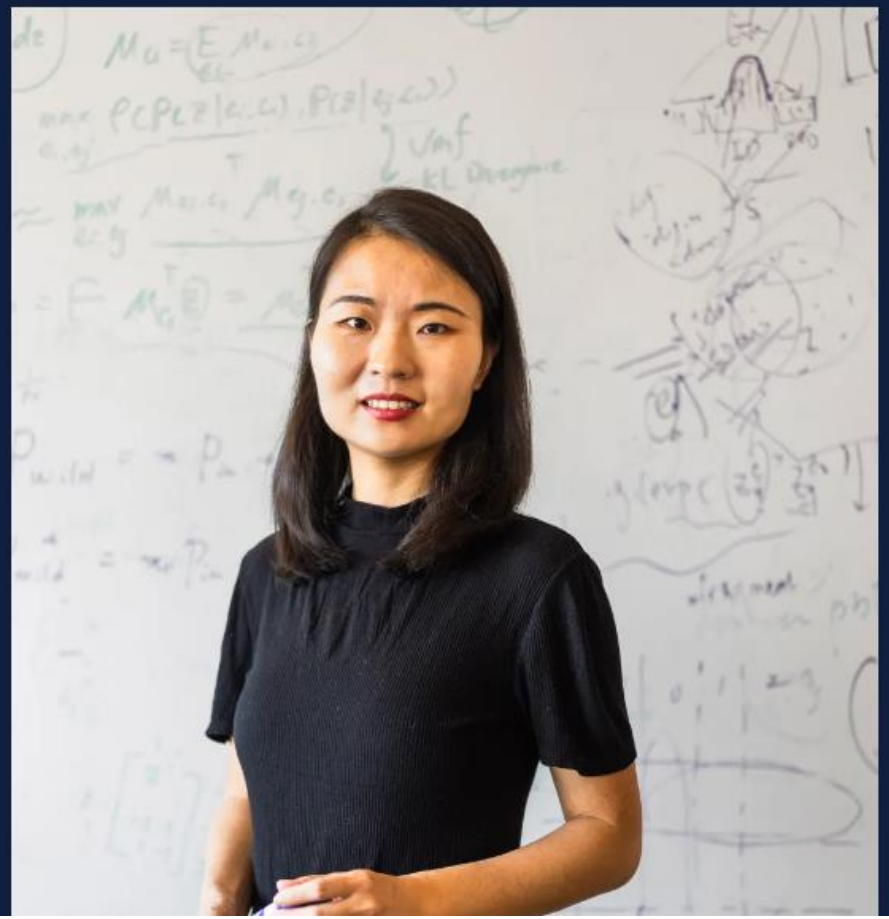
ARTIFICIAL INTELLIGENCE

2023 Innovator of the Year: As AI models are released into the wild, Sharon Li wants to ensure they're safe

Li's research could prevent AI models from failing catastrophically when they encounter unfamiliar scenarios.

By Melissa Heikkilä

September 12, 2023

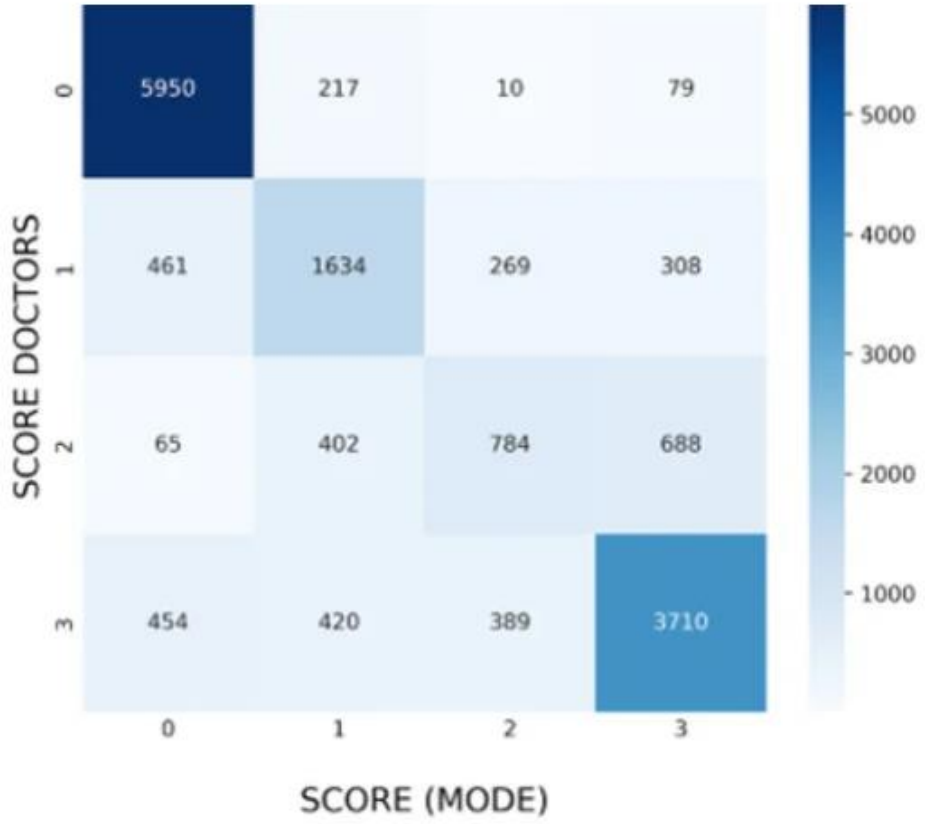


https://pages.cs.wisc.edu/~sharonli/publication

RETO 4) DATOS Y ETIQUETAS

¿Dónde conseguir buenos datos?
¿Cómo han sido tratados?
¿Son representativos?
¿Tienen sesgos?

La calidad de los modelos de IA dependen de la calidad de los datos usados para entrenar



En aprendizaje supervisado, el etiquetado es realizado por humanos, tendrá errores y sesgos. Es difícil lograr con la IA mejores resultados que los humanos que han etiquetado los casos.

Inter-Rater Variability in the Evaluation of Lung Ultrasound in Videos Acquired from COVID-19 Patients. Joaquin L. Herraiz et al. Appl. Sci. 2023

RETO 5) ARMONIZACIÓN

PROBLEMA CUANDO SE USAN BASES DE DATOS HETEROGENEAS

- **Common pitfalls and recommendations for using machine learning to detect and prognosticate for COVID-19 using chest radiographs and CT scans (https://www.nature.com/articles/s42256-021-00307-0)**

"Our search identified 2,212 studies, of which 415 were included after initial screening and, after quality screening, 62 studies were included in this systematic review. Our review finds that none of the models identified are of potential clinical use due to methodological flaws and/or underlying biases. "

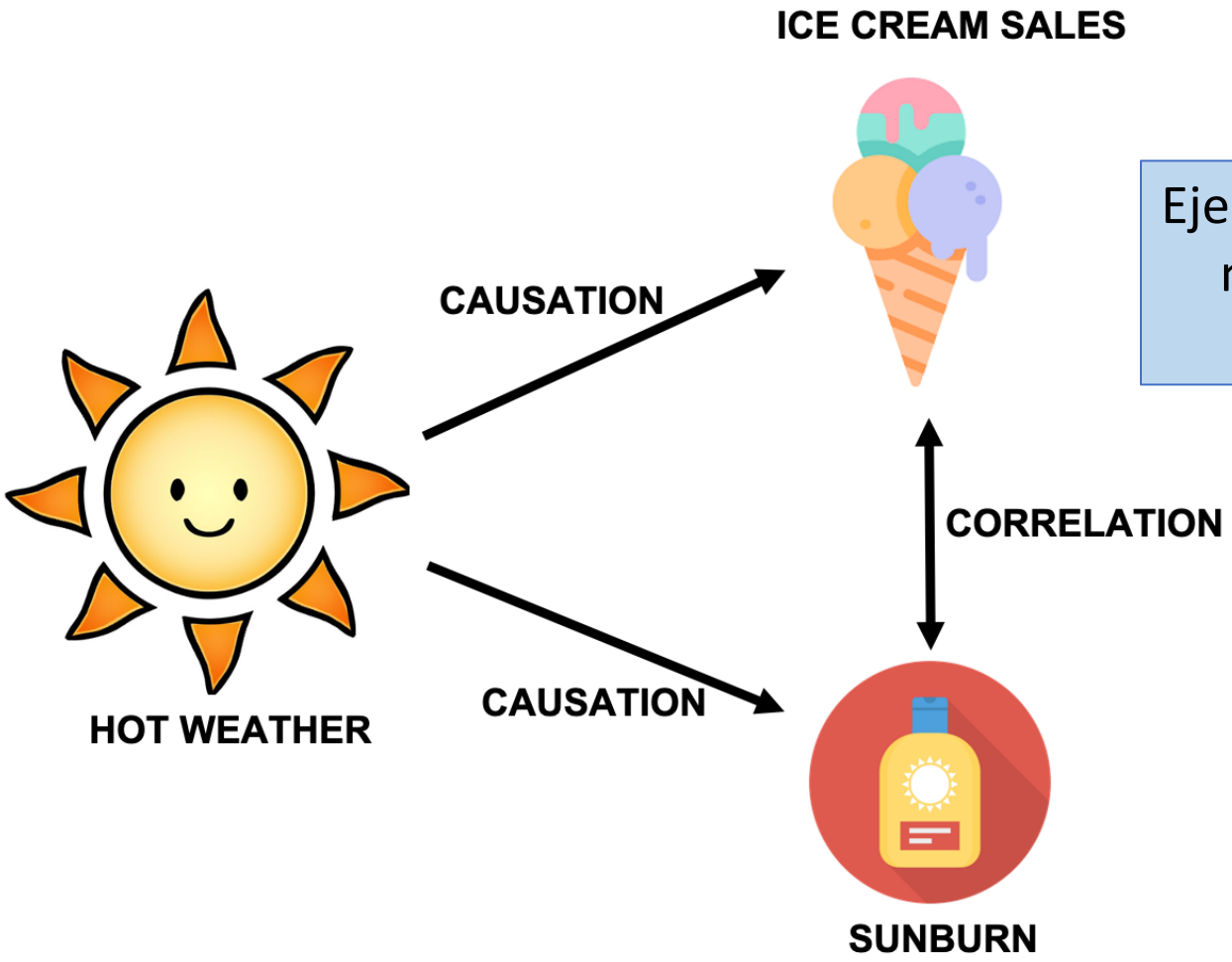
SOLUCION 1 - ARMONIZACIÓN – HACER DATOS COMPATIBLES
SOLUCION 2 - DESARROLLAR MODELO PARA 1 CASO CONCRETO

RETO 6) Multidisciplinaridad

La IA permite trabajar en temas sin ser experto.
¡Pero eso no significa que debas hacerlo!

SOLUCIÓN: Colaborar con expertos de distintas áreas y aprender de todos

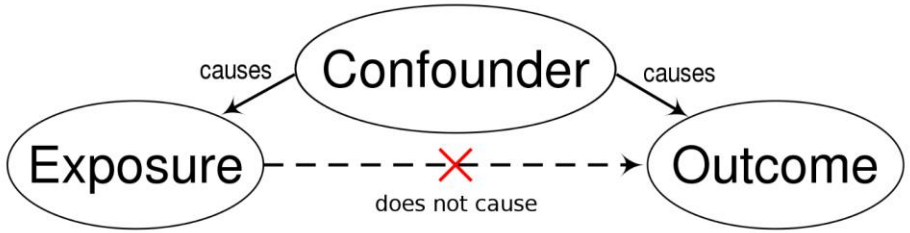
RETO 7)
¡Correlación no implica Causalidad!



Ejemplo: Pacientes con otras enfermedades tienen menos probabilidad de tener complicaciones. ¡Porque reciben mejores tratamientos!

IA es muy buena encontrando correlaciones, pero ¡Eso no significa que detecte relación causa-efecto!

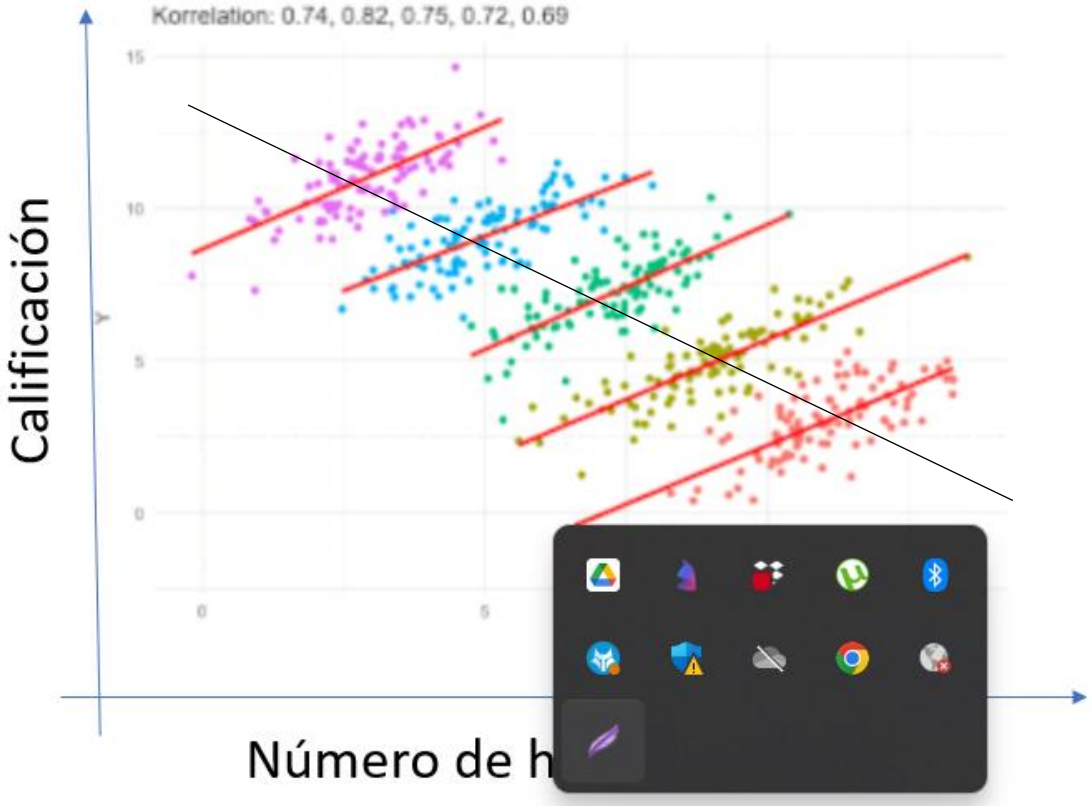
RETO 7) Simpson's Paradox



https://en.wikipedia.org/wiki/Simpson%27s_paradox

Lurking or confounding variables
(Variable de Confusión)

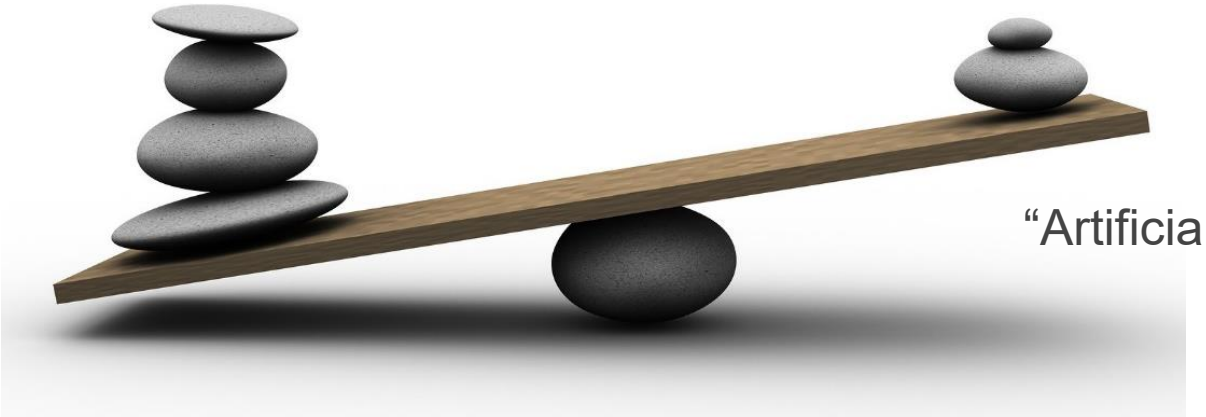
Treatment / Stone size	Treatment A	Treatment B
Small stones	*Group 1* **93% (81/87)**	*Group 2* 87% (234/270)
Large stones	*Group 3* **73% (192/263)**	*Group 4* 69% (55/80)
Both	78% (273/350)	**83% (289/350)**



*C. R. Charig; D. R. Webb; S. R. Payne; J. E. Wickham (29 March 1986). "Comparison of treatment of renal calculi by open surgery, percutaneous nephrolithotomy, and extracorporeal shockwave lithotripsy". Br Med J (Clin Res Ed). **292** (6524): 879–882.*

8) Otros Retos Éticos de la I.A.

- ¿Qué pasa con los casos especiales?
- ¿Quién es responsable de decisiones guiadas por I.A.?
- ¿Cómo elegir la mejor métrica para evaluarla?
- ¿Qué hacer con todo el conocimiento previo?
- Considerar todos los posibles escenarios
- Datos Balanceados – Edad, Raza, Género..
- ¿Qué efecto tendrá en el empleo?



"Artificial Intelligence in Radiology Ethical Considerations (Brady et al. Diagnostics, 2020)"

Ejemplo de IA: Riesgos de entrenar IA sin filtro

- Una herramienta de lenguaje que sea entrenada con Twitter, Reddit.. “será posiblemente un reflejo del lenguaje vertido en esas redes.

En el 60 % de los casos, la I.A. de generación de texto, GPT-3 creó frases que vinculaban a los musulmanes con disparos, bombas, asesinatos y violencia.

[1] Dave Gershorn, For Some Reason I’m Covered in Blood’: GPT-3 Contains Disturbing Bias Against Muslims, January 2021.

OPT-175B tiene una alta propensión a generar lenguaje tóxico (racista, sexista, violento) y reforzar estereotipos tóxicos.

Computer Science > Computation and Language

[Submitted on 2 May 2022 (v1), last revised 21 Jun 2022 (this version, v4)]

OPT: Open Pre-trained Transformer Language Models

https://arxiv.org/abs/2205.01068

Susan Zhang, Stephen Roller, Naman Goyal, Mikel Artetxe, Moya Chen, Shuohui Chen, Christopher Dewan, Mona Diab, Xian Li, Xi Victoria Lin, Todor Mihaylov, Myle Ott, Sam Shleifer, Kurt Shuster, Daniel Simig, Punit Singh Koura, Anjali Sridhar, Tianlu Wang, Luke Zettlemoyer

Large language models, which are often trained for hundreds of thousands of compute days, have shown remarkable capabilities for zero- and few-shot learning. Given their computational cost, these models are difficult to replicate without significant capital. For the few that are available through APIs, no access is granted to the full model weights, making them difficult to study. We present Open Pre-trained Transformers (OPT), a suite of decoder-only pre-trained transformers ranging from 125M to 175B parameters, which we aim to fully and responsibly share with interested researchers. We show that OPT-175B is comparable to GPT-3, while requiring only 1/7th the carbon footprint to develop. We are also releasing our logbook detailing the infrastructure challenges we faced, along with code for experimenting with all of the released models.

EL PAÍS

Tecnología

TU TECNOLOGÍA · CIBERSEGURIDAD · PRIVACIDAD · INTELIGENCIA ARTIFICIAL · INTERNET · GRANDES TECNOLÓGICAS ·

INTELIGENCIA ARTIFICIAL >

¿Puede una máquina decidir cuánto debes cobrar?

Las empresas recurren a la inteligencia artificial para su política salarial, pero aún es imprescindible la supervisión humana para evitar sesgos o errores

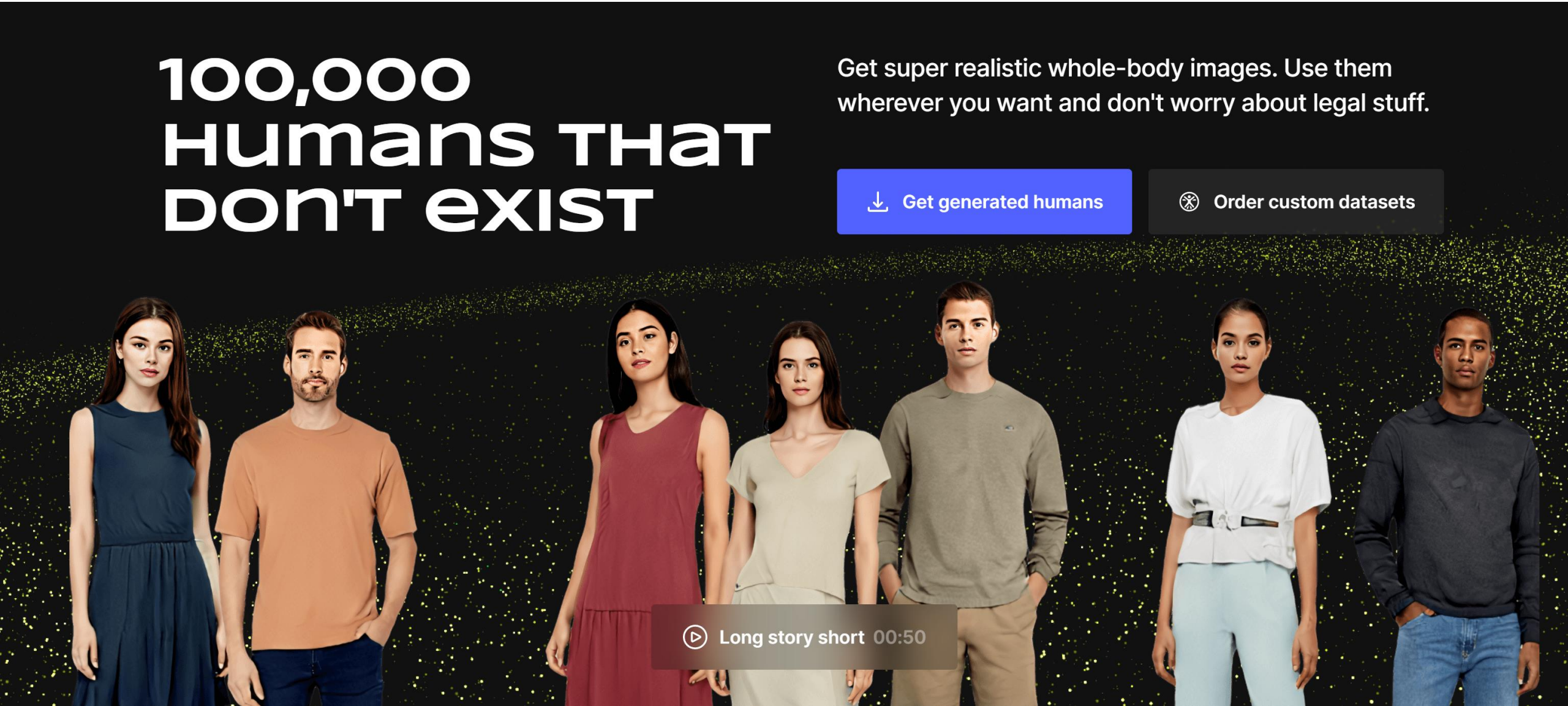
RAÚL LIMÓN

04 OCT 2022 - 05:20 CEST

https://elpais.com/tecnologia/2022-10-04/puede-una-maquina-decidir-cuanto-debes-cobrar.html

https://generated.photos/humans

¿Impacto en Empleo?

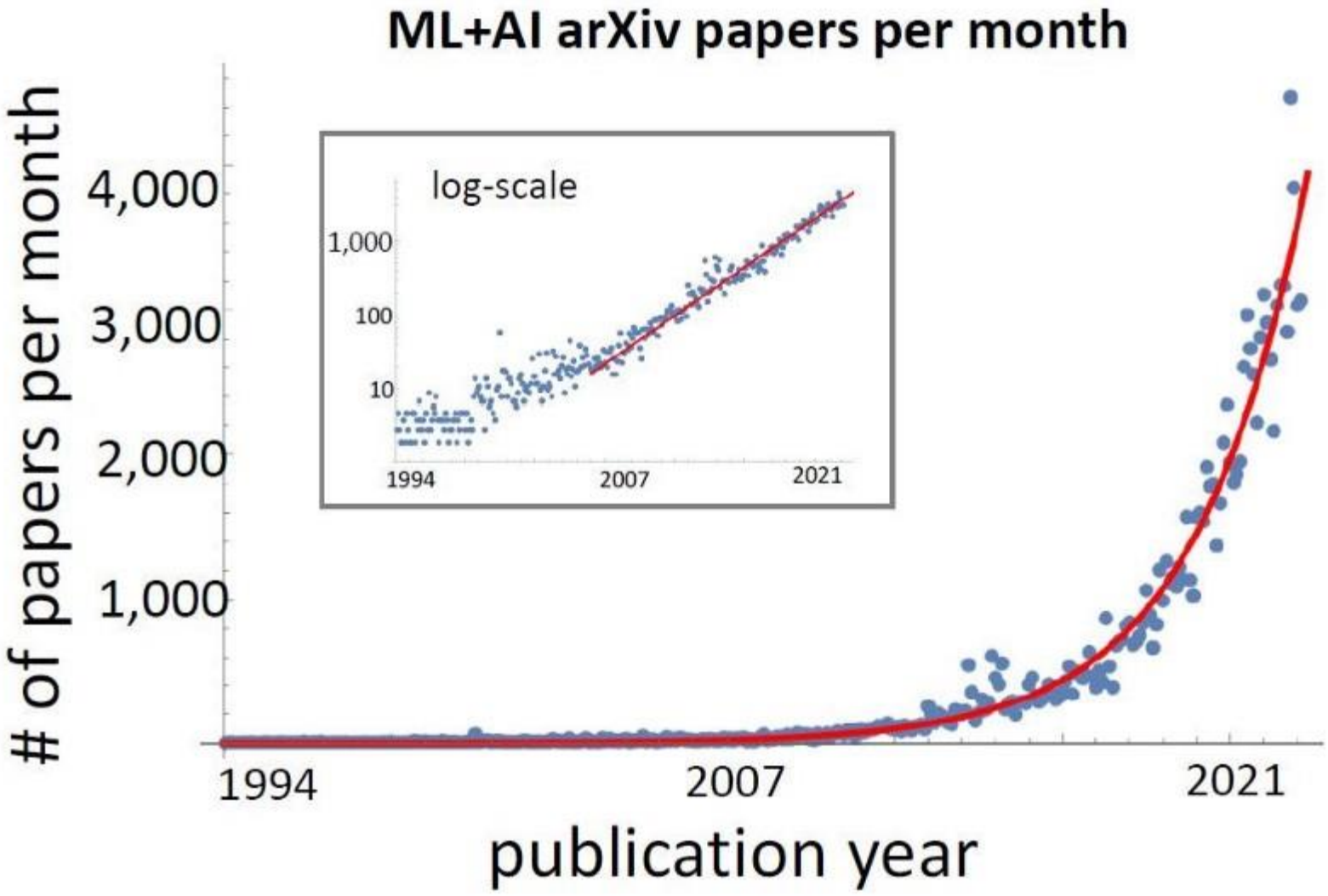


IMPACTO EN DERECHOS DE IMAGEN,Y COPYRIGHT

Imagen generada con Stable Diffusion:
“MRI scanner with Brad Pitt as a patient dressed in a blue gown in a white, bright room”

https://www.pcmag.com/news/ai-image-creator-dall-e-2-comes-to-microsoft-software-including-bing

¿La investigación acelerada en AI está siendo suficientemente rigurosa?



Feature

IS AI LEADING TO A REPRODUCIBILITY CRISIS IN SCIENCE?

Scientists worry that ill-informed use of artificial intelligence is driving a deluge of unreliable or useless research. **By Philip Ball**

22 | Nature | Vol 624 | 7 December 2023

https://towardsdatascience.com/do-you-also-feel-ai-is-going-too-fast-56bf56ab43b3

Nature 7 dic 2023

INVERSIÓN EN MEDICINA VS MILITAR

14.6 USD BILLION 2023-e

102.7 USD BILLION 2028-e

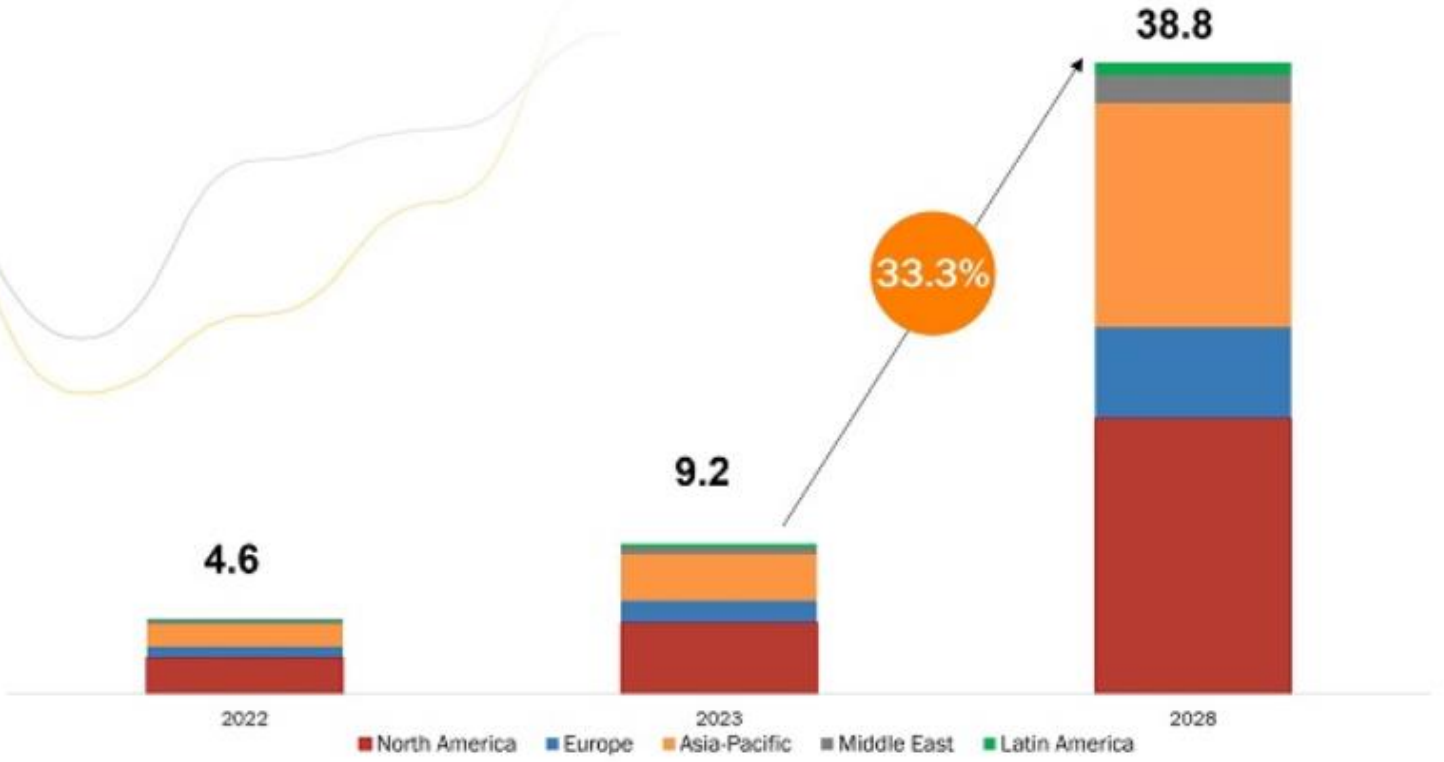
CAGR of 47.6%

The AI in Healthcare market is estimated to reach USD 102.7 billion by 2028, registering a CAGR of 47.6% from 2023 to 2028.

AI IN MILITARY MARKET
GLOBAL FORECAST TO 2028 (USD BN)

CAGR OF
33.3%

The AI in Military Market is expected to be worth USD 38.8 billion by 2028, growing at a CAGR of 33.3% during the forecast period.



Peligros Graves

Dual use of artificial-intelligence-powered drug discovery
Fabio Urbina et al. *Nature Machine Intelligence* **volume 4**, pages189–191 (March 2022)

Una herramienta de I.A. creó 40000 posibles armas químicas en 6 horas

Creación sintética de Virus:
¿Se debería permitir su investigación? ¿Cómo controlarlo?

Manipulación genética:
¿Dónde está el límite?

Distopias

RETO: ¿Existe alguna obra de ciencia ficción que muestre un futuro no distópico?

Experimento: Universo 25

https://psicologiaymente.com/social/universo-25

¿Es necesario poner freno a la IA?

- ¿Será necesario renunciar a ciertos avances de la IA?
- ¿Es posible?

Pause Giant AI Experiments: An Open Letter

We call on all AI labs to immediately pause for at least 6 months the training of AI systems more powerful than GPT-4.

Signatures
33709

Add your signature

Published
March 22, 2023

https://futureoflife.org

nature

Explore content | About the journal | Publish with us | Subscribe

nature > news explainer > article

NEWS EXPLAINER | 23 November 2023

What the OpenAI drama means for AI progress – and safety

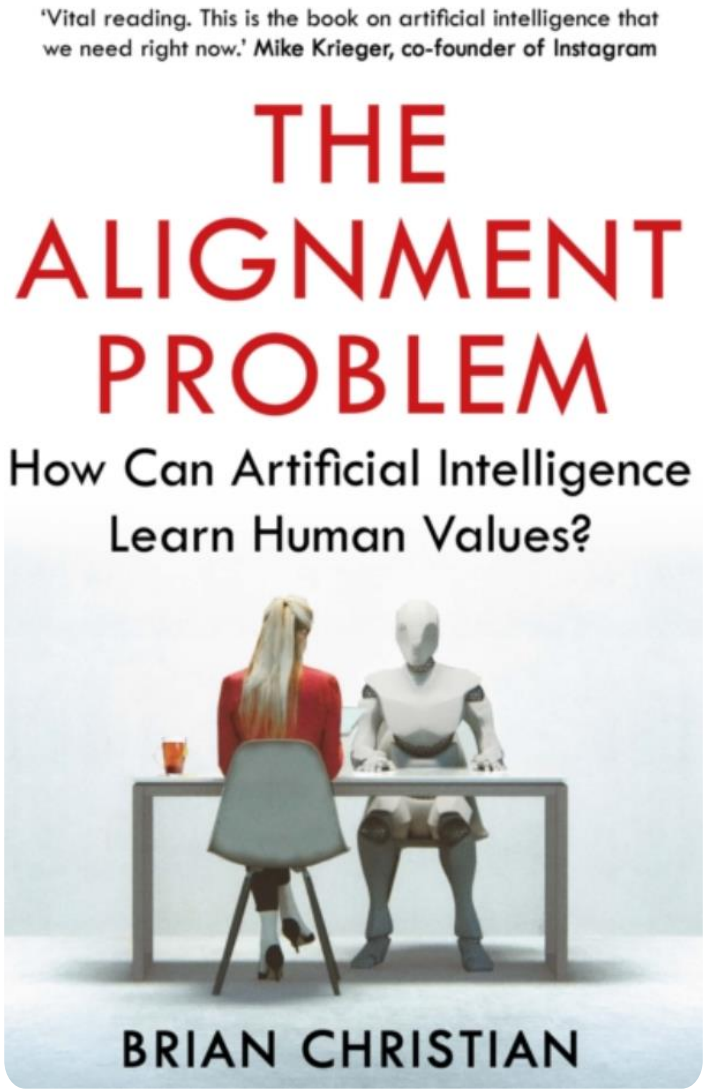
A debacle at the company that built ChatGPT highlights concern that commercial forces are acting against the responsible development of artificial-intelligence systems.

NOTICIA
13-DIC-2023

La ofensiva contra el móvil en los colegios llega al ámbito nacional: el Gobierno ya estudia regularlo

La ministra de Educación ha propuesto este miércoles regular a nivel nacional el uso de teléfonos en periodo lectivo

BLOQUE 3: ¿Qué se puede hacer?



The alignement problem

- ¿Debería una I.A. decir insultos?
- ¿Y comentarios machistas o racistas?
- ¿Y comentarios sobre religión? ¿sexo? ¿fútbol?
- ¿Cómo enseñar que aprenda de los humanos sólo aquellos que consideramos políticamente correcto y acorde con nuestros valores?
- ¿Qué pasa si las conclusiones del análisis de la I.A. están sesgadas?

¿Qué se puede hacer?

Fairlearn | Get Started | User Guide | API Docs | Example Notebooks | Contributor Guide | FAQ | About Us

Improve fairness of AI systems

Fairlearn is an open-source, community-driven project to help data scientists improve fairness of AI systems.

Learn about AI fairness from our guides and use cases. Assess and mitigate fairness issues using our Python toolkit. Join our community and contribute metrics, algorithms, and other resources.

Get Started

https://fairlearn.org/

USE CASE | CREDIT-CARD LOANS

Assessment and mitigation of fairness issues in credit-card default models

When making a decision to approve or decline a loan, financial services organizations use a variety of models, including a model that predicts the applicant's probability of default. These predictions are sometimes used to automatically reject or accept an application, directly impacting both the applicant and the organization.

In this scenario, fairness-related harms may arise when the model makes more mistakes for some groups of applicants compared to others. We use Fairlearn to assess how different groups, defined in terms of their sex, are affected and how the observed disparities may be mitigated.

Jupyter Notebook

Mitigating Unfairness in ML models:

Define fairness metrics based on harms

Specify fairness constraint as a model parameter

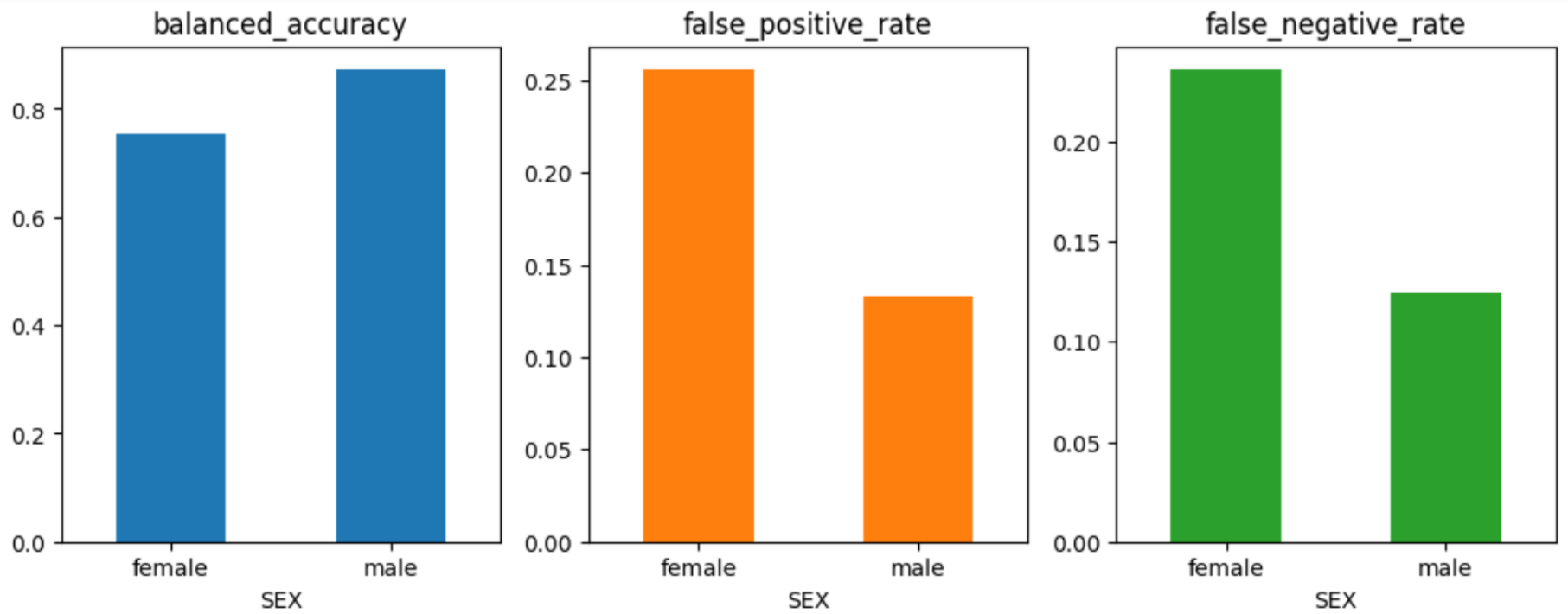
PRÁCTICA

Fairlearn

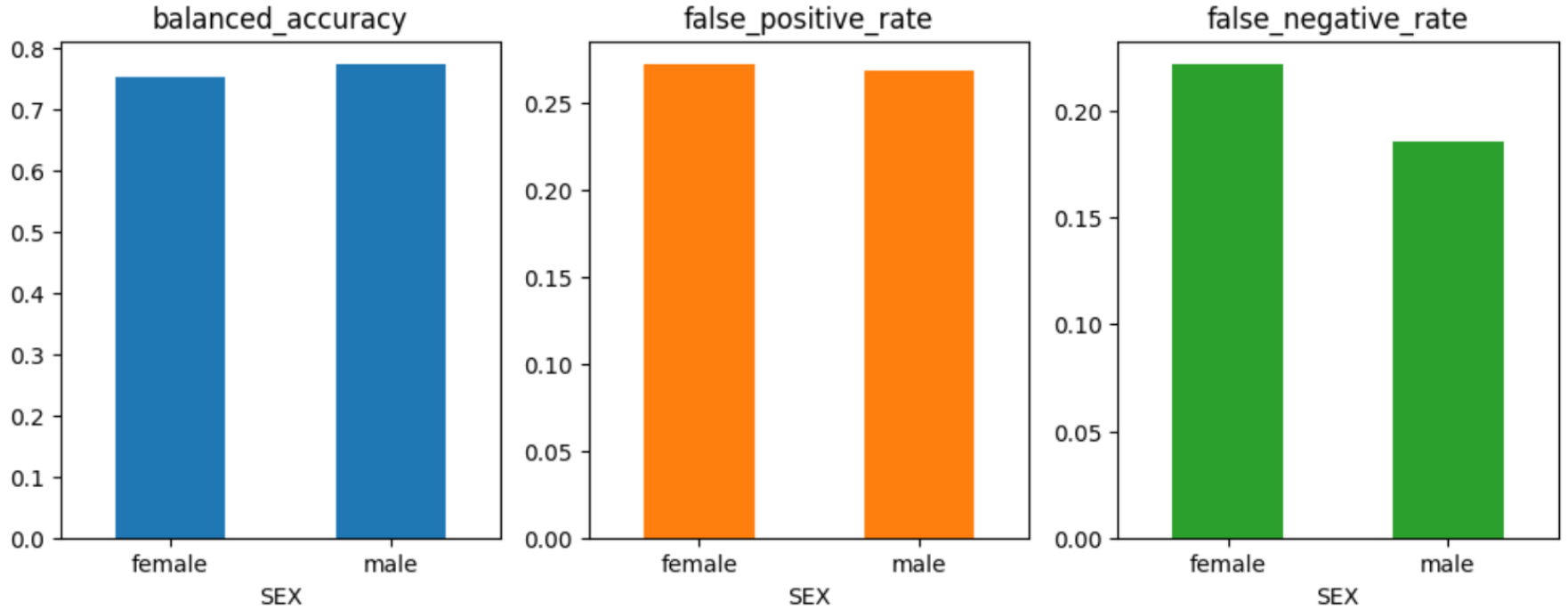
https://t.ly/Uw8SP

PRÁCTICA

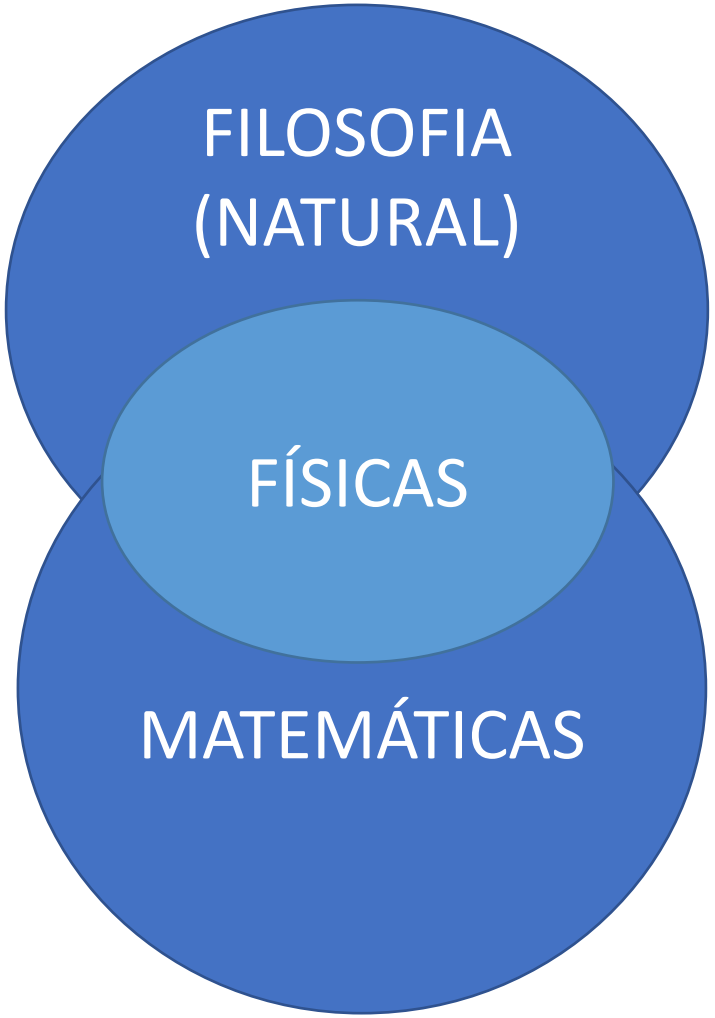
MODELO INJUSTO

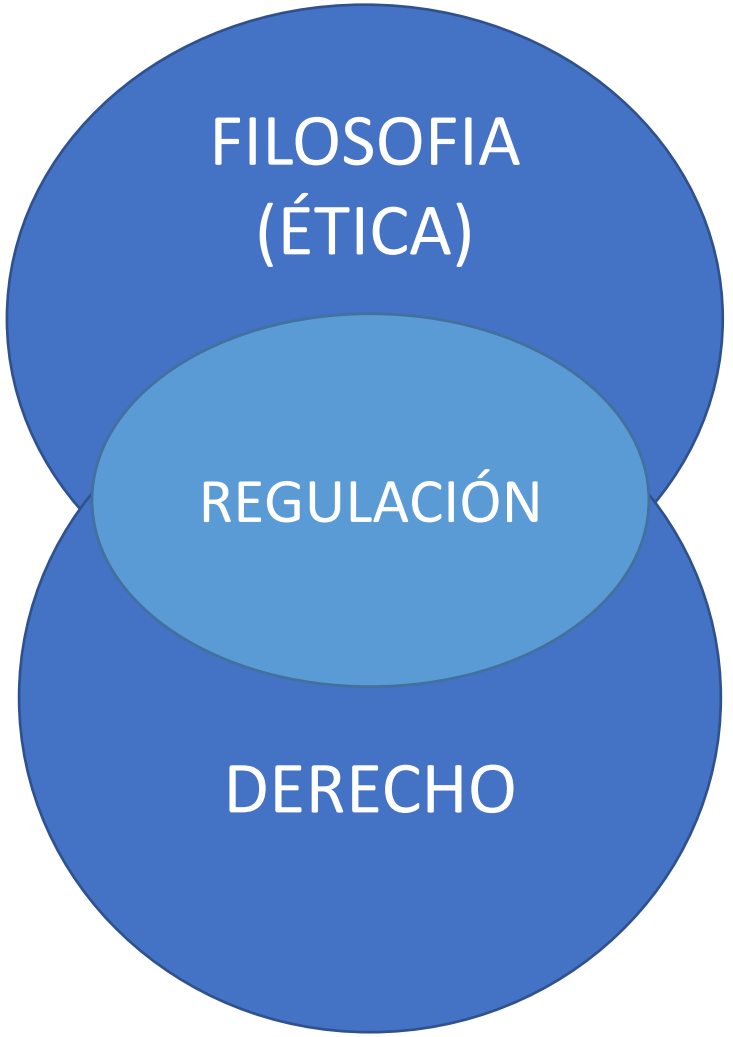


MODELO JUSTO



BLOQUE 4: ÉTICA Y REGULACIÓN





Recomendaciones, Autoregulación y Regulación

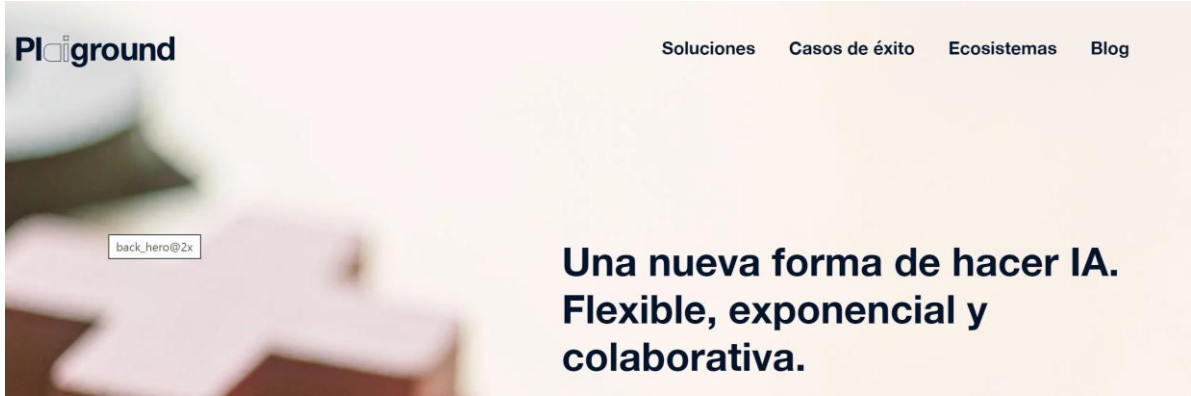
- **Recomendaciones** de las instituciones internacionales y de las colaboraciones público-privadas. UNESCO y otras.
- **Autorregulación** - Empresas consideran, detectan y mitigan los posibles riesgos sociales y éticos de su actividad con la IA. Los riesgos éticos y sociales se tienen en cuenta durante todo el ciclo de vida del sistema y, cuando se detectan, se mitigan mediante acciones específicas. Si es imposible mitigar los riesgos previstos, las empresas podrían y deberían abstenerse de desplegarlo hasta que se conozcan mejor los riesgos y se puedan evitar.
- **Regulación –** Ciertos usos de la tecnología están regulados por ley. El RGPD (Reglamento general de protección de datos) europeo y la Regulación de IA (EU I.A. ACT).

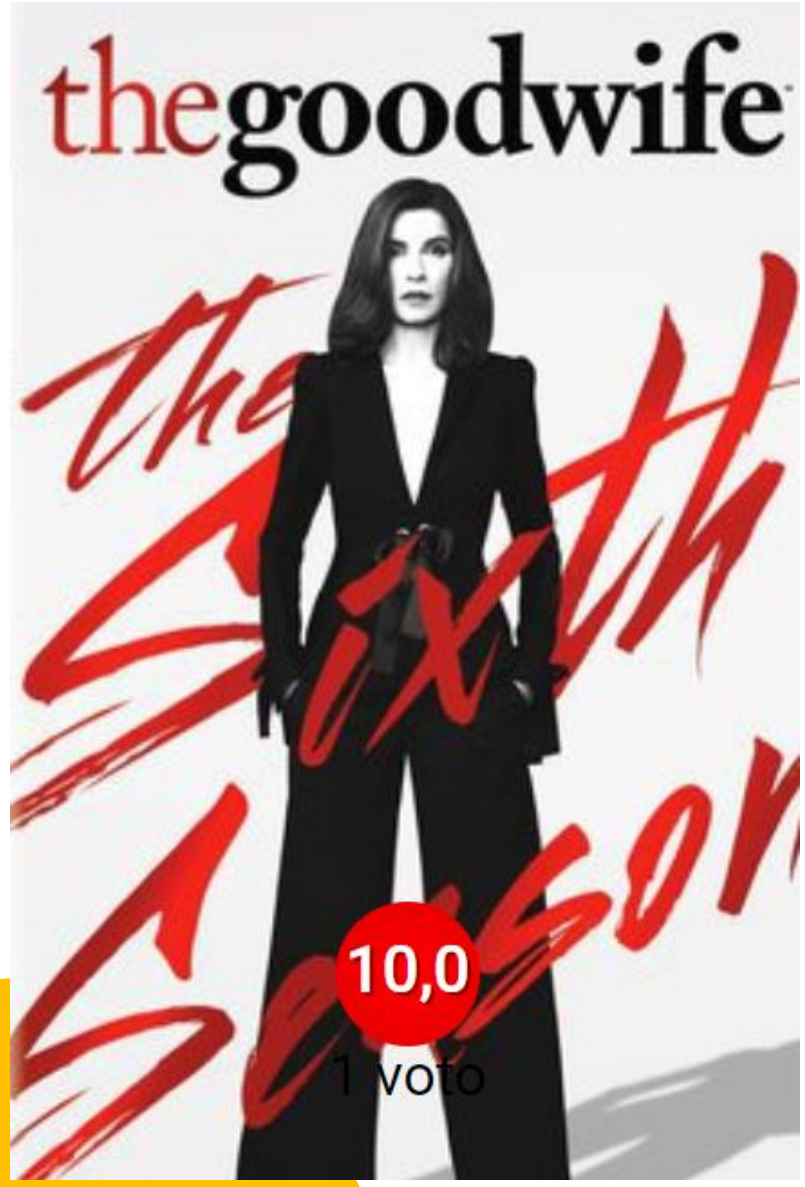
I.A. responsable como parte de la política de la empresa

septiembre 2022

IA Responsable: así podemos hacer que nuestra Inteligencia Artificial impacte positivamente

https://plaiground.ai/es/ia-responsable-asi-podemos-hacer-que-nuestra-inteligencia-artificial-impacte-positivamente/

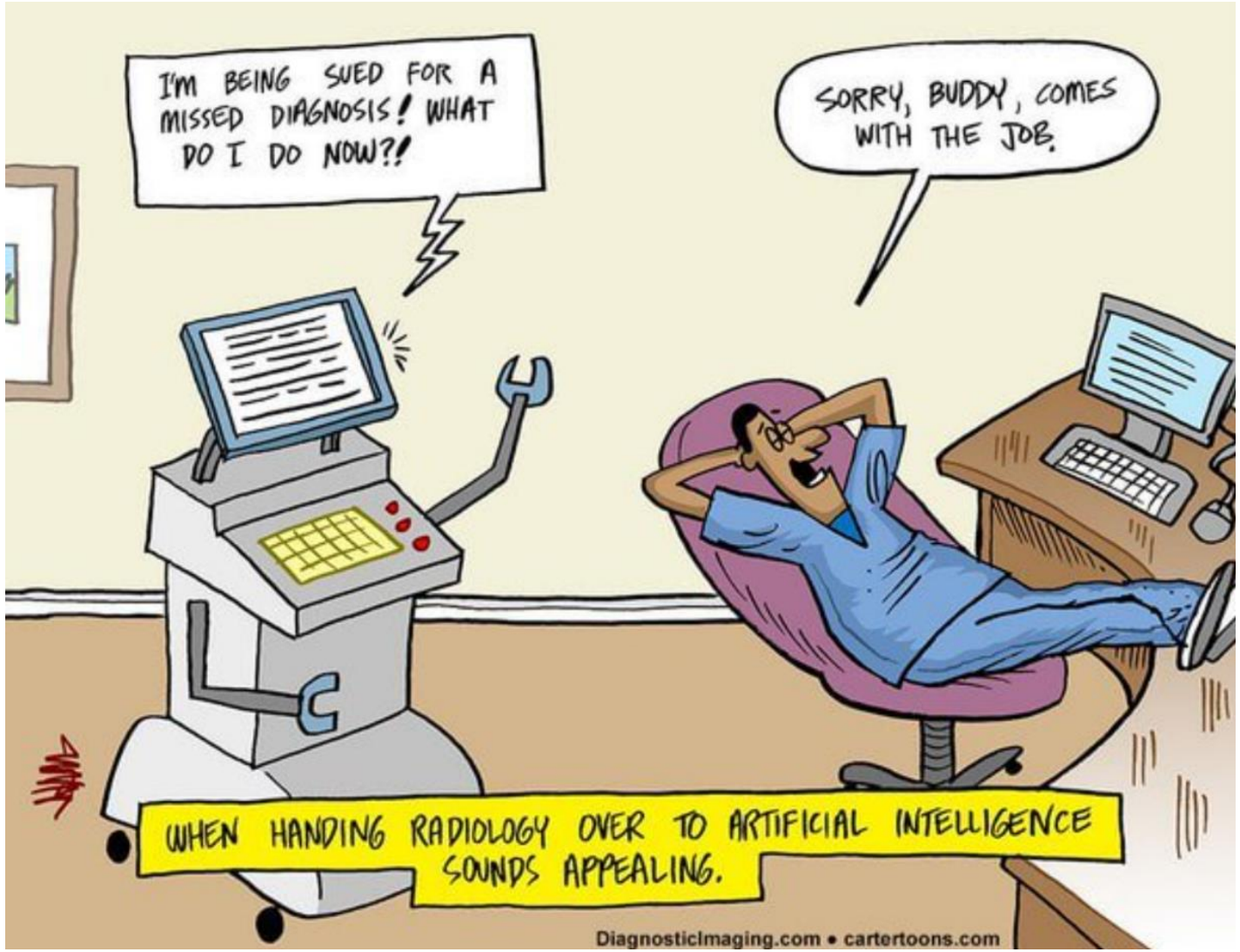




Ética y Regulacón

- ¿Quién es responsable ante la ley de las recomendaciones / acciones de una I.A.?
  - Programador
  - Usuario
  - Regulador
  - Sociedad
  - Otros
- Ejemplos: Recomendación sesgada de rutas en coche, accidentes,…

Regulación de IA de la FDA



https://www.fda.gov/medical-devices/software-medical-device-samd/artificial-intelligence-and-machine-learning-software-medical-device

A.I. ACT

https://artificialintelligenceact.eu

What is the EU AI Act?

The AI Act is a proposed European law on artificial intelligence (AI) – the first law on AI by a major regulator anywhere. The law assigns applications of AI to three risk categories. First, applications and systems that create an **unacceptable risk**, such as government-run social scoring of the type used in China, are banned. Second, **high-risk applications**, such as a CV-scanning tool that ranks job applicants, are subject to specific legal requirements. Lastly, applications not explicitly banned or listed as high-risk are largely left unregulated.

septiembre 2022

La opinión del experto: AI Act

EU Artificial Intelligence Act

https://plaiground.ai/es/la-opinion-del-experto-ai-act/

Leticia Gomez Rivero

SHARE THIS POST

Javier Valls, evaluador de Ética de la IA en la Comisión Europea, comparte con nosotros su visión sobre la nueva normativa AI Act que entrará en vigor en 2023 y cómo afectará a las organizaciones europeas.

/eu-

CIENCIA | INTELIGENCIA ARTIFICIAL

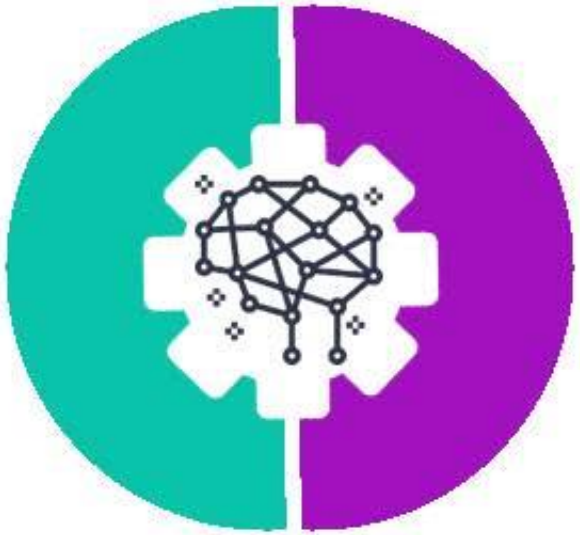
La UE pacta la primera ley sobre inteligencia artificial del mundo: ¿qué significa?

La Ley de Inteligencia Artificial, conocida como IA Act, se presentó por primera vez en abril de 2021. Uno de los aspectos más controvertidos ha sido el uso de sistemas de identificación biométrica, dadas sus implicaciones en el control gubernamental y los derechos ciudadanos.

Actualizado a 12 de diciembre de 2023, 08:36

AI is good …

- For citizens
- For business
- For the public interest

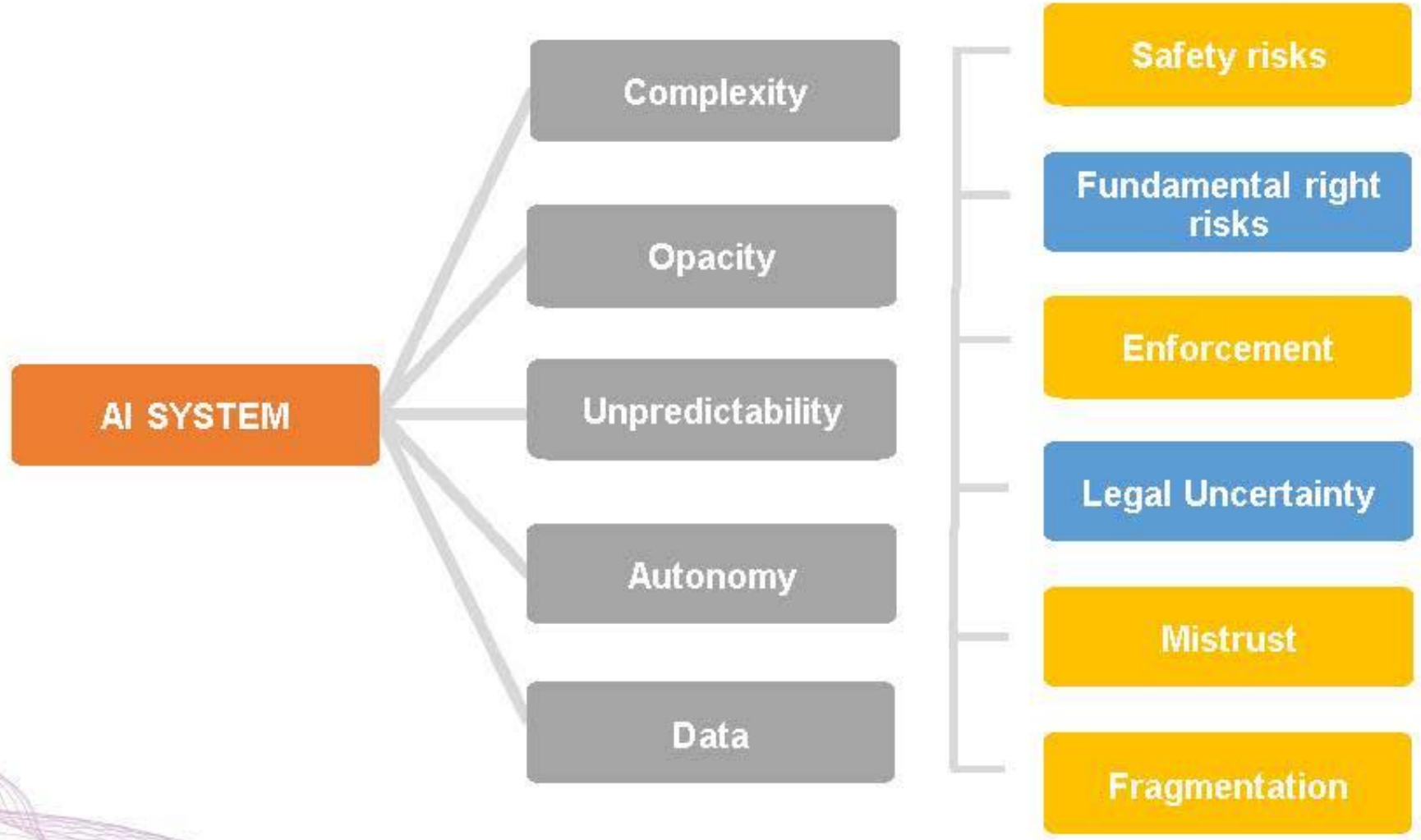


… but creates some risks

- For the safety of consumers and users
- For fundamental rights

European Commission

Why do we regulate AI use cases?



European Commission

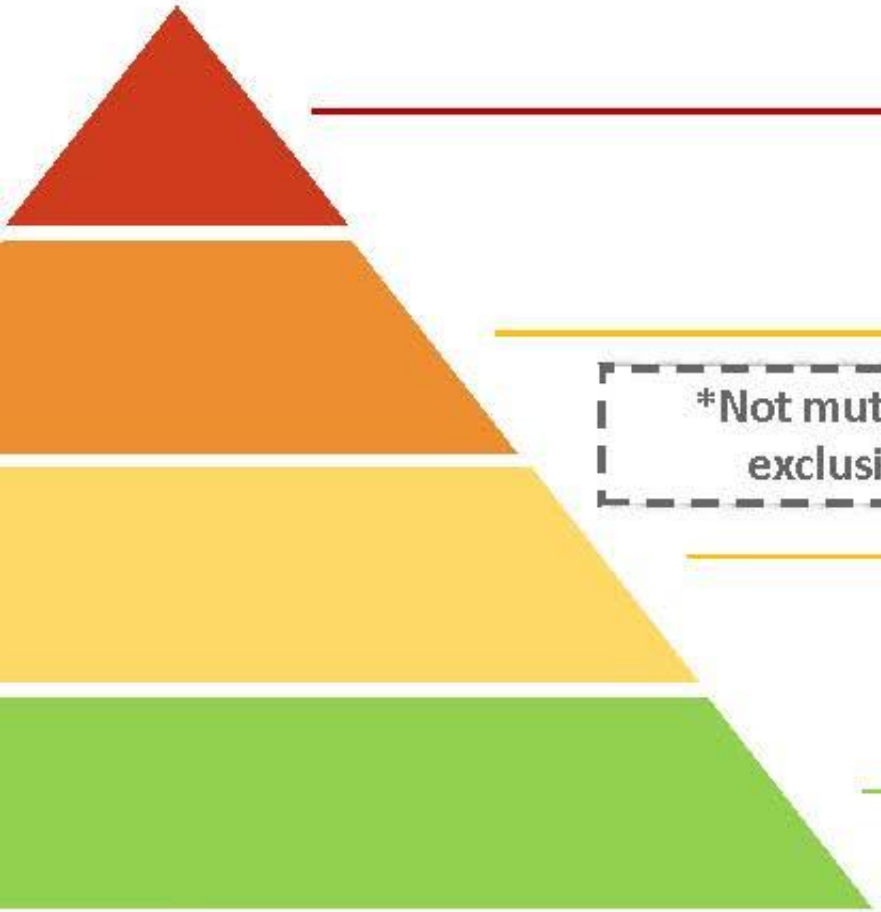
Definition and technological scope of the regulation (Art. 3)

Definition of Artificial Intelligence

- Definition of AI should be **as neutral as possible** in order to cover techniques which are not yet known/developed
- **Overall aim is to cover all AI**, including traditional symbolic AI, Machine learning, as well as hybrid systems
- **Annex I:** list of AI techniques and approaches should provide for legal certainty (adaptations over time may be necessary)

> *"a software that is developed with one or more of the techniques and approaches listed in Annex I and can, for a given set of human-defined objectives, generate outputs such as content, predictions, recommendations, or decisions influencing the environments they interact with"*

A risk-based approach to regulation



*Not mutually exclusive

Unacceptable risk
e.g. social scoring

Prohibited

High risk
e.g. recruitment, medical devices

Permitted subject to compliance with AI requirements and ex-ante conformity assessment

AI with specific transparency obligations
'Impersonation' (bots)

Permitted but subject to information/transparency Obligations

Minimal or no risk

Permitted with no restrictions

European Commission

Most AI systems will not be high-risk (Titles IV, IX)

New transparency obligations for certain AI systems (Art. 52)

- **Notify humans** that they are **interacting with an AI system** unless this is evident
- Notify humans that emotional recognition or biometric categorisation systems are applied to them
- Apply **label to deep fakes** (unless necessary for the exercise of a fundamental right or freedom or for reasons of public interests)

Possible voluntary codes of conduct for AI with specific transparency requirements (Art. 69)

- No mandatory obligations
- Commission and Board to encourage drawing up of codes of conduct intended to foster the **voluntary application of requirements to low-risk AI systems**

High-risk Artificial Intelligence Systems (Title III, Annexes II and III)

Certain applications in the following fields:

1 **SAFETY COMPONENTS OF REGULATED PRODUCTS**

(e.g. medical devices, machinery) which are subject to third-party assessment under the relevant sectorial legislation

2 **CERTAIN (STAND-ALONE) AI SYSTEMS IN THE FOLLOWING FIELDS**

- ✓ Biometric identification and categorisation of natural persons
- ✓ Management and operation of critical infrastructure
- ✓ Education and vocational training
- ✓ Employment and workers management, access to self-employment
- ✓ Access to and enjoyment of essential private services and public services and benefits
- ✓ Law enforcement
- ✓ Migration, asylum and border control management
- ✓ Administration of justice and democratic processes

Requirements for high-risk AI (Title III, chapter 2)

Establish and implement **risk management** processes

&

In light of the **intended purpose** of the AI system

- Use high-quality **training, validation and testing data** (relevant, representative etc.)
- Establish **documentation** and design logging features (traceability & auditability)
- Ensure appropriate certain degree of **transparency** and provide users with **information** (on how to use the system)
- Ensure **human oversight** (measures built into the system and/or to be implemented by users)
- Ensure **robustness**, **accuracy** and **cybersecurity**

Overview: obligations of operators (Title III, Chapter 3)

Provider obligations

- Establish and Implement **quality management** system in its organisation
- Draw-up and keep up to date **technical documentation**
- **Logging** obligations to enable users to monitor the operation of the high-risk AI system
- Undergo **conformity assessment** and potentially re-assessment of the system (in case of significant modifications)
- Register AI system in EU database
- Affix CE marking and sign declaration of conformity
- Conduct **post-market monitoring**
- **Collaborate** with market surveillance authorities

User obligations

- Operate AI system in accordance with **instructions of use**
- Ensure **human oversight** when using of AI system
- **Monitor** operation for possible risks
- **Inform the provider or distributor about any serious incident** or any malfunctioning
- **Existing legal obligations** continue to apply (e.g. under GDPR)

Lifecycle of AI systems and relevant obligations

Design in line with requirements ▸ Ensure AI systems **perform consistently for their intended purpose** and are **in compliance with the requirements** put forward in the Regulation

Conformity assessment ▸ **Ex ante** conformity assessment

Post-market monitoring ▸ Providers to **actively and systematically collect, document and analyse relevant data** on the reliability, performance and safety of AI systems throughout their lifetime, and to **evaluate continuous compliance of AI systems with the Regulation**

Incident report system ▸ **Report serious incidents as well as malfunctioning leading to breaches to fundamental rights** (as a basis for investigations conducted by competent authorities).

New conformity assessment ▸ **New conformity assessment** in case of **substantial modification** (modification to the intended purpose or change affecting compliance of the AI system with the Regulation) by providers or any third party, including when changes are **outside the "predefined range" indicated by the provider for continuously learning AI systems.**

AI that contradicts EU values is prohibited (Title II, Article 5)

✗ **Subliminal manipulation** resulting in physical/ psychological harm

Example: An **inaudible sound** is played in truck drivers' cabins to push them to **drive longer than healthy and safe**. AI is used to find the frequency maximising this effect on drivers.

✗ **Exploitation of children or mentally disabled persons** resulting in physical/psychological harm

Example: A doll with an integrated **voice assistant** encourages a minor to **engage in progressively dangerous behavior** or challenges in the guise of a fun or cool game.

✗ **General purpose social scoring**

Example: An AI system **identifies at-risk children** in need of social care **based on insignificant or irrelevant social 'misbehavior'** of parents, e.g. missing a doctor's appointment or divorce.

✗ **Remote biometric identification for** law enforcement purposes in publicly accessible spaces (with exceptions)

Example: All faces captured live by video cameras checked, in real time, against a database to identify a terrorist.

A.I. ACT: MEJORAS POSIBLES

How can it be improved?

There are **several loopholes and exceptions** in the proposed law. These shortcomings limit the Act's ability to ensure that AI remains a force for good in your life. Currently, for example, facial recognition by the police is banned unless the images are captured with a delay or the technology is being used to find missing children.

In addition, **the law is inflexible**. If in two years' time a dangerous AI application is used in an unforeseen sector, the law provides no mechanism to label it as "high-risk". For more detailed analyses, please see this page.

Future of Life Institute

The Future of Life Institute (FLI), an independent nonprofit with the goal of maximising the benefits of technology and reducing its associated risks, shared its recommendations for the EU AI Act with the European Commission. It argues that the Act should ensure that AI providers consider the impact of their applications on society at large, not just the individual. AI applications that cause negligible harms to individuals could cause significant harms at the societal level. For example, a marketing application used to influence citizens' voting behaviour could affect the results of elections. Read more of the recommendations here.

University of Cambridge institutions

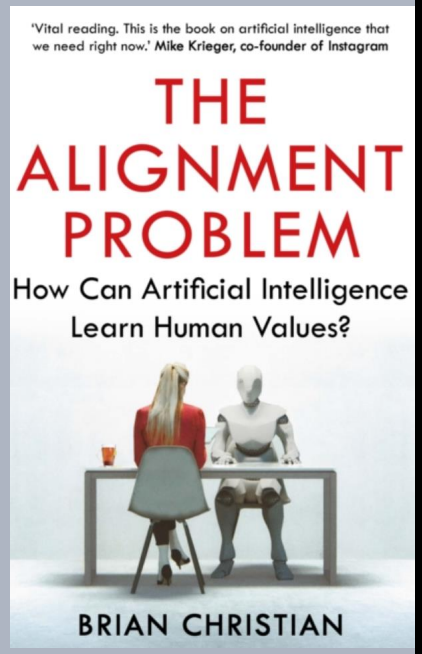
Leverhulme Centre for the Future of Intelligence and Centre for the Study of Existential Risk, two leading institutions at the University of Cambridge, provided their feedback on the proposed EU AI law to the European

REFERENCIAS:

- The alignment problem – Brian Christian - 2021
- *https://fairlearn.org/*
- https://artificialintelligenceact.eu/

EXTRA: EJEMPLOS DE APLICACIONES

- https://indico.cern.ch/event/1283970



CONCLUSIONES: Toda I+D+i tiene consecuencias No debemos trabajar sin ser conscientes de ello

- Eso es especialmente importante con tecnologías punteras como la I.A.
- ¿Qué consecuencias (positivas y negativas) tendrá aquello que estamos desarrollando?
  - ¿Cómo podemos potenciar sus efectos positivos?
  - ¿Cómo podemos mitigar/adaptarse a sus efectos negativos?
  - Hay que anticiparse para realizar una I+D+I responsible.