INTELIGENCIA ARTIFICIAL

La inteligencia artificial silenciosa ha entrado en nuestras vidas

Francisco Herrera

DaSCI

UNIVERSIDAD DE GRANADA

Inteligencia artificial

Inteligencia artificial generativa

Eclosión de la IA, con riesgos

Etica y regulación. Auditabilidad

IA fiable y segura

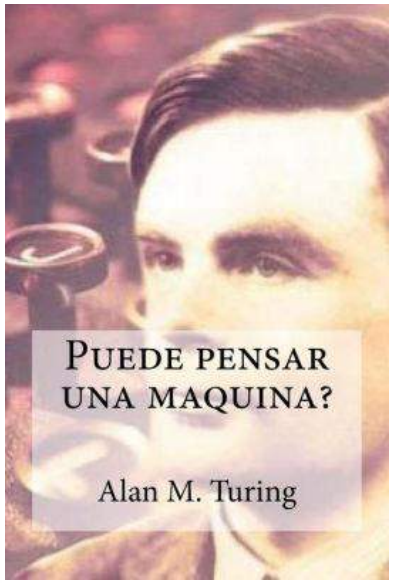
Concluyendo

Inteligencia artificial

John McCarthy (1955) (Stanford)

Conferencia de Dartmouth (1956)

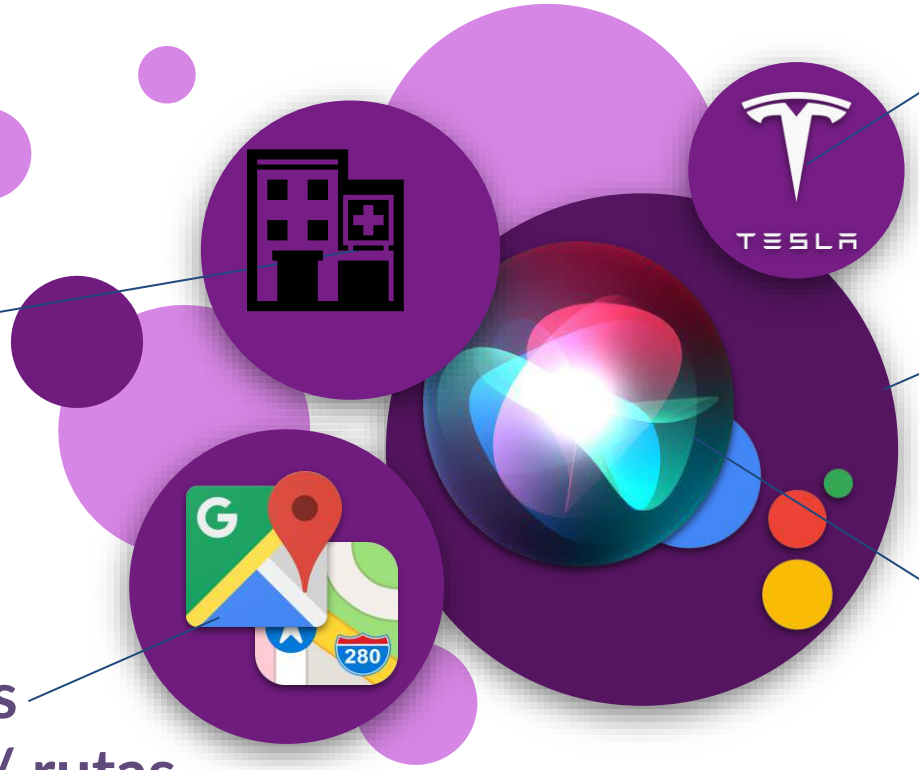
Inteligencia Artificial: "la ciencia e ingeniería de hacer máquinas que se comporten de una forma que llamaríamos inteligente si el humano tuviese ese comportamiento"



Alan Turing
Pionero de la IA
Test de Turing

Inteligencia artificial silenciosa, ha entrado en nuestras vidas

1996 | Primer robot aspirador:
Electrolux Trilobite



Automoción

Diagnóstico
de enfermedades

Sistema de
Recomendaciones

amazon
NETFLIX

Asistentes
de navegación / rutas

Asistentes virtuales: Siris,
Alexa, google Assistant, ...

Inteligencia artificial silenciosa, ha entrado en nuestras vidas

La inteligencia artificial ha entrado en el salón de casa.

Radio años 20 (siglo XX)
Televisión años 50 (siglo XX)
IA años 20 (siglo XXI) (conversaciones, recomendaciones, …)

Breve recorrido histórico. Inteligencia artificial. 70 años de historia

"Máquinas no pensantes cada vez más capaces"

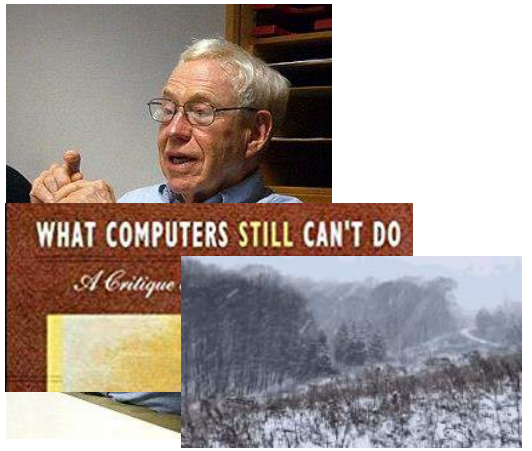
Edad de ORO
1960 - Años sesenta

Invierno de la IA
1970 - Años setenta

1980 – IA débil (propósito específico)
La inteligencia artificial empieza a rendir y asombrar

Grandes Expectativas

HAL 9000
2001: Una odisea del espacio
Stanley Kubrick
1968



Hubert Dreyfus

Ajedrez

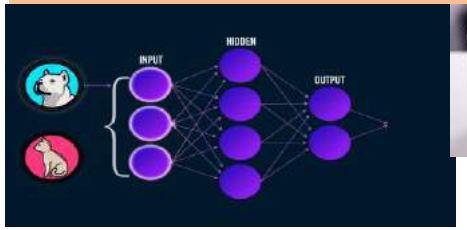
1996 – 1997
Kasparov vs. Deep Blue

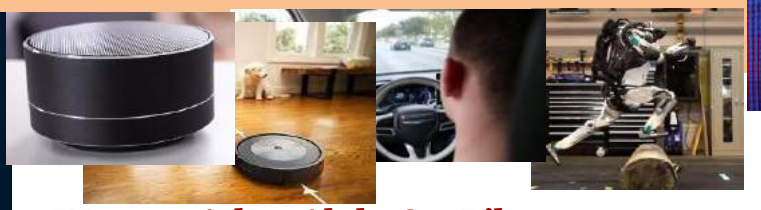
Coche Autónomo

2005 - Stanley (Standord) DARPA grand Challenge (S. Thrun)

Desierto de Mojave,
212 kilómetros, 6 horas y 54 minutos

2001 – … Inteligencia Artificial entra en nuestras vidas. Era del Big Data (macrodatos). Deep Learning. El poder de los datos





Watson, Atlas, AlphaGo, Libratus…

2022 - Era de la IA Generativa

Inteligencia artificial. Invierno de la IA

Expectativas no cumplidas en los 60, “IA General o fuerte”, grandes retos, asistentes conversacionales y sistemas inteligentes al nivel de los humanos, ej. Hal 9000

Hal 9000: ejemplo de IA General o IA Fuerte

Inteligencia artificial. Impulso (IA de propósito específico, IA débil)

Años 80 y 90, impulso en base a abordar problemas complejos concretos, IA débil. Gran hito ajedrez 96-97 (DeepBlue, ej. de "IA débil", solo juega al ajedrez)

El día que los microchips vencieron a las neuronas:

10/2/1996

Deep Blue (una máquina que analizaba 100 millones de jugadas por segundo) derrotó al entonces campeón mundial en la primera partida del match; el humano se repuso y ganó el duelo. Cómo recuerda aquel encuentro el ajedrecista nacido en Bakú

■ Ajedrez

La máquina despiadada

'Deep Blue' vence a Kasparov en diecinueve movimientos

EFE • NUEVA YORK

El ordenador 'Deep Blue', que ayer conducía las piezas blancas, derrotó y humilló al campeón del mundo de la Asociación Profesional de Ajedrez (PCA), el ruso Garí Kasparov, en 19 movimientos, en la sexta y última partida del encuentro que han disputado en Nueva York.

El estupor fue general. Público, aficionados, maestros, y seguidores de la partida a través de Internet en todo el mundo, no podían dar crédito a lo que estaba pasando en la sexta partida del encuentro entre la máquina y el hombre. La partida comenzó con la variante Nimzovitch de la antigua defensa Caro Kan, desarrollada por los maestros Caro y Kan a finales del siglo pasado y que es una de las preferidas de Anatoli Karpov y otros jugadores de élite.

Sorpresa

La sorpresa llegó en la octava jugada con un sacrificio de la máquina que, aunque es teórico, no se suele practicar en el ajedrez de alto nivel por los riesgos que conlleva de partida abierta y de difícil cálculo. Pero la segunda sorpresa, aún mayor, llegó en el movimiento 17 cuando Kasparov se dejó la dama a cambio de una torre y un alfil pero lo peor es que no había contrajuego debido a la mala situación del rey de Kasparov en el centro del tablero. La jugada 19 de 'Deep Blue' –c4– dejaba al campeón del mundo sin esperanza alguna, ni siquiera de aspirar a las tablas. Una derrota sorprendente ante el mejor del mundo. ¿Error humano, cansancio, 'stress', demasiada confianza en sí mismo?

REUTERS

Kasparov muestra su desesperación frente al ordenador.

12/5/1997

Inteligencia artificial. Cambio de paradigma

“Cambio de paradigma, desde algoritmo al dato. La era del big data (macrodatos)”

2005 - Driverless car

Stanley (Standord), DARPA grand Challenge - 2005 (S. Thrun)

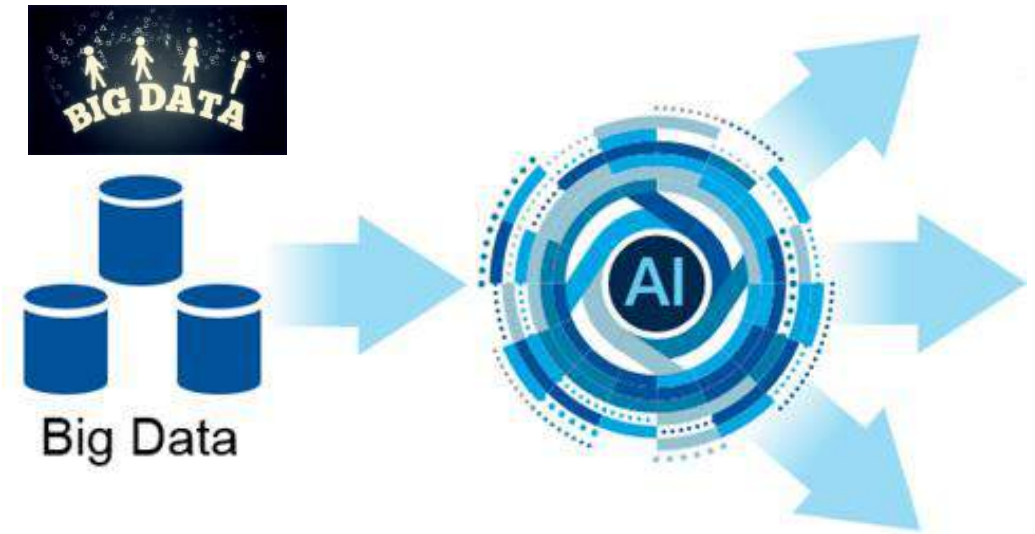
Desierto de Mojave,

212 kilómetros, 6 horas y 54 minutos

https://elpais.com/tecnologia/2005/10/10/actualidad/1128932879_850215.html

Datos e Inteligencia Artificial

El Big Data aporta una enorme cantidad de datos que alimentan los algoritmos de Machine Learning (Inteligencia Artificial), y permiten crear modelos más complejos y con una mayor precisión.



Sistema Inteligente

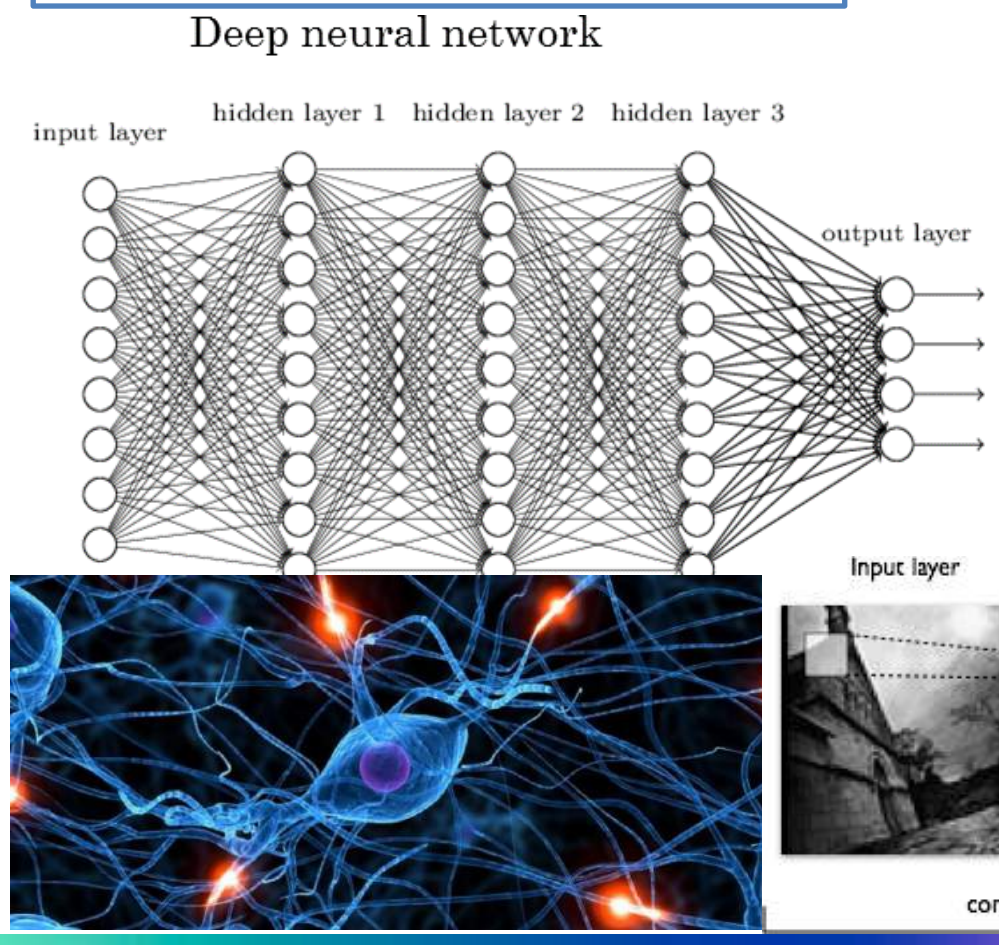
Conocimiento

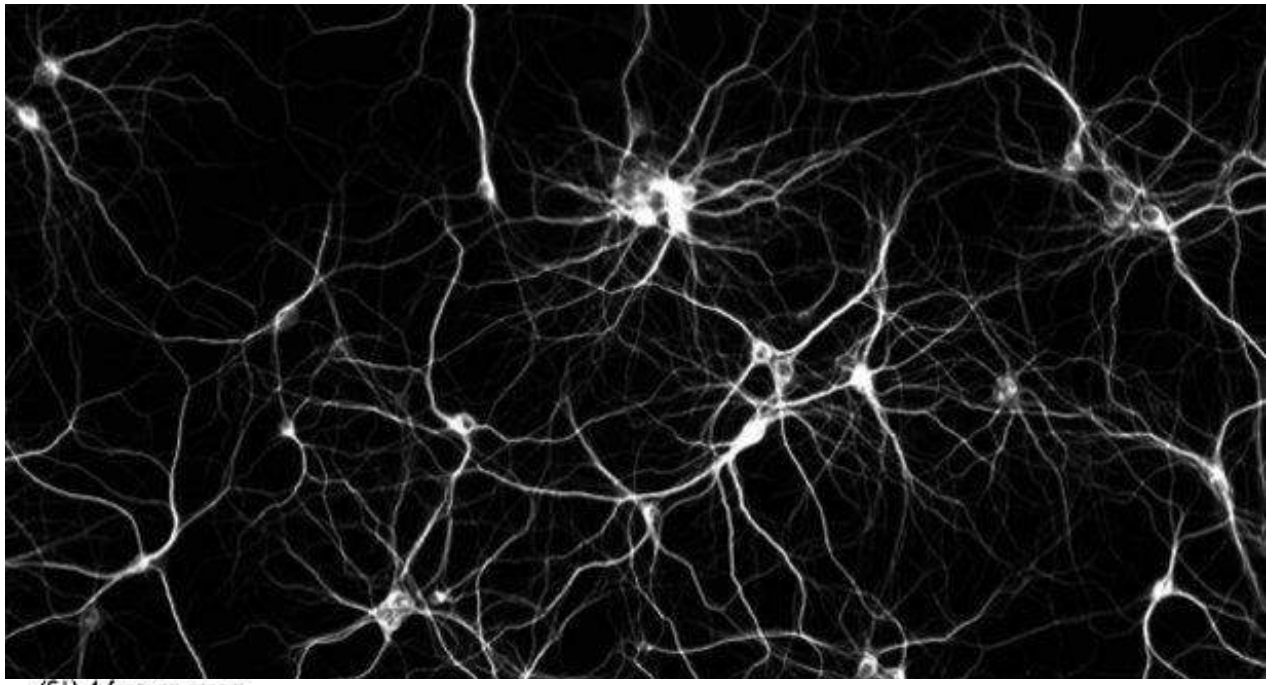
Aplicaciones

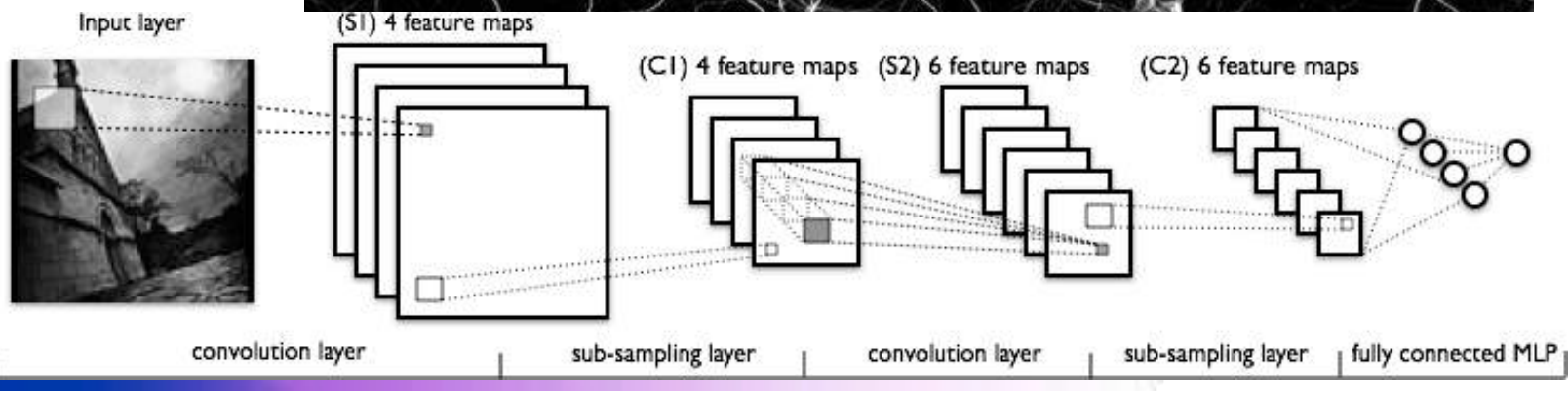
Sistemas Inteligentes alimentados por datos para convertir los datos en conocimiento

Inteligencia artificial, la era del big data. Datos e IA

Deep learning







Inteligencia artificial, la era del big data. El poder de los datos

Unos días después el director llamó al padre para disculparse.

Respuesta conciliadora del padre:

"He estado hablando con mi hija –dijo el padre– Resulta que en mi casa han tenido lugar ciertas actividades de las que yo no estaba del todo informado. Mi hija sale de cuentas en agosto. Soy yo el que les debe una disculpa".

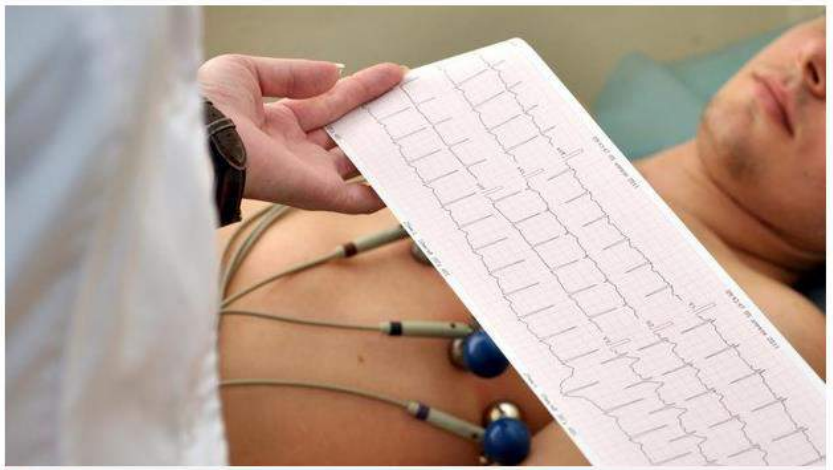
Fuente: Big Data. La revolución de los datos masivos.

Inteligencia artificial, la era del big data. El poder de los datos

Una IA predice con éxito si moriremos pronto, aunque nadie sabe cómo lo hace

MADRID Actualizado:15/11/2019 02:13h

- La Inteligencia Artificial logró predecir la muerte de personas, en el plazo de un año, con solo ver sus electrocardiogramas



El software se basa en los resultados de pruebas cardíacas - Adobe Stock

Fornwalt y sus colegas "alimentaron" a la Inteligencia Artificial con una gran cantidad de datos históricos: 1,77 millones de resultados de electrocardiogramas de casi 400.000 personas, y le pidieron que predijera quiénes tenían más probabilidades de morir en los 12 meses siguientes. Un electrocardiograma registra la actividad eléctrica del corazón, y sus patrones cambian debido a las afecciones cardíacas, incluidos los infartos y la fibrilación auricular.

Una **Inteligencia Artificial** acaba de demostrar que es capaz de predecir, con pasmosa exactitud, **las posibilidades de que una persona muera en el plazo de un año**, basándose únicamente en los resultados de sus pruebas cardíacas. El sistema fue incluso capaz de anunciar la muerte de pacientes con valores que los médicos habían considerado normales. Cómo se las arregla la IA para predecir estas muertes resulta un misterio. El impactante estudio, dirigido por Brandon Fornwalt, del Centro Médico Geisinger, en Pennsylvania, se presentará este mismo sábado durante las sesiones científicas de la American Heart Association, en Dallas.

https://www.abc.es/ciencia/abci-predice-exito-si-moriremos-pronto-aunque-nadie-sabe-como-hace-201911141623_noticia.html
https://www.nature.com/articles/s41591-020-0870-z

Inteligencia artificial

Inteligencia artificial generativa

Eclosión de la IA, con riesgos

Etica y regulación. Auditabilidad

IA fiable y segura

Concluyendo

Inteligencia artificial generativa: 2022 eclosión

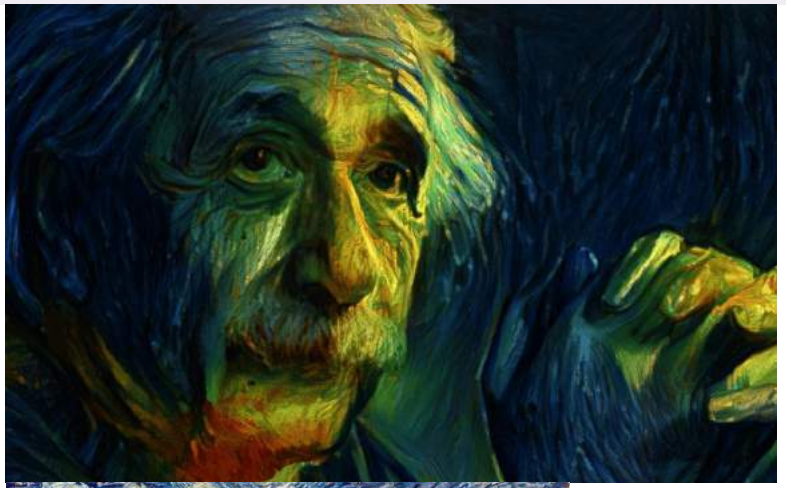
En 2022, la inteligencia artificial (IA) se volvió creativa.

Inteligencia Artificial GENERATIVA

2022 Creación a partir de instrucciones "Prompt"

Midjourney

DALL·E 2



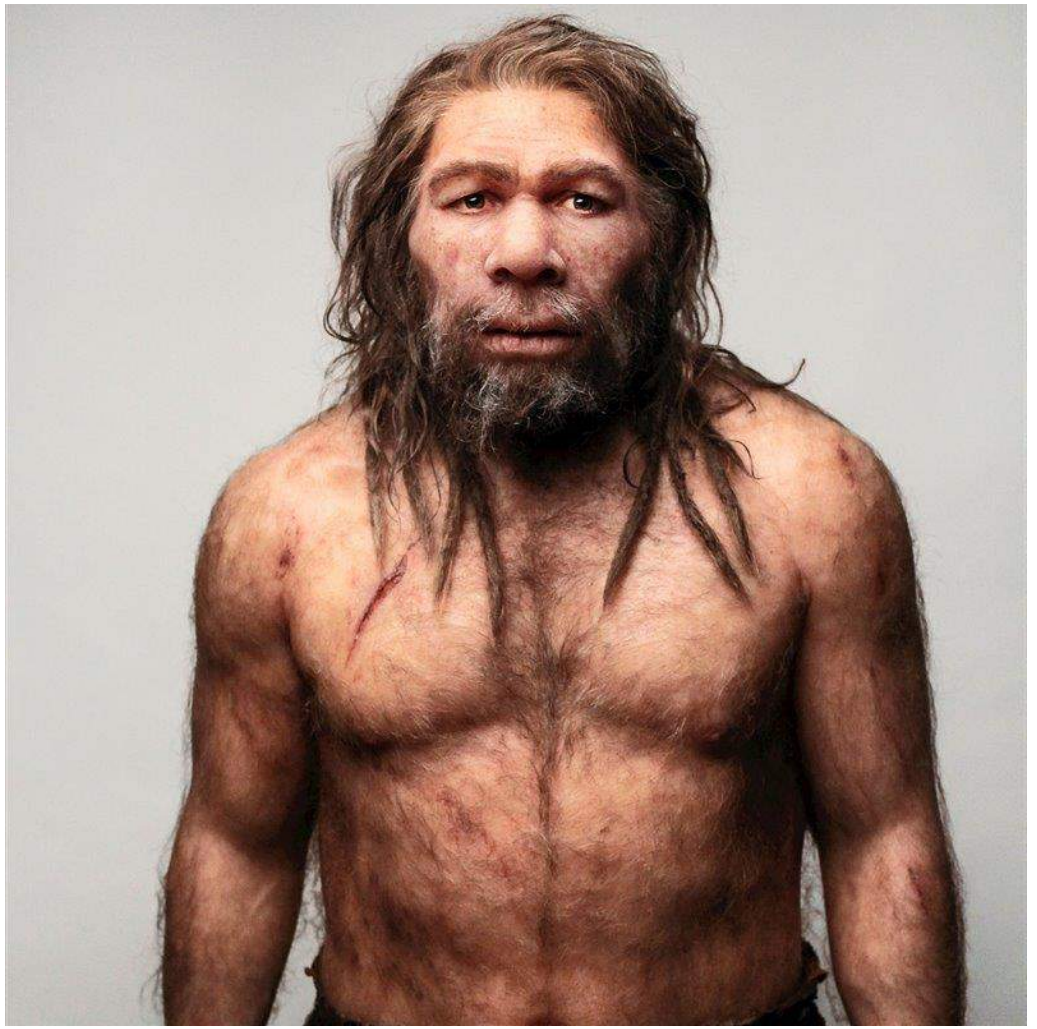
LA NOCHE ESTRELLADA, VINCENT VAN GOGH

DeepArt.io

Jason Allen's A.I.-generated work, **"Théâtre D'opéra Spatial,"** took first place in the digital category at the Colorado State Fair.

Midjourney

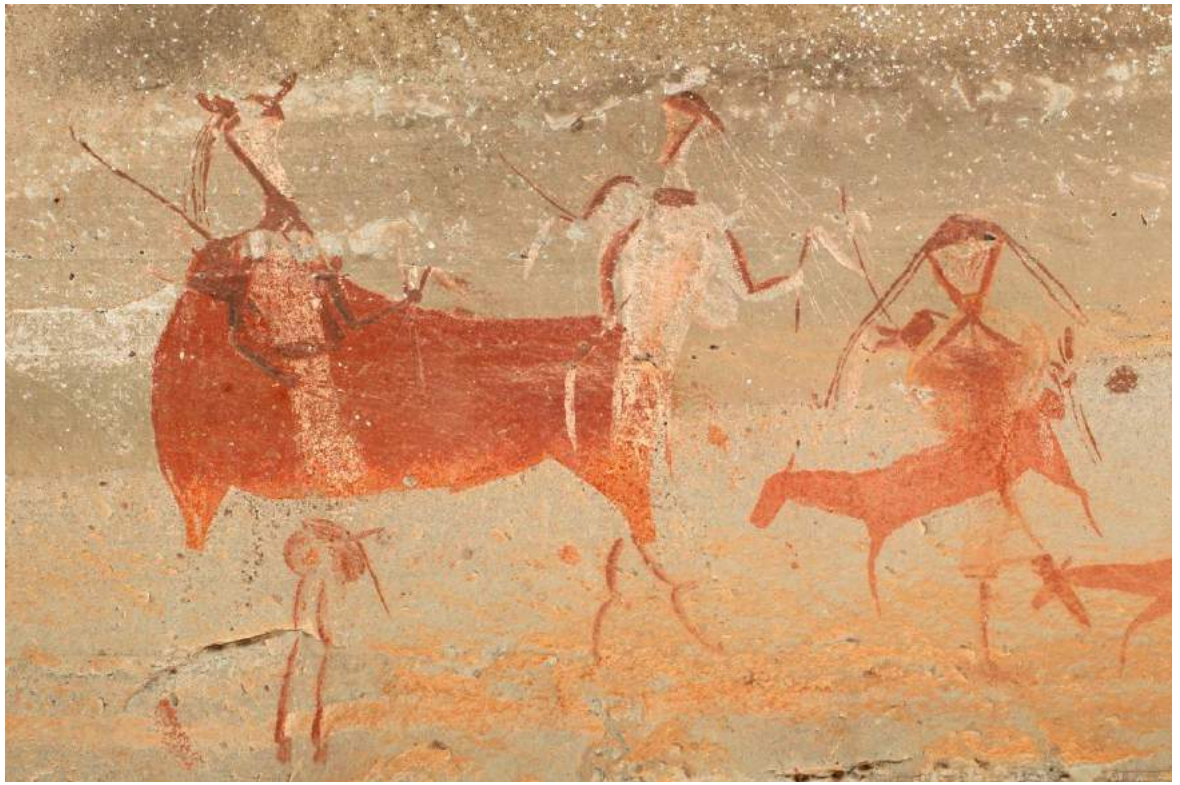
Homo sapiens



El **arte** es una de las manifestaciones de la creatividad humana, es la herramienta con la cual *Homo sapiens* desarrolla su cultura, unión y fuerza como pueblo.

El l**enguaje** permite la comunicación
El lenguaje humano pudo aparecer hace más de 50.000 años

Homo sapiens



20 000 años de antigüedad

https://arthearty.com/significance-of-lascaux-cave-paintings

ARTE DIGITAL

“Teatro espacial de la ópera”

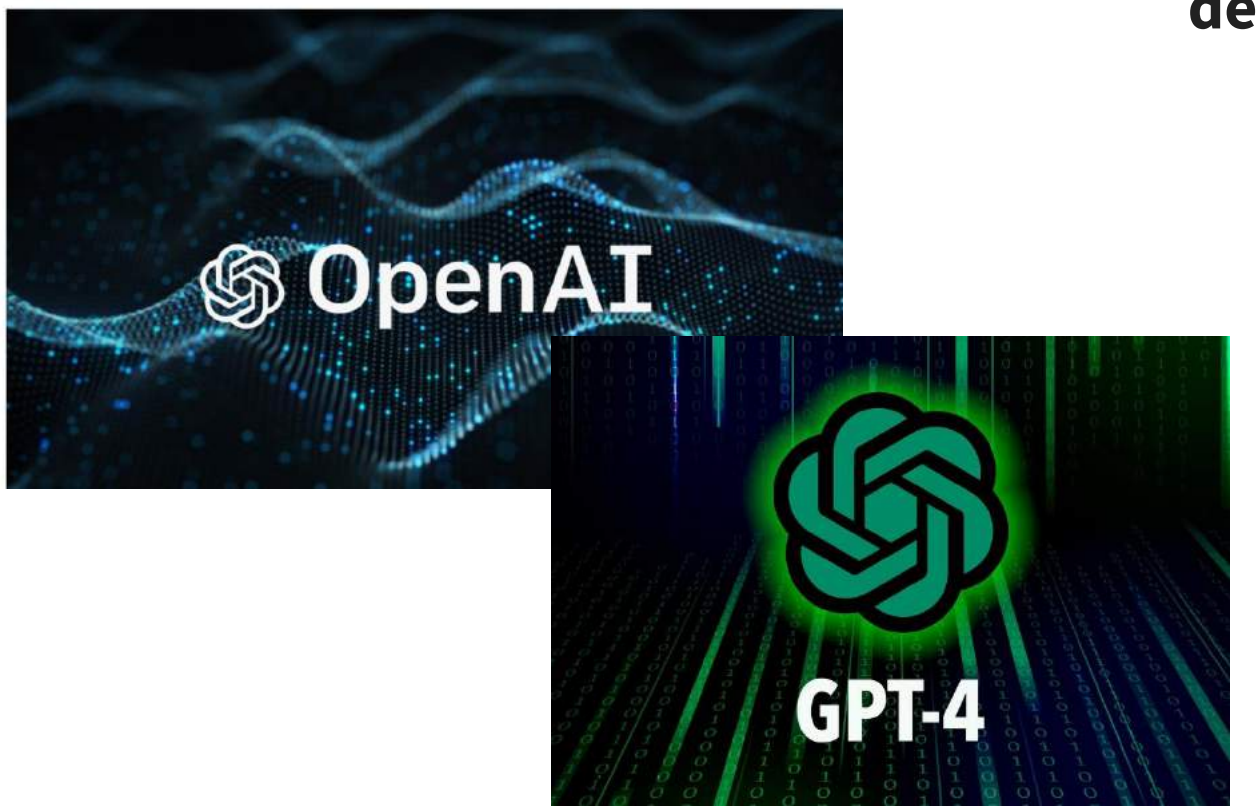
ARTE DIGITAL

“Teatro espacial de la ópera”

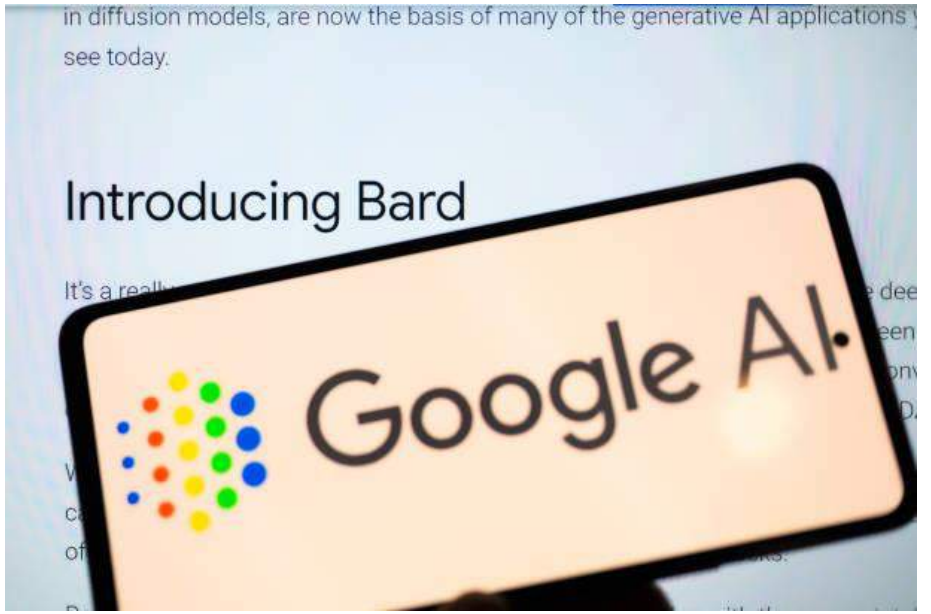
Midjourney
Creación a partir de instrucciones (Prompt)
Ganadora Premio Colorado, Agosto 2022

IA Generativa: Lenguaje

GPT-3, el nuevo modelo de lenguaje de OpenAI, es capaz de programar, diseñar y hasta conversar sobre política o economía



Google presenta Bard: el nuevo modelo de lenguaje AI con amplias aplicaciones



https://www.xataka.com/robotica-e-ia/gpt-3-nuevo-modelo-lenguaje-openai-capaz-programar-disenar-conversar-politica-economia
https://www.xataka.com/basics/google-bard-que-como-funciona-que-puedes-hacer-inteligencia-artificial-buscador

IA Generativa (GPT-4 y ChatGPT)

Cornell University

the Sim

arXiv > cs > arXiv:2303.12712

Search...

Help | Advanced

Computer Science > Computation and Language

[Submitted on 22 Mar 2023 (v1), last revised 24 Mar 2023 (this version, v2)]

Sparks of Artificial General Intelligence: Early experiments with GPT-4

Sébastien Bubeck, Varun Chandrasekaran, Ronen Eldan, Johannes Gehrke, Eric Horvitz, Ece Kamar, Peter Lee, Yin Tat Lee, Yuanzhi Li, Scott Lundberg, Harsha Nori, Hamid Palangi, Marco Tulio Ribeiro, Yi Zhang

Microsoft lo tiene claro y en un último informe habla de que su IA, GPT-4, es una forma temprana de inteligencia artificial general (AGI).

Computer Hoy

LO ÚLTIMO ANÁLISIS GUÍAS DE COMPRA LOS MEJORES TUTORIALES PRODUCTOS

Investigadores de Microsoft afirman que la última versión de ChatGPT muestra indicios de inteligencia humana

TECNOLOGÍA Carolina González Valenzuela | 27 mar. 2023 12:45h.

https://computerhoy.com/investigadores-microsoft-afirman-ultima-version-chatgpt-muestra-indicios-inteligencia-humana-1221122

IA Generativa:
Sobre las alucinaciones

Uno de los riesgos de la IA generativa se llama alucinación.

Una de las mayores preocupaciones de los sistemas de IA Generativa es cuando no entienden las preguntas, las malinterpretan y como no pueden generar respuestas correctas, empiezan a inventarlas en un proceso llamado “Alucinación de Inteligencia Artificial”.
Alucinación es el término empleado para el fenómeno en el que los algoritmos de IA y las redes neuronales de aprendizaje profundo producen resultados que no son reales, que no coinciden con ningún dato en el que se haya entrenado el algoritmo u otro patrón identificable.
No puede explicarse por su programación, información de entrada, otros factores como la clasificación incorrecta de datos, capacitación inadecuada, incapacidad para interpretar preguntas en diferentes idiomas, incapacidad para contextualizar preguntas.
Las alucinaciones pueden ocurrir en todo tipo de datos sintéticos, como texto, imágenes, audio, video y código informático.

ESTADOS UNIDOS

Un abogado admite que usó ChatGPT para un escrito y éste se inventó precedentes legales

El letrado se enfrenta a posibles sanciones y deberá explicar a la Corte por qué no debería ser castigado

https://www.elmundo.es/internacional/2023/05/27/64727a83fdddff68bf8b459a.html

Inteligencia artificial generativa: 2022 eclosión

En 2022, la inteligencia artificial (IA) se volvió creativa.

Cada vez hay más tareas en las que cuesta distinguir si algo ha salido de un cerebro humano o de un sistema de IA (ajedrez, arte, descubrimiento proteínas, conversaciones).

Ventajas/beneficios que nos traerá, riesgo de que se usen mal.
Ambos enfoques tienen parte de razón

Inteligencia artificial

Inteligencia artificial generativa

Eclosión de la IA, con riesgos

Etica y regulación. Auditabilidad

IA fiable y segura

Concluyendo

Eclosión económica asociada a la inteligencia artificial

COMPUTERWORLD

La IA y los datos generarán 16.500 millones de euros para 2025 en la industria española

Instituciones, expertos académicos y empresarios afirman que el uso de la tecnología basada en datos e inteligencia artificial (IA) es la gran oportunidad para el crecimiento del país, ya que generará ingresos en la industria de 16.000 millones de euros para 2025.

CincoDías

El objetivo del Gobierno es llegar al 40% a finales de 2025

DIGITALIZACIÓN

La economía digital crece 3 puntos en España y ya representa el 22% de PIB

COMBUSTIBLE PARA EL CRECIMIENTO

accenture

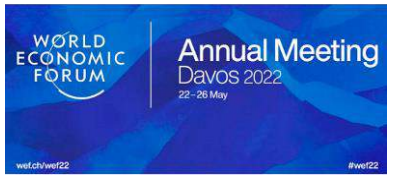
El estudio de Accenture sobre el impacto de la Inteligencia Artificial en 12 economías desarrolladas revela que la IA podría llegar a duplicar las tasas anuales de crecimiento económico en 2035, cambiando la naturaleza del trabajo y estableciendo una nueva relación entre el hombre y la máquina. Se prevé que el impacto de la IA en los negocios aumentará la productividad del trabajo hasta en un 40% y permitirá a las personas hacer un uso más eficiente de su tiempo.

La Inteligencia Artificial hará aumentar la economía mundial en 15,7 billones de dólares en 2030

Archivado en: Economía · DES2021 · Inteligencia Artificial

Expertos en Inteligencia Artificial analizarán en DES2021 los retos y usos prácticos de esta tecnología para aumentar la competitividad de las empresas.

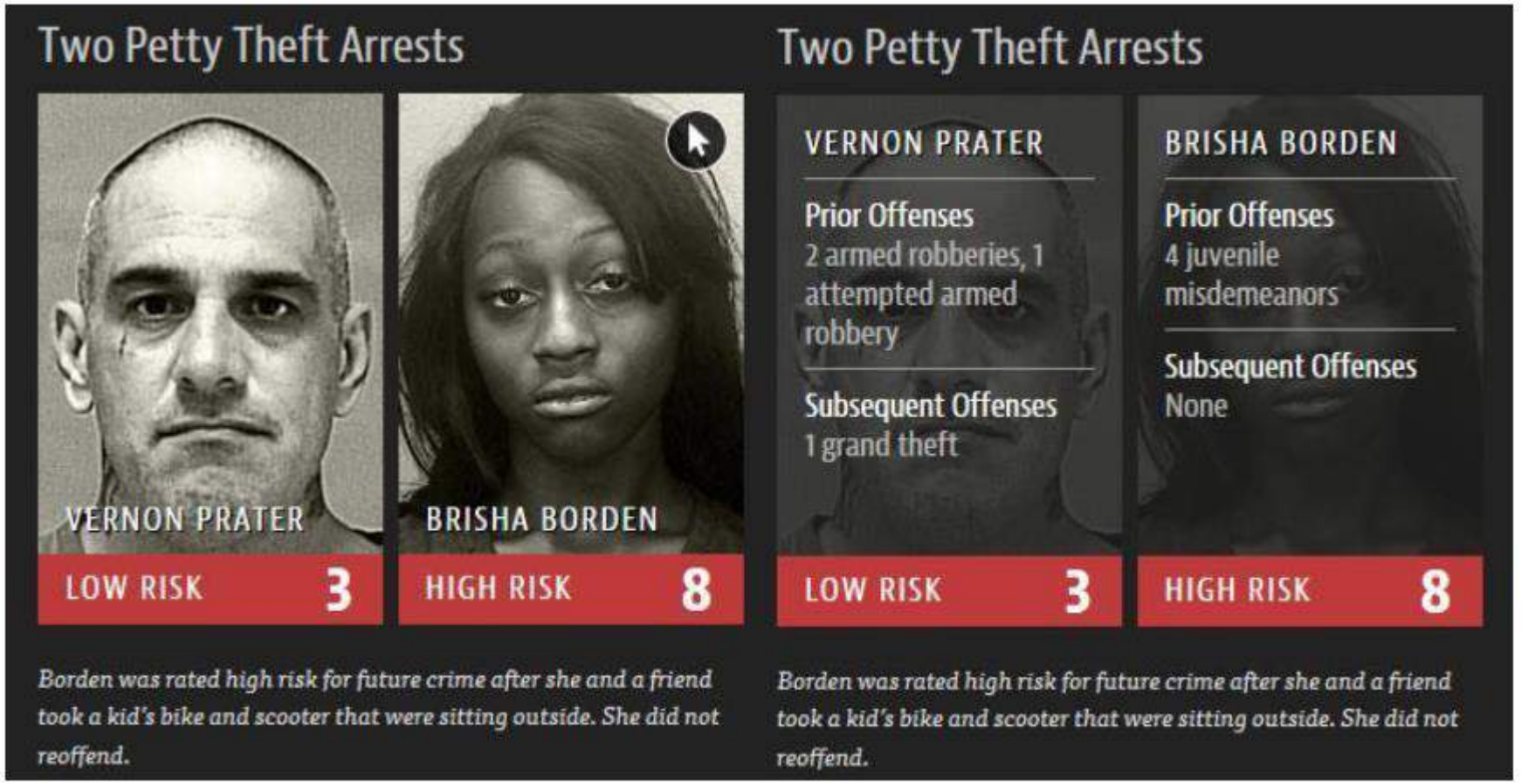
Colaboración entre humanos, máquinas y algoritmos



Según un informe del **Foro Económico Mundial,** se espera **que 85 millones de puestos de trabajo sean desplazados por un cambio en la división del trabajo entre humanos y máquinas para 2025. La buena noticia es que pueden surgir 97 millones de nuevos cargos más adaptados a la nueva división del trabajo** entre humanos, máquinas y algoritmos, según el mismo informe.

Eclosión de la inteligencia artificial, con riesgos

Compas. Sistema inteligente para calcular el riesgo de reincidencia en el delito



https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing
http://theconversation.com/why-using-ai-to-sentence-criminals-is-a-dangerous-idea-77734

Eclosión de la inteligencia artificial, con riesgos

noticiasya MERCADO Suscríbete

Home / Noticias Locales

Amazon apaga sistema de reclutamiento porque discriminaba a mujeres

AMAZON

Por Saddam Aguayo

Tu opinión cuenta, escríbenos a **socialmedia@entravision.com** →

octubre 15, 2018, 7:23 pm PST

Amazon.com

elDiario.es

Hola, Francisco422

Política Internacional Economía Opinión Cultura Educación Clima Desalambre Igualdad Estatuto

Sesgos de género en los algoritmos: un círculo perverso de discriminación en línea y en la vida real

La "condición algorítmica" está alterando los derechos humanos, que prohíben la discriminación por razón de sexo o género. La industria tecnológica no está haciendo lo suficiente para abordar estos sesgos. El

A nurse in front of a hospital

A doctor in front of a hospital

Eclosión de la inteligencia artificial, con riesgos

'He Would Still Be Here': Man Dies by Suicide After Talking with AI Chatbot, Widow Says

The incident raises concerns about guardrails around quickly-proliferating conversational AI models.

By Chloe Xiang

March 30, 2023, 9:59pm Share Tweet Snap

Listen to this article

A Belgian man recently died by suicide after chatting with an AI chatbot on an app called Chai, Belgian outlet *La Libre* reported.

https://www.vice.com/en/article/pkadgm/man-dies-by-suicide-after-talking-with-ai-chatbot-widow-says

Eclosión de la inteligencia artificial, con riesgos

Toju Duke, exdirectora de IA de Google: “La inteligencia artificial amplifica injusticias sistémicas que ya deberíamos haber eliminado”

La exdirectora de IA responsable en Google cree que el debate sobre esta tecnología se ha centrado en si la humanidad correrá peligro mañana, cuando el problema es que hoy ya discrimina a la población

20 de octubre de 2023 • 16:25

Manuel G. Pascual

EL PAÍS

https://www.lanacion.com.ar/tecnologia/toju-duke-exdirectora-de-ia-de-google-la-inteligencia-artificial-amplifica-injusticias-sistemicas-nid20102023/

Inteligencia artificial. Necesitamos regular

Written Testimony
of
Sam Altman
Chief Executive Officer
OpenAI
Before the U.S. Senate Committee on the Judiciary
Subcommittee on Privacy, Technology, & the Law

"Es esencial regular la inteligencia artificial, y que esas normas garanticen que el público acceda a los muchos beneficios de esta tecnología",
"Mi peor miedo es que esta tecnología salga mal. Y si sale mal, puede salir muy mal".

https://www.washingtonpost.com/documents/0668f6f4-d957-4b94-a745-2aa9617d1d60.pdf?itid=lk_inline_manual_18

EL PAÍS Tecnología

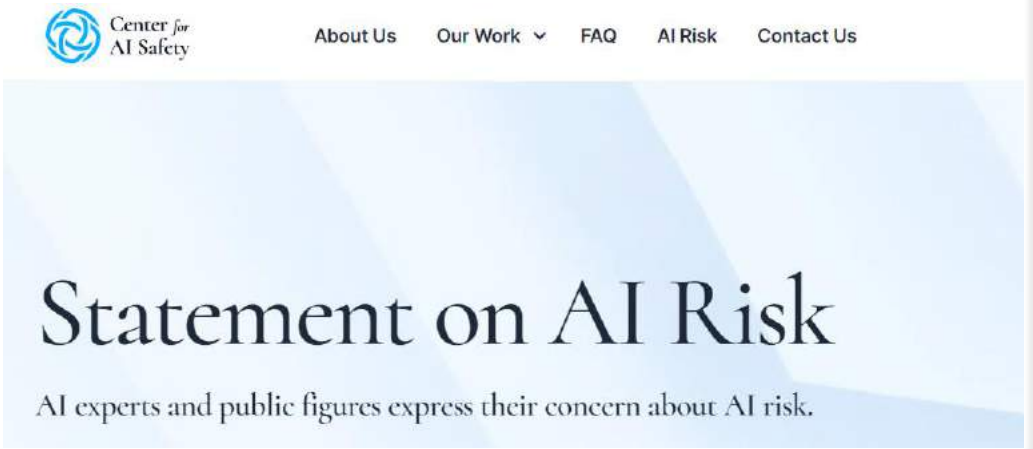
Sam Altman (ChatGPT), en el Capitolio: "Si la inteligencia artificial sale mal, puede salir muy mal"

El cofundador de OpenAI comparece ante el Senado estadounidense para defender los beneficios y alertar de los riesgos de la tecnología revolucionaria

Washington - 16 MAY 2023 - 19:10 CEST

https://elpais.com/tecnologia/2023-05-16/sam-altman-chatgpt-en-el-capitolio-si-la-inteligencia-artificial-sale-mal-puede-salir-muy-mal.html

Inteligencia artificial. Necesitamos regular



Mitigating the risk of extinction from AI should be a global priority alongside other societal-scale risks such as pandemics and nuclear war.

Signatories:

☑ AI Scientists ☑ Other Notable Figures

Geoffrey Hinton
Emeritus Professor of Computer Science, University of Toronto

Yoshua Bengio
Professor of Computer Science, U. Montreal / Mila

Demis Hassabis
CEO, Google DeepMind

Sam Altman
CEO, OpenAI

Dario Amodei
CEO, Anthropic

https://www.safe.ai/statement-on-ai-risk

Inteligencia artificial

Inteligencia artificial generativa

Eclosión de la IA, con riesgos

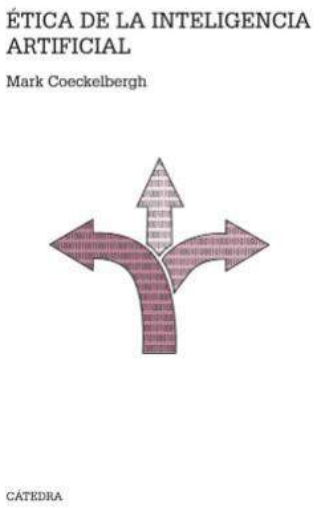
Etica y regulación. Auditabilidad

IA fiable y segura

Concluyendo

IA: Etica y regulación. Debate ético y Principios éticos de la IA

Aunque los problemas de sesgo y discriminación siempre han estado presentes en la sociedad, la preocupación es que la IA pueda perpetuar estos problemas y acrecentar su impacto



Aproximación filosófica:

- El papel de la IA y los humanos
- Análisis la omnipresencia de la IA
- Análisis ético y social de su aplicación
- Políticas de IA de cara al futuro
- Vigilar el discurso público con respecto a la IA

Principios éticos de Naciones Unidas (Nov. 2021)

1. Proporcionalidad y no amenaza
2. Seguridad
3. Equidad y no discriminación
4. Sostenibilidad
5. Derecho a la privacidad, y protección de datos
6. Control Humano
7. Transparencia y explicabilidad
8. Responsabilidad y rendición de cuentas
9. Conciencia y educación sobre la IA
10. Gobernanza y colaboración adaptativa y de múltiples partes interesadas

Inteligencia artificial. Cumbre para la regulación y seguridad

• Últimas noticias

El Confidencial

Iniciar sesión

LA IA YA ESTÁ CAMBIANDO LAS GUERRAS

Cumbre en UK contra el riesgo "catastrófico" de la IA. ¿Puede acabar con la humanidad?

Cientos de líderes mundiales, académicos, investigadores y empresarios, incluido el hombre más rico del mundo, Elon Musk, se dan cita esta semana en Reino Unido para abordar los retos que plantea la inteligencia artificial

Policy paper

The Bletchley Declaration by Countries Attending the AI Safety Summit, 1-2 November 2023

Published 1 November 2023

AI SAFETY SUMMIT

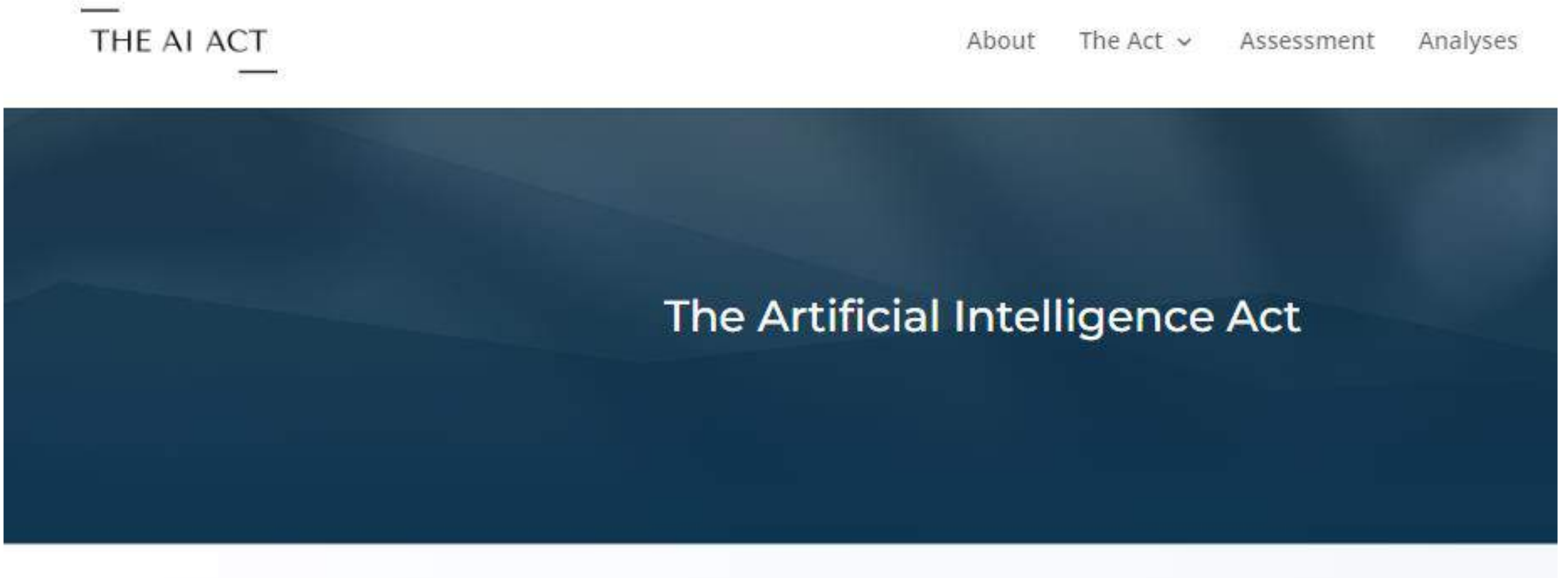
En el contexto de nuestra cooperación, y para informar la acción a nivel nacional e internacional, nuestra agenda para abordar el riesgo fronterizo de la IA se centrará en:

- **identificar los riesgos de seguridad de la IA de interés compartido, ...**
- **construir políticas respectivas basadas en riesgos en todos nuestros países para garantizar la seguridad a la luz de dichos riesgos,...**

https://www.elconfidencial.com/tecnologia/2023-11-01/uk-congreso-seguridad-ia-inteligencia-artificial-reino-unido_3765886/

IA: Etica y regulación

Marco regulatorio europeo



https://artificialintelligenceact.eu/

Alto riesgo

Regulación europea: AI Act

Inaceptables

~~Social Scoring (Black Mirror)~~

Alto riesgo

- identificación biométrica y categorización de personas físicas,
- gestión y funcionamiento de infraestructuras esenciales,
- educación y formación profesional,
- empleo, gestión de los trabajadores y acceso al autoempleo,
- acceso y disfrute de servicios públicos y privados esenciales y sus beneficios,
- asuntos relacionados con la aplicación de la ley,
- gestión de la migración, el asilo y el control fronterizo, o
- administración de justicia y procesos democráticos.

IA: Etica y regulación. Auditabilidad

Establecer e implementar un **sistema de gestión de riesgos**

a la luz del **propósito previsto** del sistema de IA.

- Utilizar **datos** de entrenamiento, validación y pruebas de **alta calidad** (relevantes, representativos, etc.)
- Elaborar **documentación técnica y configurar capacidades de registro** (trazabilidad y auditabilidad)
- Garantizar un grado adecuado de **transparencia** y proporcionar a los usuarios información sobre las capacidades y limitaciones del sistema y cómo utilizarlo
- Garantizar la **supervisión humana** (medidas integradas en el sistema y/o a implementar por los usuarios)
- Garantizar **robustez, precisión y ciberseguridad**

IA: Regulación y derechos e implicaciones sociales (coches autónomos)

¿Pautas legales para juzgar a un sistema inteligente?

COCHE AUTÓNOMO TESLA ›

Tesla reconoce el segundo accidente mortal en EE UU con un coche que circulaba en piloto automático

El siniestro tuvo lugar hace una semana en California cuando un Model X chocó contra un muro después de que el sistema avisase al conductor para que pusiera las manos en el volante

EL PAIS

Madrid - 31 MAR 2018 - 13:33 CEST

Los servicios de rescate trabajan en el lugar donde se estrelló el Model X, en California. STRINGER (REUTERS)

29 Marzo 2018 – Segundo accidente mortal de un coche Tesla con piloto automático

https://elpais.com/tecnologia/2018/03/31/actualidad/1522488697_843295.html

IA: Regulación y derechos e implicaciones sociales (Impacto social en el trabajo)

EL PAIS

NEGOCIOS

EL FUTURO DEL TRABAJO > OPINIÓN

¿Debemos preocuparnos porque los robots se queden con nuestro trabajo?

No cabe duda de que muchos empleos vinculados a la economía del conocimiento pueden ser sustituibles

Robots y personas trabajan en un almacén industrial.
LOCUS ROBOTICS/PR NEWSWIRE (EUROPA PRESS)

Paul Krugman es premio Nobel de Economía.
© The New York Times, 2022.

¿qué pasaría si las máquinas pudieran hacerse cargo de una gran parte de lo que tradicionalmente hemos considerado trabajo del conocimiento?

ChatGPT, tecnología que parece capaz de realizar tareas que requieren una considerable formación académica.

¿cuánto de lo que hacemos los seres humanos es creativo de verdad o aporta una comprensión en profundidad?

https://elpais.com/economia/negocios/2022-12-10/debemos-preocuparnos-porque-los-robots-se-queden-con-nuestro-trabajo.html?mid=DM158851&bid=1399054736

Inteligencia artificial

Inteligencia artificial generativa

Eclosión de la IA, con riesgos

Etica y regulación. Auditabilidad

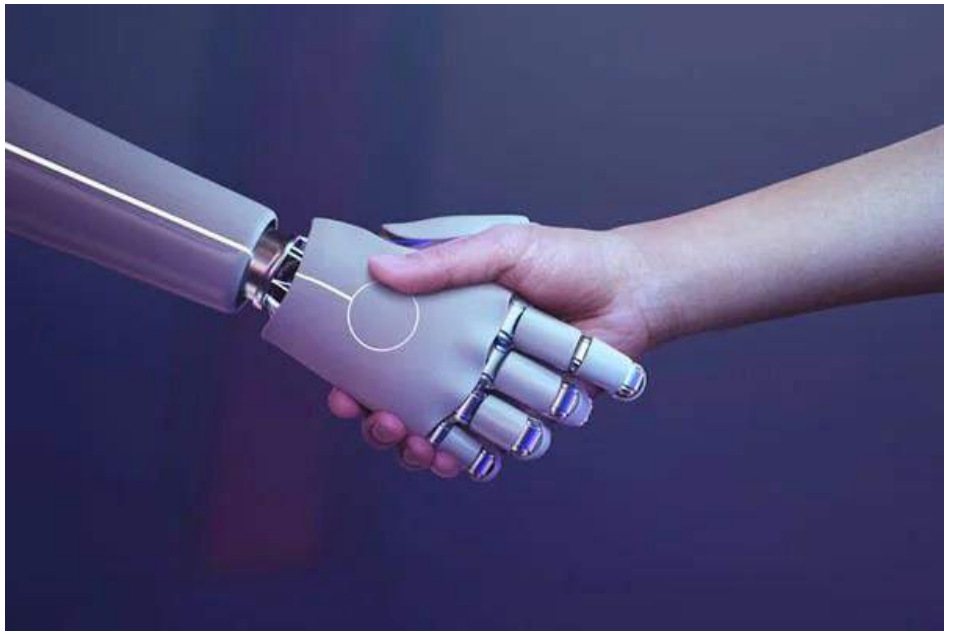
IA fiable y segura. Tecnología de IA

Concluyendo

IA: Regulación, Inteligencia artificial fiable

Inteligencia artificial fiable
(Trustworthy AI)

1. Lícita (cumple la ley)
2. Ética (principios éticos)
3. Robusta (segura en su uso, control de fallos, …)



Inteligencia artificial fiable

Tres pilares básicos de una IA fiable:

- ***Conforme a la ley.*** La IA tiene que ser lícita
- ***Ética.*** El respeto de la ley no es siempre garante del respeto de principios éticos que salvaguarden los derechos inalienables de las personas.

(Ético es: Proteger la igualdad, la no discriminación y la solidaridad entre los ciudadanos. Respetar la democracia, la justicia y el estado de Derecho. Respetar la dignidad humana. Respetar la libertad individual. Respetar los derechos de los ciudadanos)

- ***Robusta****.* los sistemas de IA deben ofrecer un correcto funcionamiento siempre y prever medidas de protección para evitar cualquier efecto adverso imprevisto.

Inteligencia artificial fiable. Etica, legal y robusta

IA cumpliendo siete requisitos esenciales

I. *Intervención y supervisión humanas.*

II. *Solidez y seguridad técnicas*

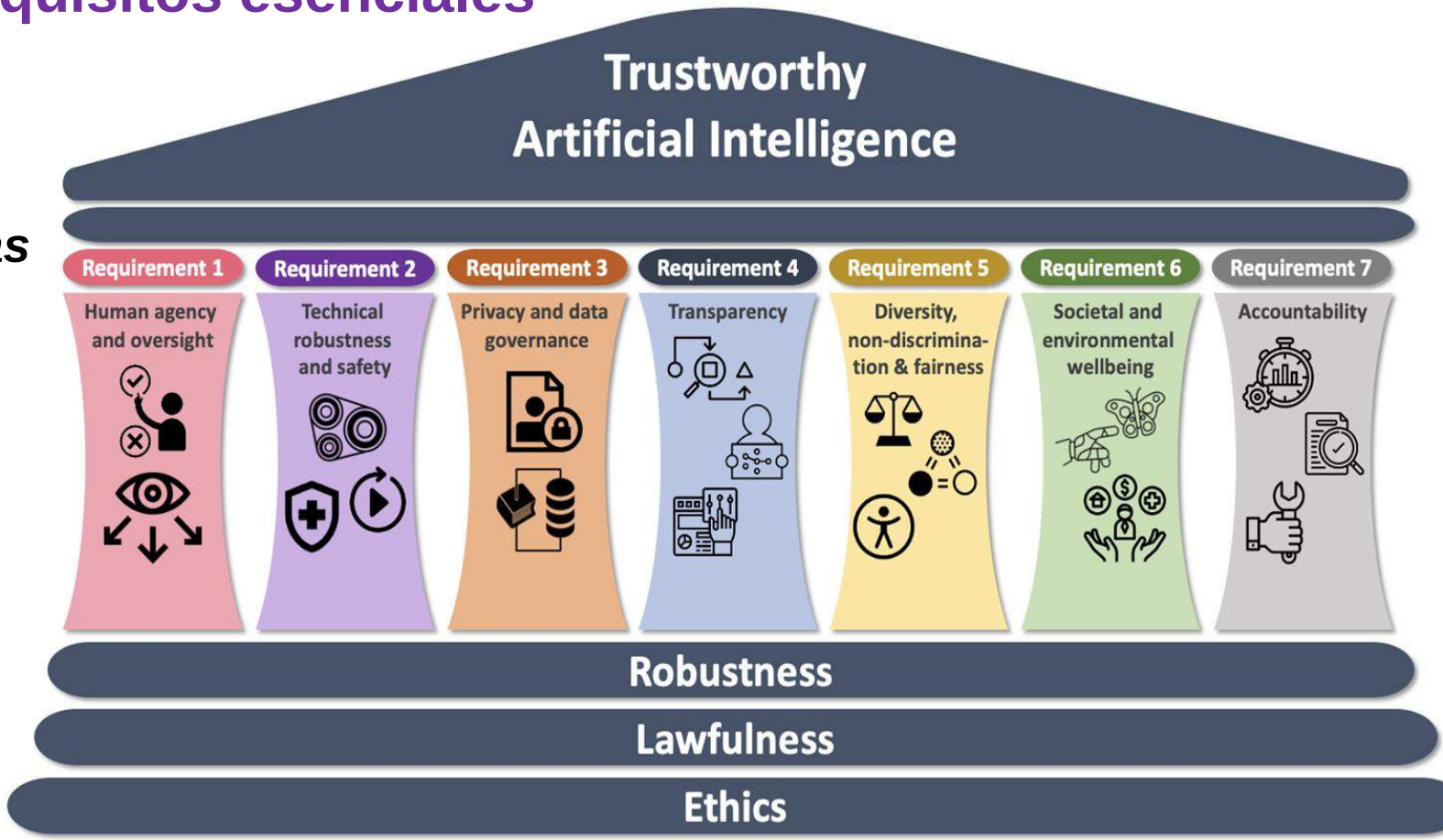
III. *Privacidad y gestión de datos.*

IV. *Transparencia.*

V. *Diversidad, no discriminación y equidad.*

VI. *Bienestar social y medioambiental.*

VII. *Rendición de cuentas.*



Díaz-Rodríguez et al. *Connecting the Dots in Trustworthy Artificial Intelligence: From AI Principles, Ethics, and Key Requirements to Responsible AI Systems and Regulation*
arXiv preprint arXiv:2305.02231 2023

Inteligencia artificial fiable

Siete requisitos esenciales que debe cumplir un sistema inteligente

I. ***Intervención y supervisión humanas.*** los sistemas de IA tienen que permitir ser gobernados o supervisados por humanos.

II. ***Solidez y seguridad técnicas.*** Los sistemas de IA tienen que garantizar robustez tecnológica e incluso considerar planes de contingencia para la adaptación ante comportamientos anómalos.

III. ***Privacidad y gestión de datos.*** Los datos tienen que estar protegidos

IV. ***Transparencia*****.** El comportamiento de los sistemas de IA debe poder ser monitorizado o trazado

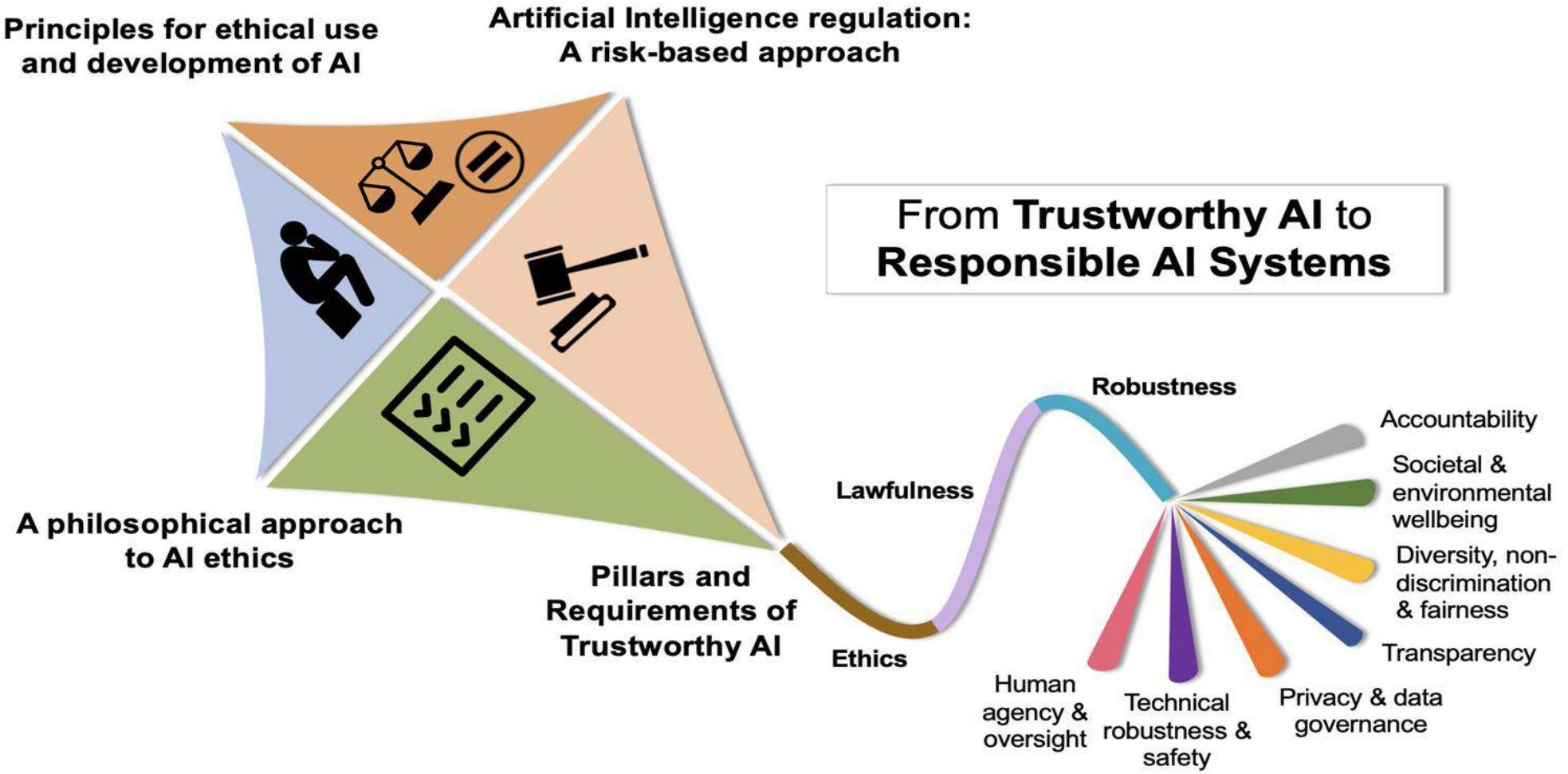
V. ***Diversidad, no discriminación y equidad.*** El proceso de adquisición y anotación de los datos tiene que preservar la igualdad y evitar la discriminación de los ciudadanos.

VI. ***Bienestar social y medioambiental.***

VII. ***Rendición de cuentas.*** Esta directriz está relacionada con el principio de responsabilidad.

Inteligencia artificial fiable

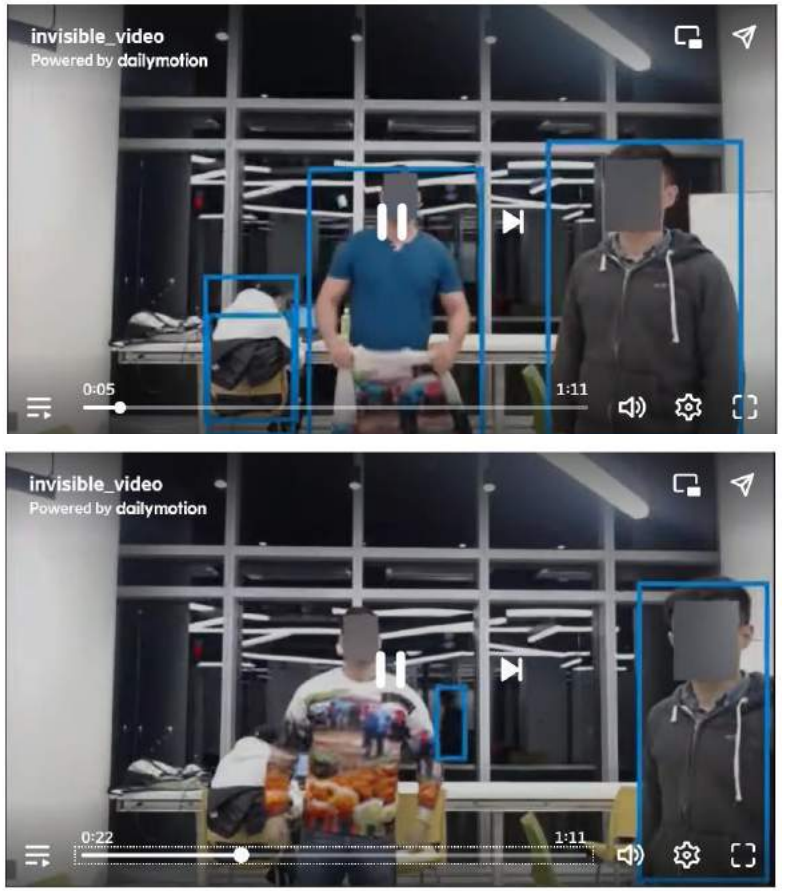
Visión holística



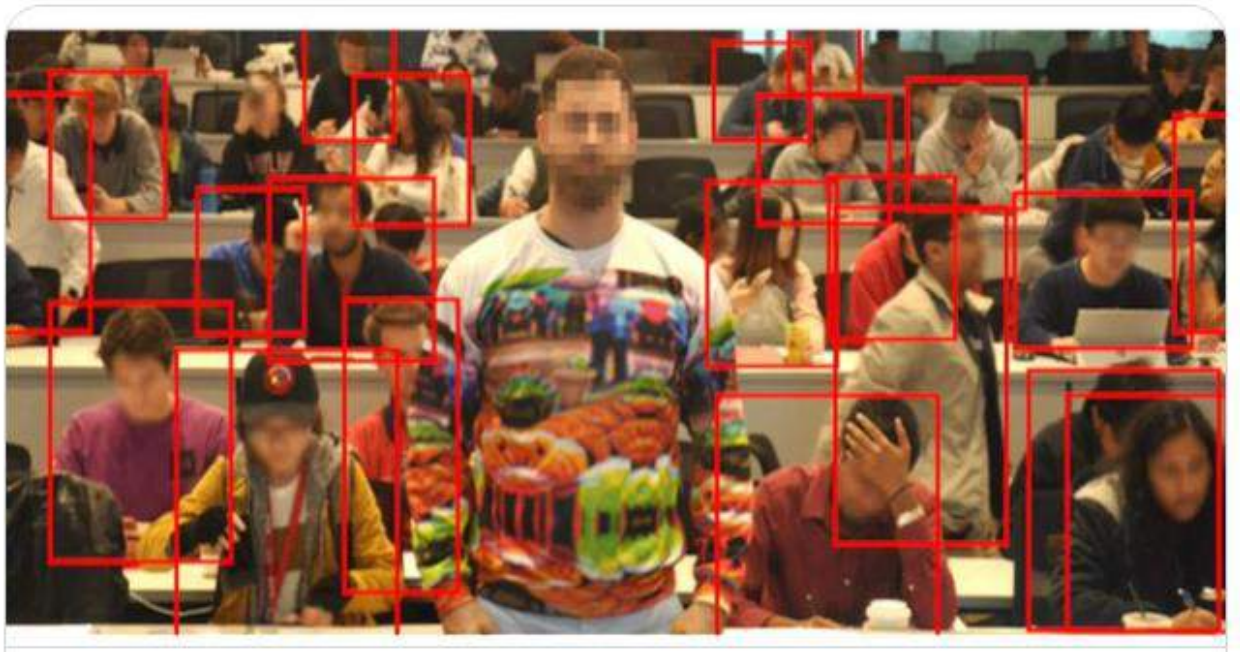
Díaz-Rodríguez et al. *Connecting the Dots in Trustworthy Artificial Intelligence: From AI Principles, Ethics, and Key Requirements to Responsible AI Systems and Regulation*
arXiv preprint arXiv:2305.02231 2023

IA segura : Reto. Robustez, evitar adversarios

Este jersey funciona como capa de invisibilidad frente la Inteligencia Artificial



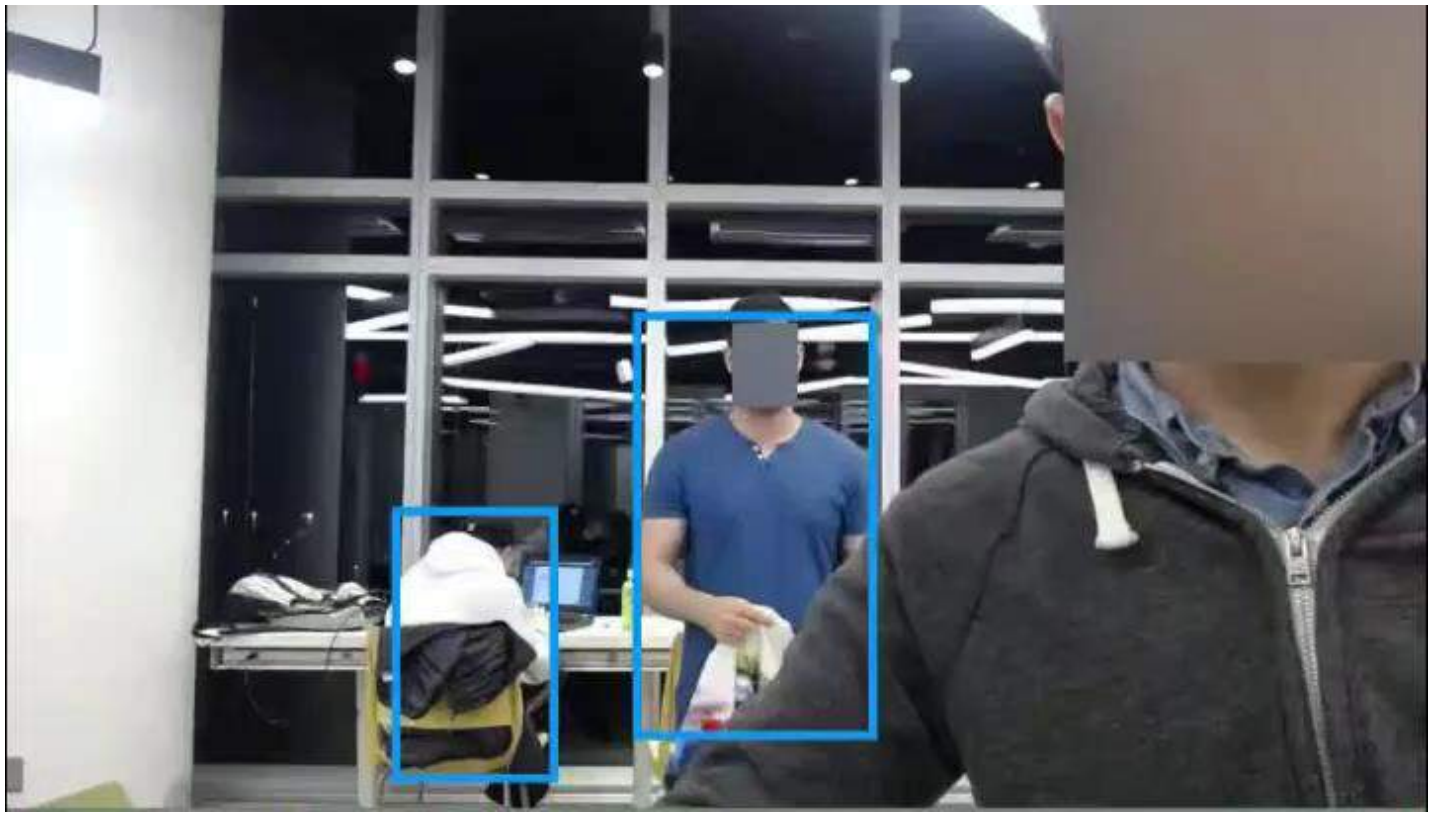
un jersey que utiliza "patrones adversos" que eliminan los patrones comunes a los que están entrenados estos sistemas y lo convierte en invisible ante la IA.



https://www.eleconomista.es/videos/noticias/12010736/10/22/Este-jersey-funciona-como-capa-de-invisibilidad-frente-la-Inteligencia-Artificial-.html

IA segura : Reto. Robustez, evitar adversarios

Este jersey funciona como capa de invisibilidad frente la Inteligencia Artificial



IA Segura:
Abarca la ética de las máquinas, la alineación de la IA con principios éticos y metas humanas, cuyo objetivo es hacer que los sistemas de IA sean morales y beneficiosos. Tecnológicamente abarca problemas para monitorear los sistemas para detectar riesgos y hacerlos altamente fiables. Más allá de la investigación en IA, implica desarrollar normas y políticas que promuevan la seguridad.

https://www.eleconomista.es/videos/noticias/12010736/10/22/Este-jersey-funciona-como-capa-de-invisibilidad-frente-la-Inteligencia-Artificial-.html

IA segura

“Necesitamos controles en dos niveles: uno, para controlar lo que los sistemas pueden hacer por sí mismos, y otro, para controlar lo que pueden hacer bajo instrucciones humanas.”

Inteligencia artificial

Inteligencia artificial generativa

Eclosión de la IA, con riesgos

Etica y regulación. Auditabilidad

IA fiable y segura

Concluyendo

Concluyendo, R1 (Una IA silenciosa que ha entrado en nuestras vidas)

«Máquinas no pensantes cada vez más capaces.
Cualquier tarea que requiera menos de diez segundos de pensamiento podrá ser hecha por una IA»

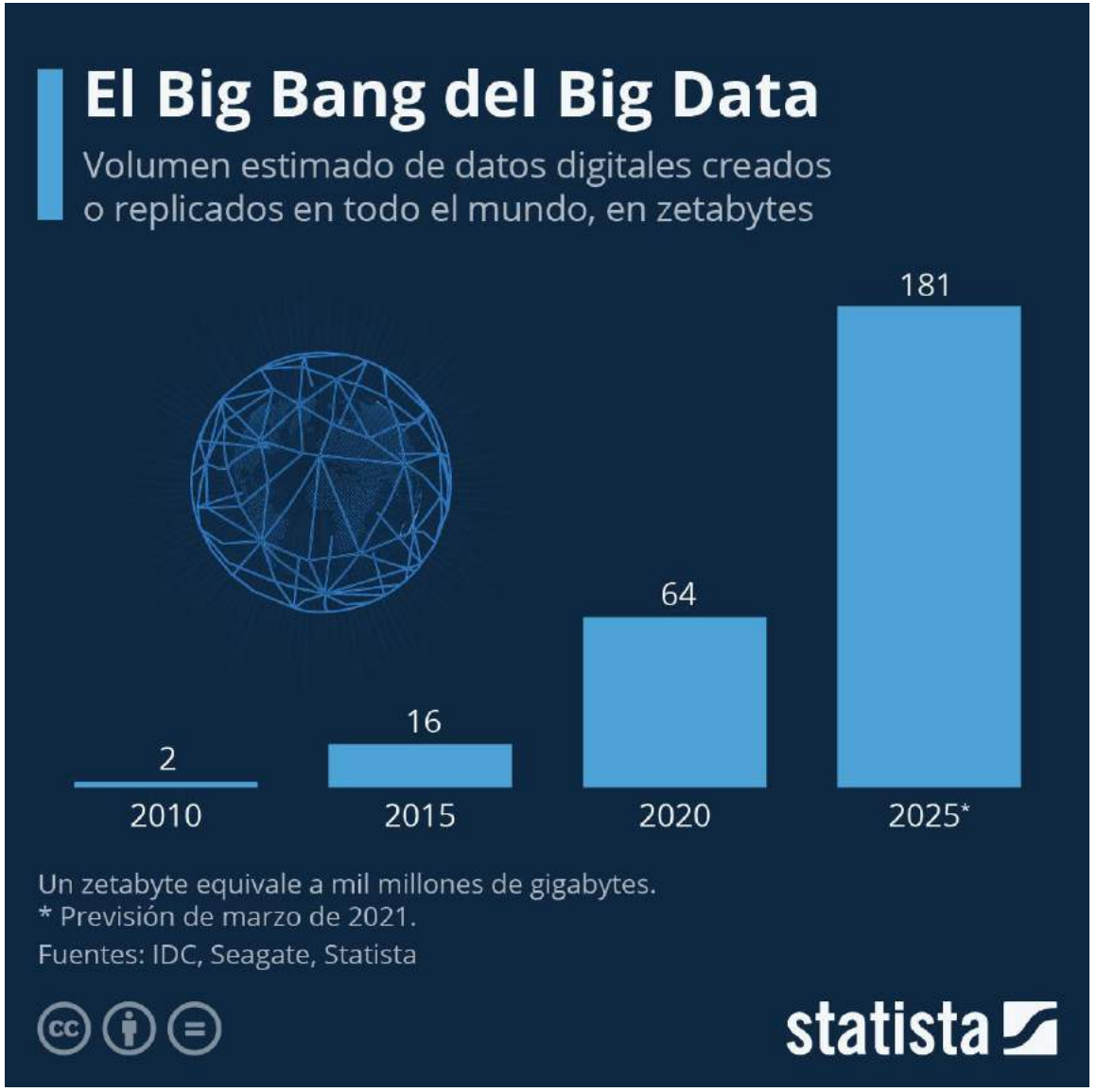
Concluyendo, R2: Los datos son la base de la inteligencia artificial y el aprendizaje automático

"Inteligencia Artificial es la nueva electricidad" A. Ng

(Big) data es "el nuevo petróleo/la nueva materia prima" del siglo XXI

Concluyendo, R2: Los datos son la base de la inteligencia artificial y el aprendizaje automático (el mundo gira alrededor de los datos)

La explosión de los datos



Concluyendo, R3:
reflexiones sobre retos/problemas a abordar

Tres grandes retos

“UN GRAN ERROR SOBRE LA IA ES PENSAR QUE ES INTELIGENTE”

SANDRA WACHTER

Professor **Sandra Wachter** is Professor of Technology and Regulation at the **Oxford** Internet Institute at the University of **Oxford**

“Siempre hay una cuestión con la **protección de datos**, porque se usan datos para alimentar el algoritmo, no es posible hacerlo sin datos.

Siempre hay una cuestión con la **explicabilidad**, en el sentido de que no entiendes completamente cómo funciona el algoritmo y por qué toma ciertas decisiones.

Y, muy a menudo, hay también un **problema de sesgo** porque, de nuevo, los **datos son históricos**, y la mayoría de conjuntos de datos no son todos los datos, son datos sesgados.”

Concluyendo: Debate. Sobre creatividad

El sueño de la máquina creativa

EL FOCO

La cuestión no es si el arte de los ordenadores lo hará mejor que nosotros, sino pensar qué podemos hacer únicamente nosotros cuando las computadoras alcanzan tal sofisticación

DANIEL INNERARITY

Catedrático de Filosofía Política, investigador Ikerbasque en la UPV/EHU y titular de la cátedra Inteligencia Artificial y Democracia en el Instituto Europeo de Florencia

Domingo, 5 febrero 2023, 00:49

La creatividad humana no puede ni imitarse ni repetirse: implica siempre una cierta trasgresión

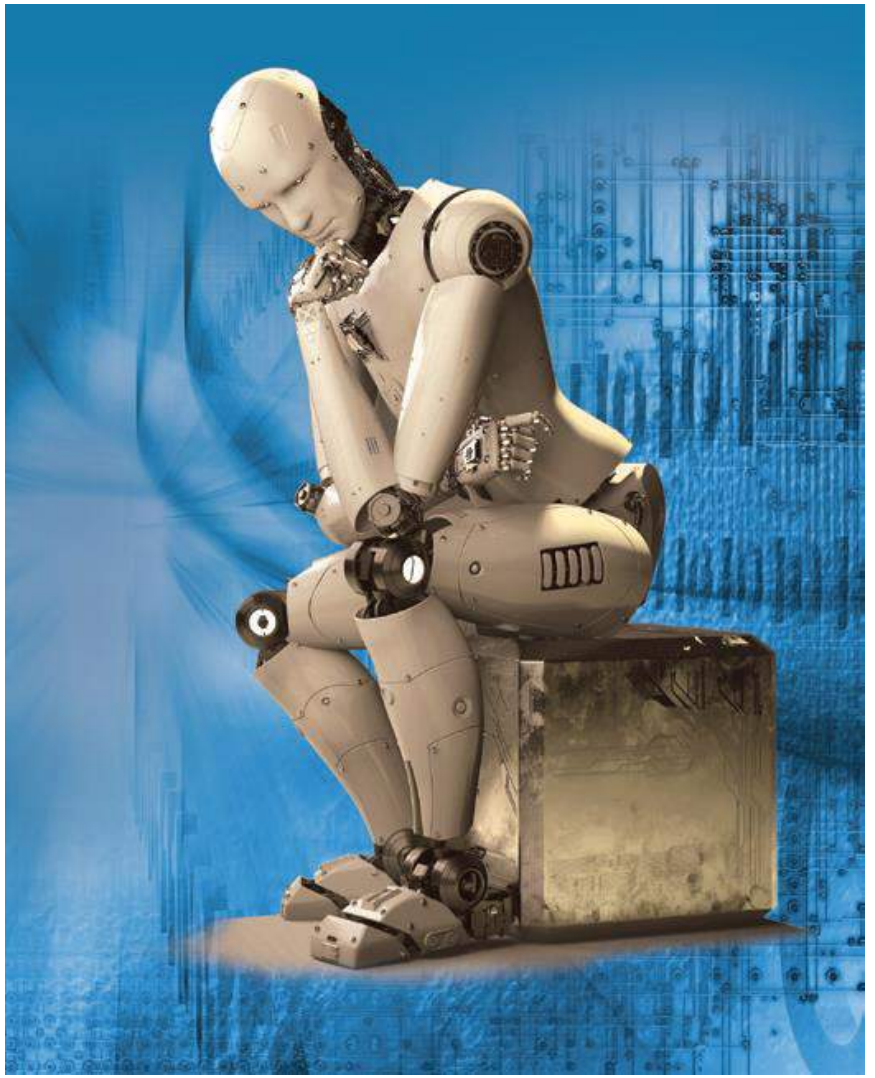
Los ordenadores tienen una forma débil de creatividad que les permite reproducir patrones de habla, sonido o formas, pero nada más. De un ordenador no puede esperarse que produzca algo radicalmente imprevisible, nada similar a lo que supuso la vanguardia o los creadores verdaderamente disruptivos en la historia de las artes.

https://www.ideal.es/opinion/sueno-maquina-creativa-20230205190531-nt.html

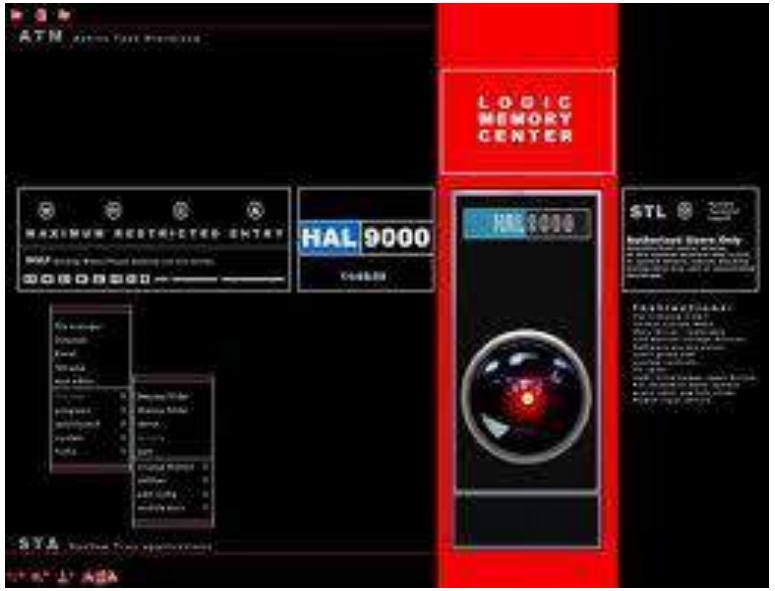
Concluyendo: Las profesiones del futuro

¿Qué harán los robots dentro de 15 años?

"Entre el 50% y 65% de los alumnos que acceden hoy a la escuela primaria trabajarán en profesiones que aún no existen"



Concluyendo: Inteligencia artificial. Ficción versus realidad



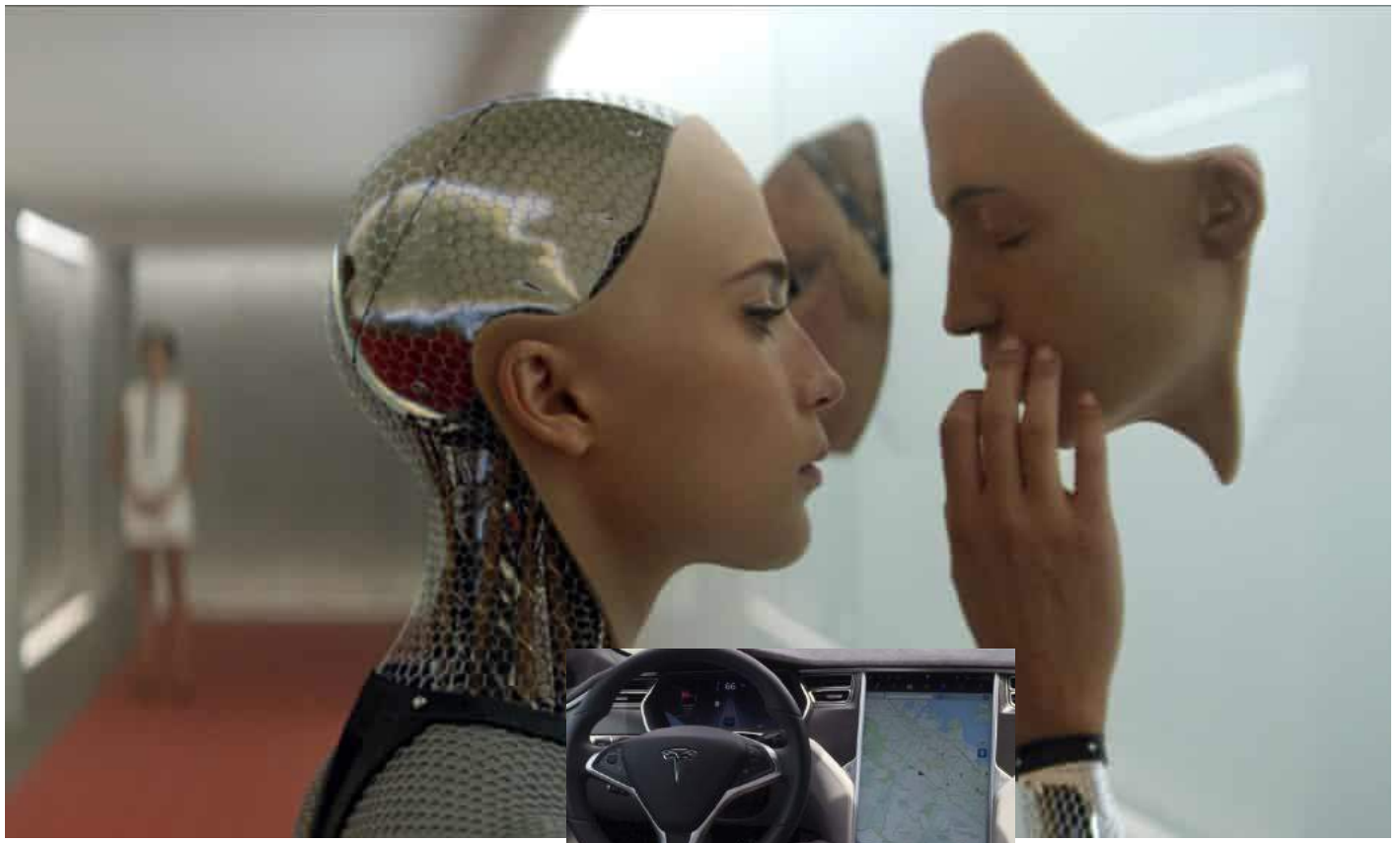
HAL 9000
2001: Una odisea del espacio
Stanley Kubrick 1968

Samantha
Her
Speake Jonze, 2013

Ava
Ex Machina
Alex Garland, 2015

Concluyendo: Inteligencia artificial. Ficción versus realidad

¿Como va a interactuar un sistema de IA con los humanos?



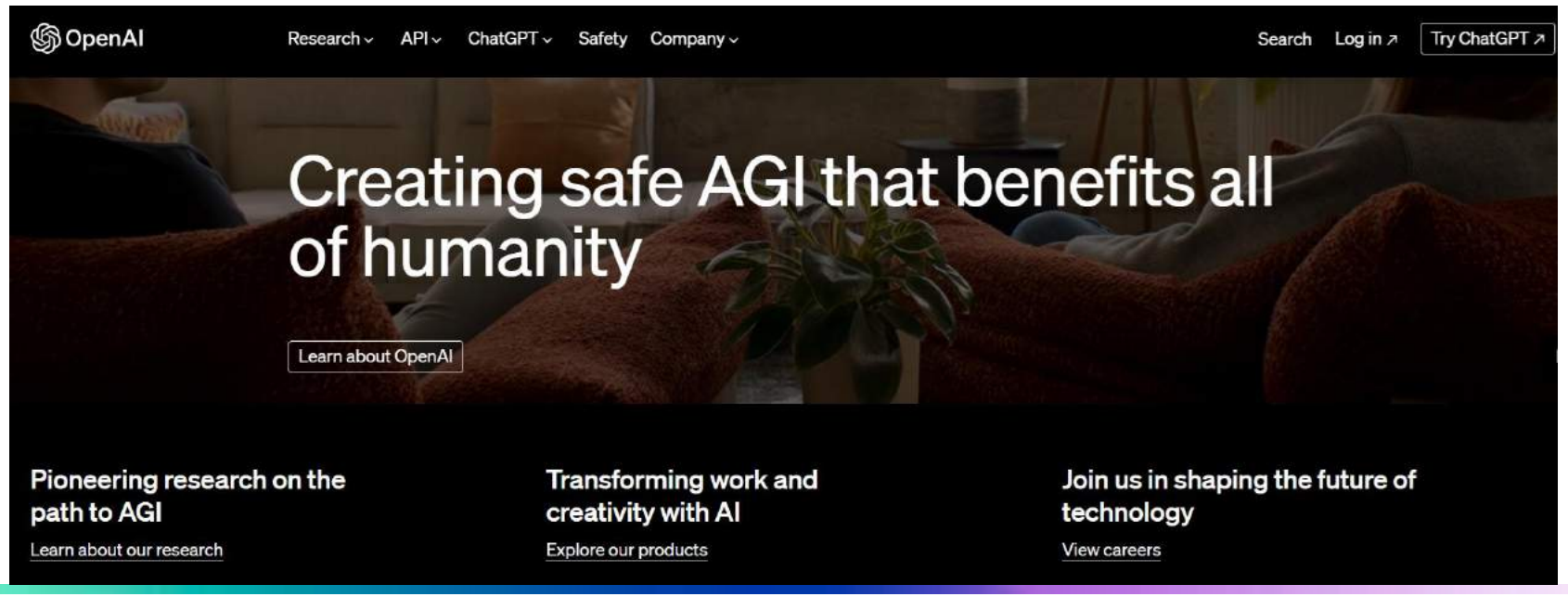
¿Estamos preparados para convivir con máquinas inteligentes?

¿Cómo actuaríamos hoy ante un coche sin conductor?

Concluyendo: Debate. ¿Alcanzaremos una IA general que iguale al humano?

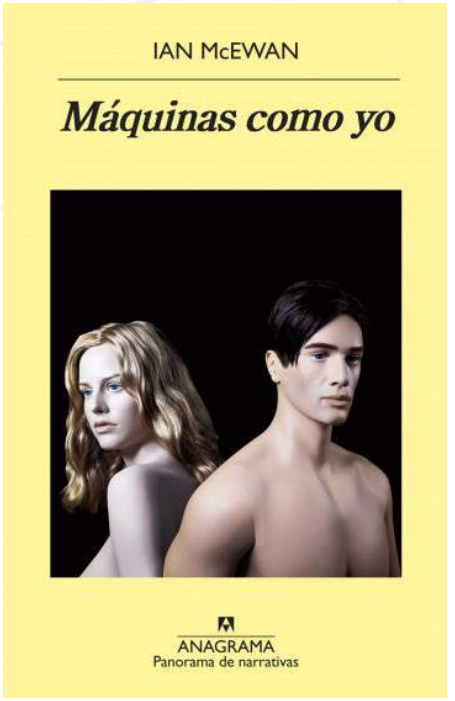
Oriol Vinyal (Responsable de Investigación en DeepMind-Google):

“Nuestra generación verá una inteligencia artificial que iguale o supere a la del ser humano”



OpenAI

La INTELIGENCIA ARTIFICIAL
silenciosa ha entrado en nuestras vidas



Lecturas (novelas) recomendadas (distopías del futuro)

GRANADA INTELIGENCIA ARTIFICIAL

OUTLINE

Artificial Intelligence hatching: Between risk and regulation

Responsible and safe AI: Trustworthy AI

Responsible AI Systems and auditability: European risk-based scenario

Trustworthy AI. A decalogue. Collaboration

Ethics, regulation and social implications: ELSEC

Conclusions

Trustworthy AI: A decalogue

KR1. Human agency and oversight

Cooperation Humans-Machines

KR2. Technical robustness and safety

Emerging technologies. General Purpose AI

KR3. Data Privacy + data quality and data governance

Sandboxes for testing in high risk scenarios. Auditability

KR4. Transparency, traceability, explainability

KR7. Accountability and responsible principle

KR5. Diversity, non-discrimination and fairness

KR6. Societal and environmental well-being

Trustworthy AI decalogue: Transparency: traceability, explainability, and communication (KR4)

Traceability is defined as the set of mechanisms and procedures aimed to keep track of the system's data, development and deployment processes

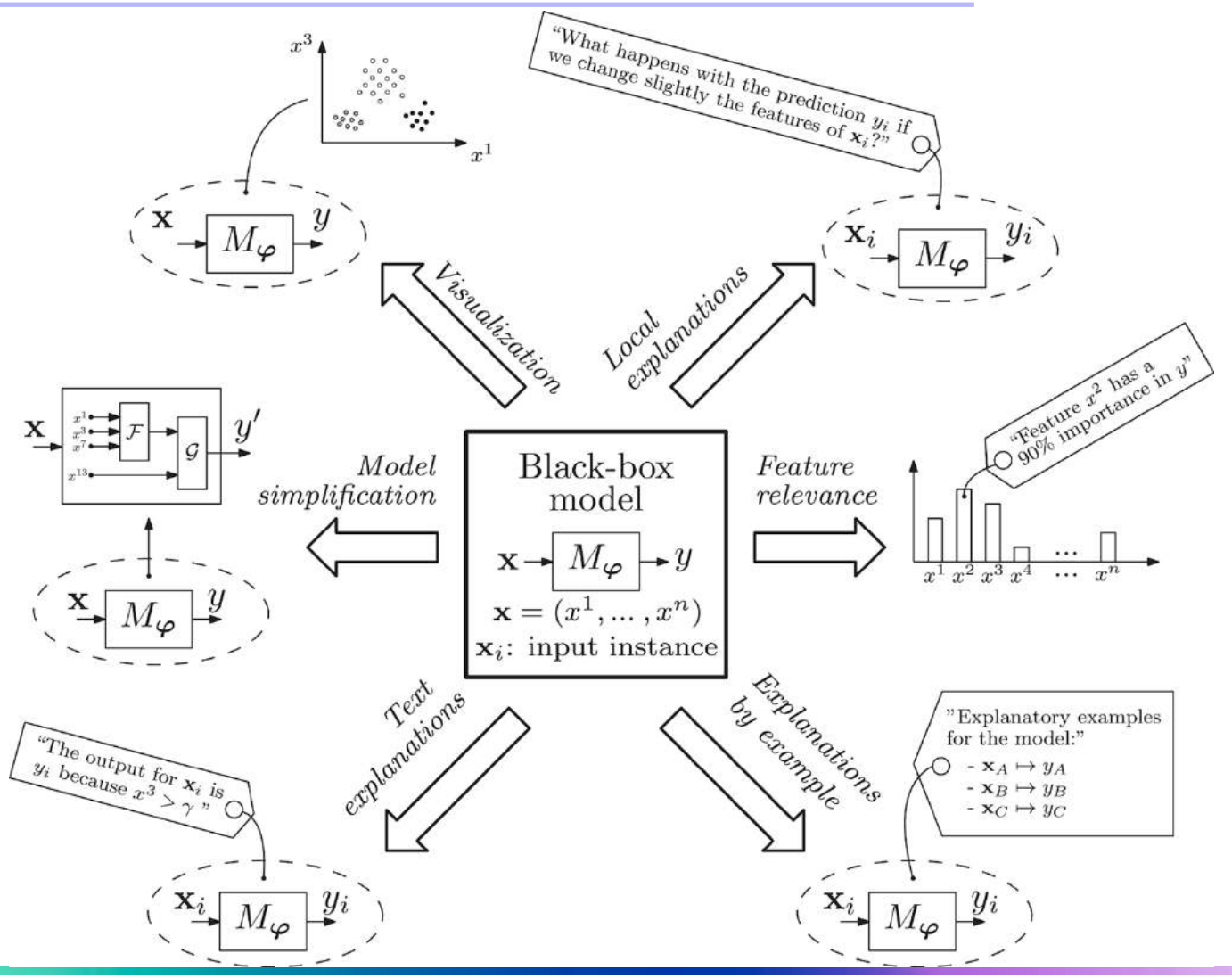
Explainability

***Given an audience**, an **explainable** AI is one that produces details or reasons to make its functioning clear or easy to understand.*

Communication is a crucial dimension, so that all aspects related to transparency are delivered to the audience in a form and format adapted to their background and knowledge. This is key to attain trust in the audience about the AI-based system at hand.

Barredo Arrieta A *et al.*. Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI Inf. Fusion, 58 (2020), pp. 82-115

Trustworthy AI decalogue: Transparency: traceability, explainability, and communication (KR4)

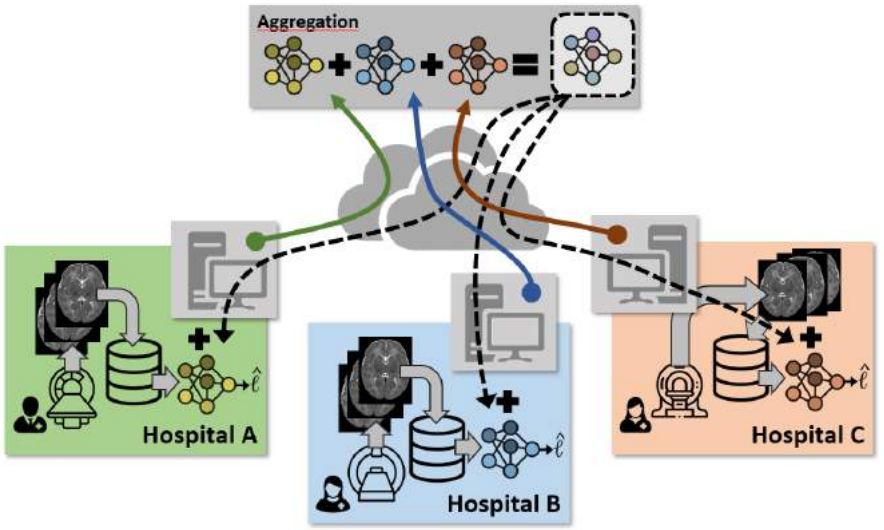


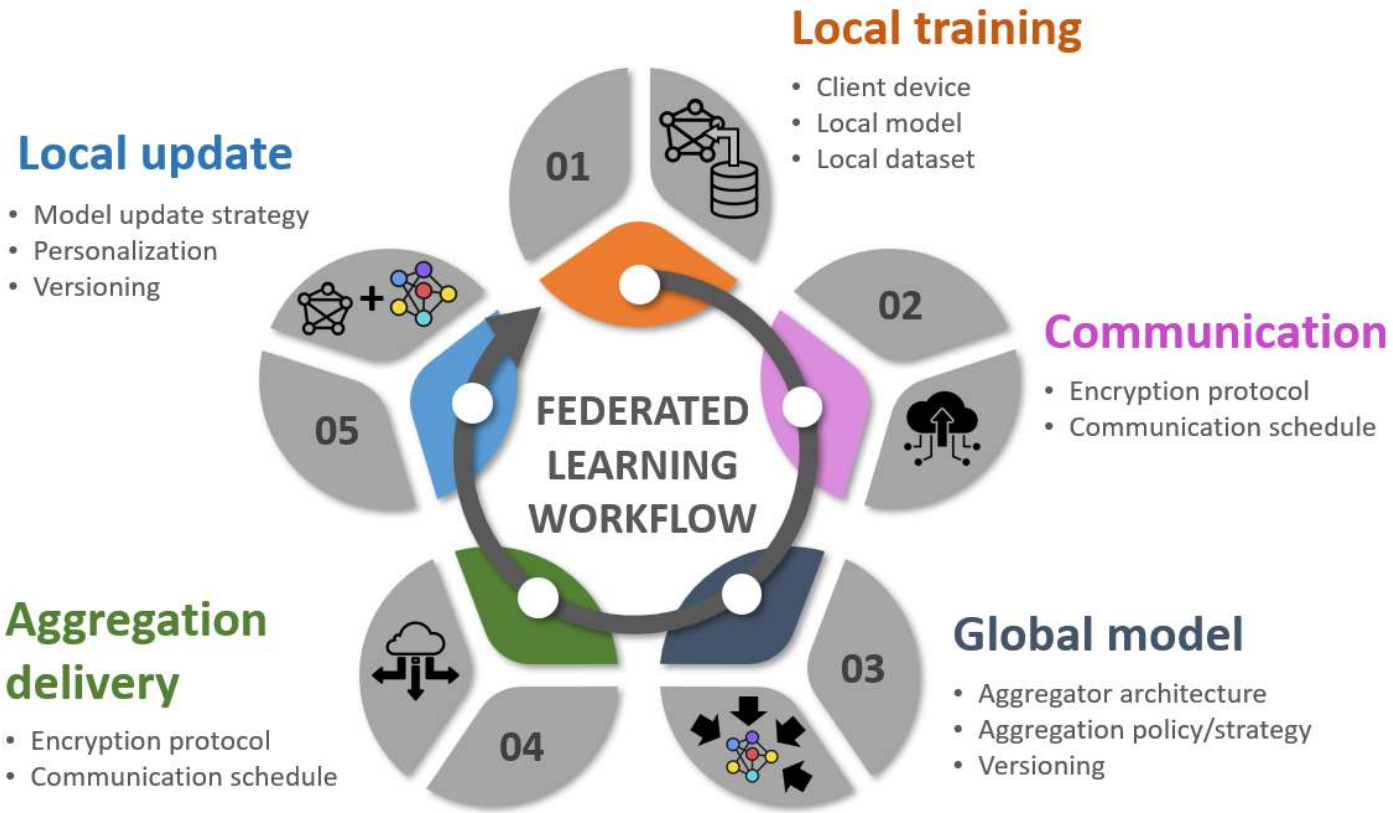
Challenge:

Post-hoc explainability approaches

Barredo Arrieta A *et al.*. Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI Inf. Fusion, 58 (2020), pp. 82-115

Trustworthy AI decalogue: Privacy (and Federated Learning) (KR3)

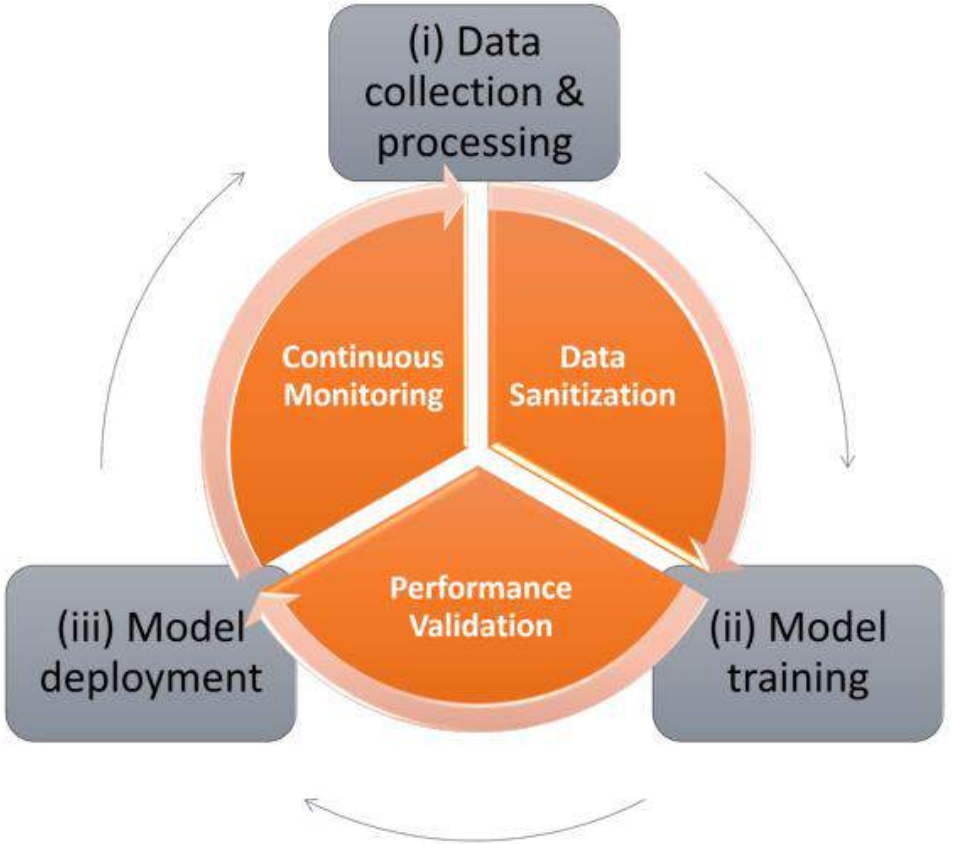




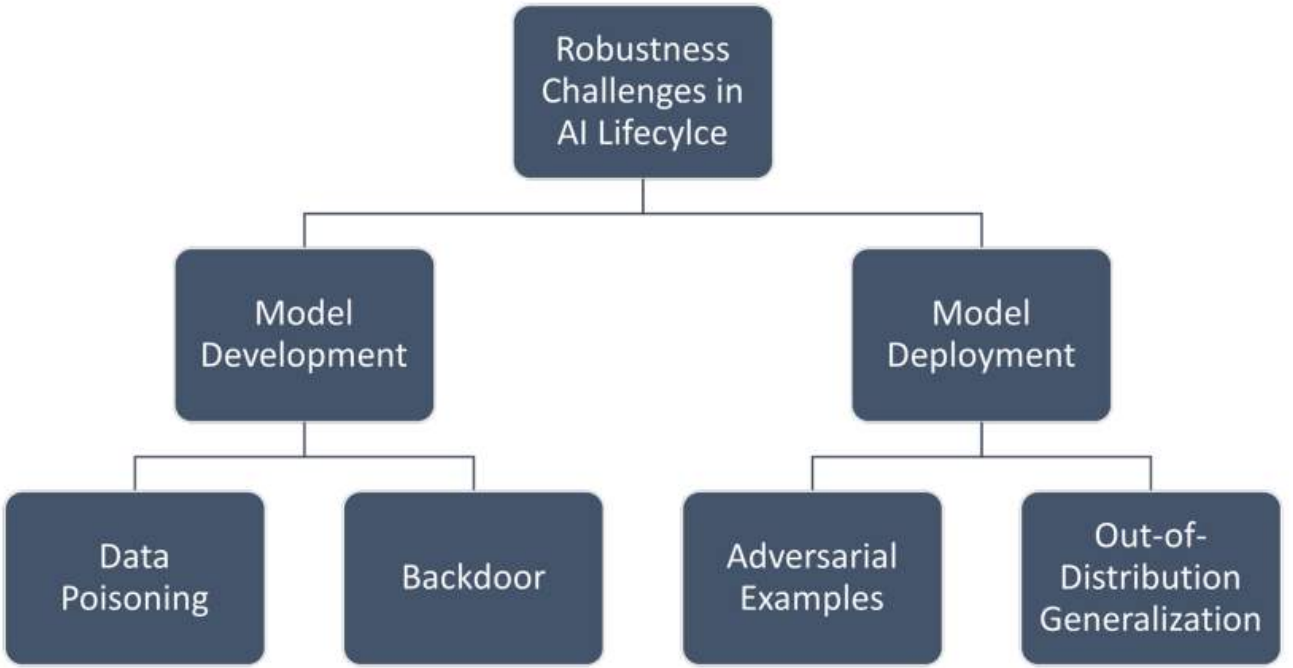
Federated learning process and workflow

A Tutorial on Federated Learning from Theory to Practice: Foundations, Software Frameworks, Exemplary Use Cases, and Selected Trends, Working in progress, 2023.

Trustworthy AI decalogue: Robustness and safety (KR2)



Robustness inspection pipeline



Highlighted robustness challenges in the AI lifecycle

Chen P.-Y., Das P., AI Maintenance: A Robustness Perspective Computer, 56 (2) (2023), pp. 48-56

Trustworthy AI decalogue:

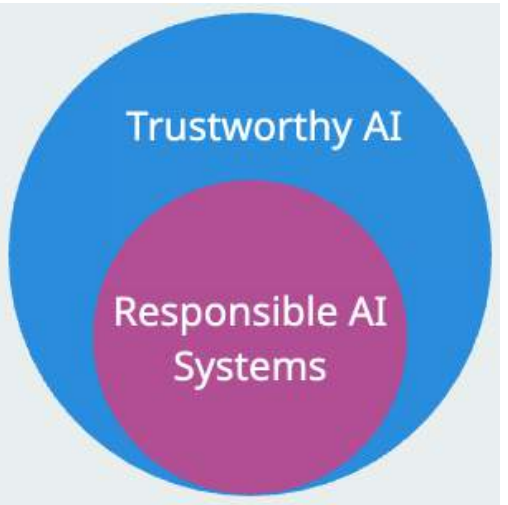
(8) Sandboxes for testing in high risk scenarios. Auditability

A Responsible AI system *"requires ensuring* ***auditability and accountability*** *during its design, development and use, according to specifications and the applicable regulation of the domain of practice to be used."*

A key element in the IA Act scenario is the concept of ***"Sandbox".***

Sandbox: Regulatory sandboxes are indeed recommended by the European AI Act (Chapt. 5, Art. 53–54).

Concretely, the AI Act establishes that **algorithms should comply with regulation and can be tested in a safe environment prior to entering the market.** The auditing process can be implemented via regulatory sandboxes.



Díaz-Rodríguez et al. *Connecting the Dots in Trustworthy Artificial Intelligence: From AI Principles, Ethics, and Key Requirements to Responsible AI Systems and Regulation* . *Information Fusion*, p.101896., 2023.
Pop, F., Adomavicius L. Sandboxes for responsible artificial intelligence. EIPA briefing september (2021).

Trustworthy AI decalogue:

(9) Regulation must be extended:

- General purpose artificial Intelligence systems (Foundation models, GenAI, ChatGPT, GPT4, Bard, …)
- Neurotechnology and AI

THE AI ACT

https://artificialintelligenceact.eu/
https://www.europarl.europa.eu/news/en/press-room/20230609IPR96212/meps-ready-to-negotiate-first-ever-rules-for-safe-and-transparent-ai

Trustworthy AI decalogue: Privacy + Data quality and governance (KR3++)

Use high-quality **training, validation and testing data** (relevant, representative etc.) **(AI Act, Title III, Chapter 2)**

Data-centric AI. It is the discipline of systematically engineering the data to build an AI system. The emphasis is on **data quality** and creating tools to facilitate building AI solutions with the highest accuracy. Some essentials to discussion:

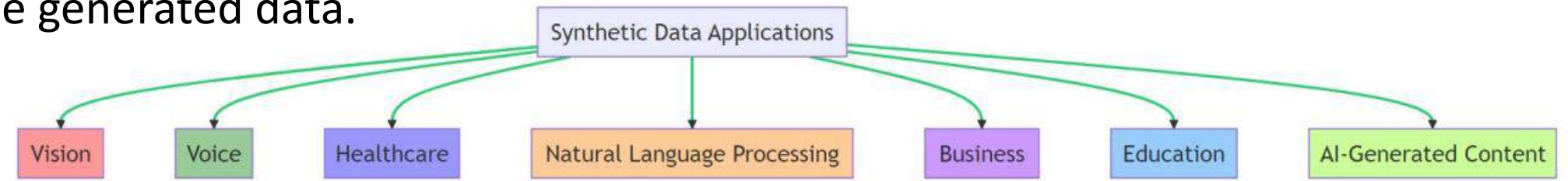
- **Data quality** needs and practice in selected sectors.
- **Data quality assessment, data preprocessing,** and **bias mitigation** in research and industry.
- **Data 'fairness',** decision making processes based on AI, objectivity and quantitative measures.

Alexandra Balahuret al.. Data quality requirements for inclusive, non-biased and trustworthy AI: Putting-Science-Into- Standards. 2022. hal-03889925
D. Zha, Z. Pervaiz Bhat, K-H. Lai, F. Yang, X. Hu, Data-centric AI: Perspectives and Challenges, arXiv:2301.04819v3, 2023.

Trustworthy AI decalogue: Privacy + Data quality and governance (KR3++)

Use high-quality **training, validation and testing data** (relevant, representative etc.) **(AI Act, Title III, Chapter 2)**

Trustworthiness in "quality" synthetic data generation. crucial aspects of privacy and fairness concerns related to synthetic data generation, outliers and corner cases generation and evaluation metrics are essentials to determine the reasonableness of the generated data.



Governance for data trust (creating datasets for AI): Data provenance, collection & presentation (minimum criteria, origin…), using and sharing data, checklist for dataset conformity, trusted data, data privacy and integrity, data maintenance and robustness.

B van der Sloot, E Keymolen, Can we trust trust-based data governance models? Data & Policy 4, e45, doi:10.1017/dap.2022.36
Marijn Janssen, Paul Brous, Elsa Estevez, Luis S. Barbosa, Tomasz Janowski, Data governance: Organizing data for trustworthy Artificial Intelligence, Government Information Quarterly 37:3, 2020, 101493
Yingzhou Lu, Minjie Shen, Huazheng Wang, Wenqi Wei, Machine Learning for Synthetic Data Generation: A Review, arXiv:2302.04062v4, 2023
B van Breugel, N Seedat, F Imrie, M van der Schaar. Can You Rely on Your Model Evaluation? Improving Model Evaluation with Synthetic Test Data , arXiv preprint arXiv:2310.16524, 2023

Trustworthy AI decalogue: Privacy + Data quality and governance (KR3++)

Use high-quality **training, validation and testing data** (relevant, representative etc.) **(AI Act, Title III, Chapter 2)**

Concluding on the data

Everyone wants to do the model work, not the data work.

The only reason machine learning works is because the data contains knowledge, not because the heuristics to extract it are perfect (as the No Free Lunch theorem demonstrates, for any heuristic there is a data set where it doesn't work!).

It seems that sometimes "we have forgotten" that **the important thing is that the data represents the distribution to be learned**, and this has an at least moderately simple structure, avoiding biases, including corner cases generation, …

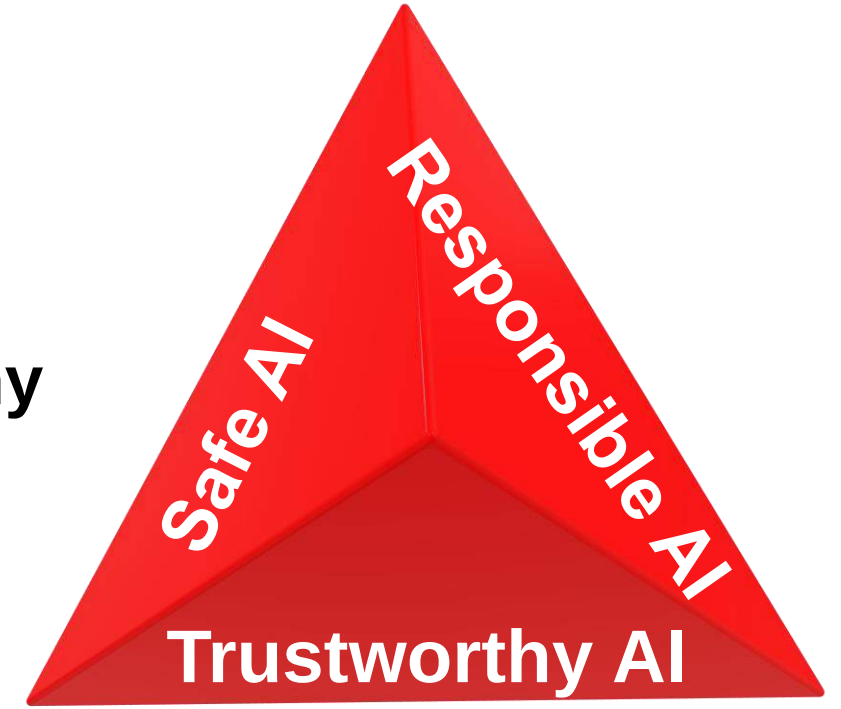
N Sambasivan, S Kapania, H Highfill, D Akrong, P Paritosh, LM Aroyo "Everyone wants to do the model work, not the data work": Data Cascades in High-Stakes AI proceedings of the 2021 CHI Conference, 2021.
L Aroyo, M Lease, P Paritosh, M Schaekermann Data excellence for AI: why should you care? Interactions 29 (2), 66-69, 2022

ELSEC: Ethics, regulation and social implications

Approach: Addressing ethical, legal, socio-economic and cultural (ELSEC) aspects of AI-based systems.

Challenge: How to define an ELSEC Trustworthy AI framework?

defining goals, challenges, and metrics aligned with the requirements of trustworthy AI and AI safety.



ELSEC AI framework

ELSEC: Ethics, regulation and social implications

Objective: A key objective is developing a **comprehensive ELSEC Trustworthy AI framework** that provides **ethical guidance** not just for the project's high-risk scenarios, but more universally applicable, **ethical theories in the context of AI.**

Highlighting two goals (among others):

- identifying **aligned values** and prescribing morally **correct actions for AI practitioners (machine ethics, AI alignment, robustness, …)**
- Using of artificial intelligence for **humanitarian obligations and work (AI for social good).**

Action: Addressing, from a global perspective, the ethical aspects, regulation and social implications of the use of AI in our society, such as the **ethical, legal, socio-economic and cultural (ELSEC) aspects of AI-based systems**, especially those aiming to mitigate humanitarian problems.

Daniel Trusilo and David Danks. Artificial intelligence and humanitarian obligations. Ethics and Information Technology (2023) 25:12
V Sekara, M Karsai, E Moro, D Kim, E Delamonica, M Cebrian, M. Luengo-Oroz, R. Moreno Jiménez, M. Garcia-Herranz Are machine learning technologies ready to be used for humanitarian work and development? arXiv preprint arXiv:2307.01891

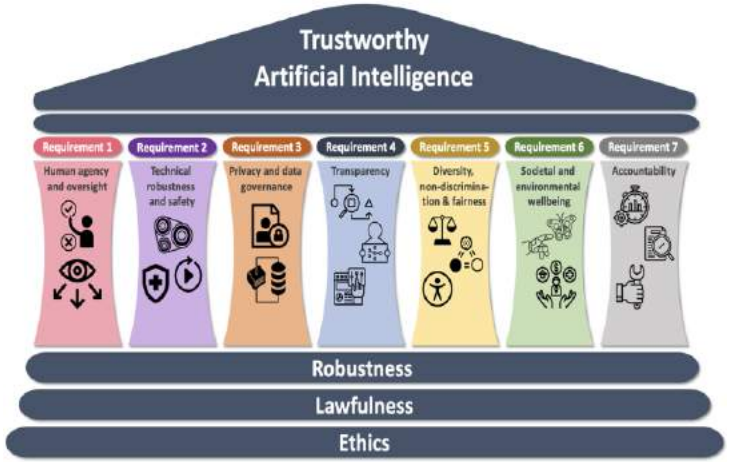
3 challenges for applying Regulation

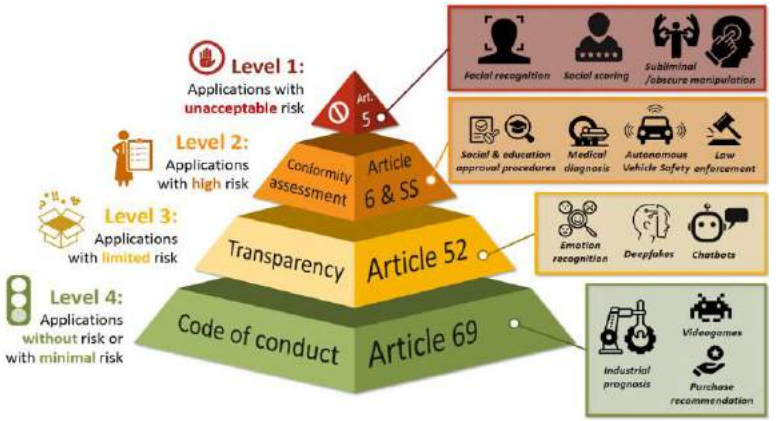
Auditability, property sought for the AI-based system

A) What must a responsible AI system accomplish by design in a specific high-risk application?

B) How to audit intelligent systems in current use by companies?

C) How to continually evaluate and address the entire life cycle of the AI system? (AI Safety)





Lectures for a reflection on trustworthiness and governance

What is the measure/level of "Trustworthy AI" that we can technically achieve? (explainability, privacy, …) → risk and trustworthiness measurement

AI and Ethics
https://doi.org/10.1007/s43681-023-00351-z

ORIGINAL RESEARCH

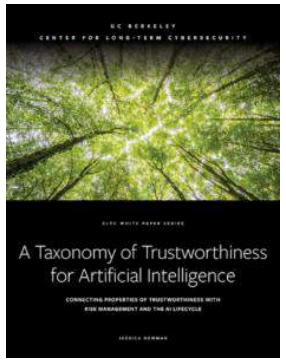
Can machines be trustworthy?

Anders Søgaard[1] Published online: 04 October 2023

Expert Systems With Applications 235 (2024) 121220
Expert Systems With Applications
journal homepage: www.elsevier.com/locate/eswa

Artificial Intelligence risk measurement

Paolo Giudici [a,*], Mattia Centurelli [b], Stefano Turchetta [c]
[a] University of Pavia, Italy
[b] Credito Emiliano, Italy
[c] Independent, Italy



White Paper / January 2023

A Taxonomy of Trustworthiness for Artificial Intelligence

CONNECTING PROPERTIES OF TRUSTWORTHINESS WITH RISK MANAGEMENT AND THE AI LIFECYCLE

By  Jessica Newman

Berkeley CLTC Center for Long-Term Cybersecurity

This paper introduces a taxonomy of trustworthiness for artificial intelligence that includes 150 properties. Each property relates to one of seven "characteristics of trustworthiness" as defined in the NIST AI RMF.

NIST Characteristics of Trustworthiness

- **Valid and Reliable**
- **Safe**
- **Fair with Harmful Bias Managed**
- **Secure and Resilient**
- **Explainable and Interpretable**
- **Privacy-Enhanced**
- **Accountable and Transparent**

https://cltc.berkeley.edu/publication/a-taxonomy-of-trustworthiness-for-artificial-intelligence/

Lectures for a reflection on trustworthiness and governance

Development of AI Governance Frameworks.

How will we ensure an AI system is designed and engineered to achieve its goals while maintaining the ability to disengage or deactivate the system if necessary?

How will we ensure an AI system would not have incentives to resist or deceive its operators?


Received: 11 April 2023 | Revised: 15 June 2023 | Accepted: 4 July 2023
DOI: 10.1111/exsy.13406
ORIGINAL ARTICLE
Expert Systems. 2023;e13406.
https://doi.org/10.1111/exsy.13406
Expert Systems WILEY

Artificial intelligence governance: Ethical considerations and implications for social responsibility

Mark Anthony Camilleri[1,2,3]


ELSEVIER
Telecommunications Policy 47 (2023) 102479
Telecommunications Policy
journal homepage: www.elsevier.com/locate/telpol

A 'biased' emerging governance regime for artificial intelligence? How AI ethics get skewed moving from principles to practices

Nicola Palladino
Trinity College Dublin, Trinity Long Room Hub, Ireland

montrealethics.ai/connecting-the-dots-in-trustworthy-artificial-intelligence-from-ai-principles-ethics-and-key-requirements-to-responsible-ai-systems-and-regulation/

Connecting the dots in trustworthy Artificial Intelligence: From AI principles, ethics, and key requirements to responsible AI systems and regulation.

https://montrealethics.ai/connecting-the-dots-in-trustworthy-artificial-intelligence-from-ai-principles-ethics-and-key-requirements-to-responsible-ai-systems-and-regulation/

ELSEVIER

Information Fusion

Volume 99, November 2023, 101896

Full length article

Connecting the dots in trustworthy Artificial Intelligence: From AI principles, ethics, and key requirements to responsible AI systems and regulation ☆

Natalia Díaz-Rodríguez [a 1], Javier Del Ser [b c 1], Mark Coeckelbergh [d], Marcos López de Prado [e f g], Enrique Herrera-Viedma [a], Francisco Herrera [a]

July 29, 2023 by MAIEI

Research Summary by **1) Natalia Díaz Rodríguez, 2) Javier Del Ser, 3) Mark Coeckelbergh, 4) Marcos López de Prado, 5) Enrique Herrera-Viedma,** and **6) Francisco Herrera**

https://www.sciencedirect.com/science/article/pii/S1566253523002129#sec4