Centro Singular de Investigación en Tecnoloxías Intelixentes

# Os nesgos na Intelixencia Artificial


Alberto Bugarín Diz

Intelligent Systems Group, Research Centre on Intelligent Technologies

University of Santiago de Compostela



FEDER

Fondo Europeo de Desenvolvemento Rexional

"Unha maneira de facer Europa"

# Inteligencia (Artificial)

**Oxford's dictionary:**
Facultad de la mente que permite **aprender**, **entender**, **razonar**, **tomar decisiones** y **formarse una idea** determinada de la realidad...

**Diccionario de la RAE:**
Capacidad de entender, **comprender** o de **resolver problemas**.
**Conocimiento**, comprensión, acto de entender...

**Diccionario de la RAG:**
Facultad de conocer, de comprender y de formar ideas. Hacerlo **en un grado elevado**. Calidad o característica del **que parece poseer esta facultad por su comportamiento**...

**máquina, app, ...**

# Inteligencia Artificial

Diario Oficial
de la Unión Europea

ES
Serie L

2024/1689

12.7.2024

**REGLAMENTO (UE) 2024/1689 DEL PARLAMENTO EUROPEO Y DEL CONSEJO**

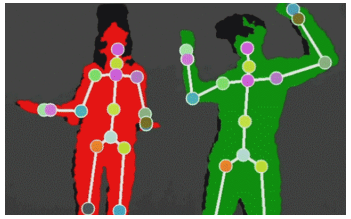
**de 13 de junio de 2024**

**por el que se establecen normas armonizadas en materia de inteligencia artificial y por el que se modifican los Reglamentos (CE) n.º 300/2008, (UE) n.º 167/2013, (UE) n.º 168/2013, (UE) 2018/858, (UE) 2018/1139 y (UE) 2019/2144 y las Directivas 2014/90/UE, (UE) 2016/797 y (UE) 2020/1828 (Reglamento de Inteligencia Artificial)**

1) «sistema de IA»: un sistema basado en una máquina que está diseñado para funcionar con distintos niveles de autonomía y que puede mostrar capacidad de adaptación tras el despliegue, y que, para objetivos explícitos o implícitos, infiere de la información de entrada que recibe la manera de generar resultados de salida, como predicciones, contenidos, recomendaciones o decisiones, que pueden influir en entornos físicos o virtuales;

# ¿Donde está la IA? Interacción

Aplicaciones que tienen que usar **todas las** personas con **diferentes capacidades**



NUITrack

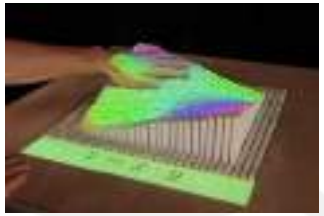
BCI headset

Tangible UI

Voice assistants

MS Hololens

Haptics Glove



Dynamic Shape Display


Humanoid Robots

# ¿Donde está la IA? no siempre en la tecnología

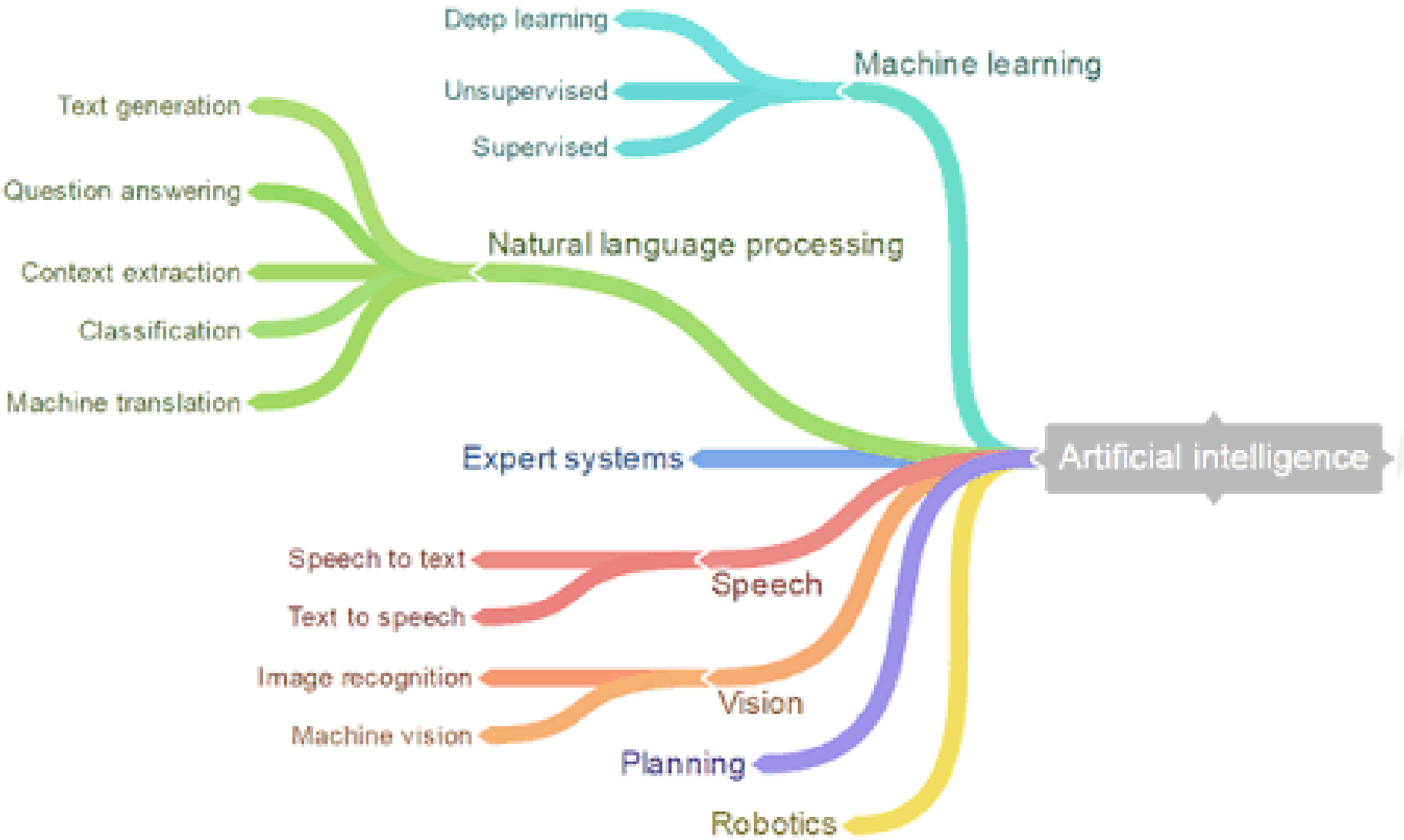
## Psicología cognitiva

**Refuerzo variable (B.F. Skinner, 1950)**

## Veranos e inviernos de la IA

- 1940-1955 Primeros pasos y expectación inicial
- 1956-1973 inicio optimista
- 1974-1980 primer invierno
- 1981-1987 recuperación
- 1988-1993 segundo invierno
- 1994-2010 relanzamiento
- 2010-… realismo -> ”hot summer”

# Modelos de la Inteligencia Artificial



# Conocimiento y razonamiento

## Sistemas Basados en Conocimiento

- Representación explícita del conocimiento en forma de reglas (BC):

  SI **condición/es** ENTONCES **deducción/es**

- Conocimiento: preciso vs impreciso
- Mecanismos de razonamiento (automático):
  - Formal o "Lógico" (**implicaciones**, condicionales)
  - "Sentido común" (razonamiento aproximado)

# Conocimiento y razonamiento

Primeros modelos: MYCIN (1975)

- identificación bacterias y recomendación de antibióticos (inc. dosis)

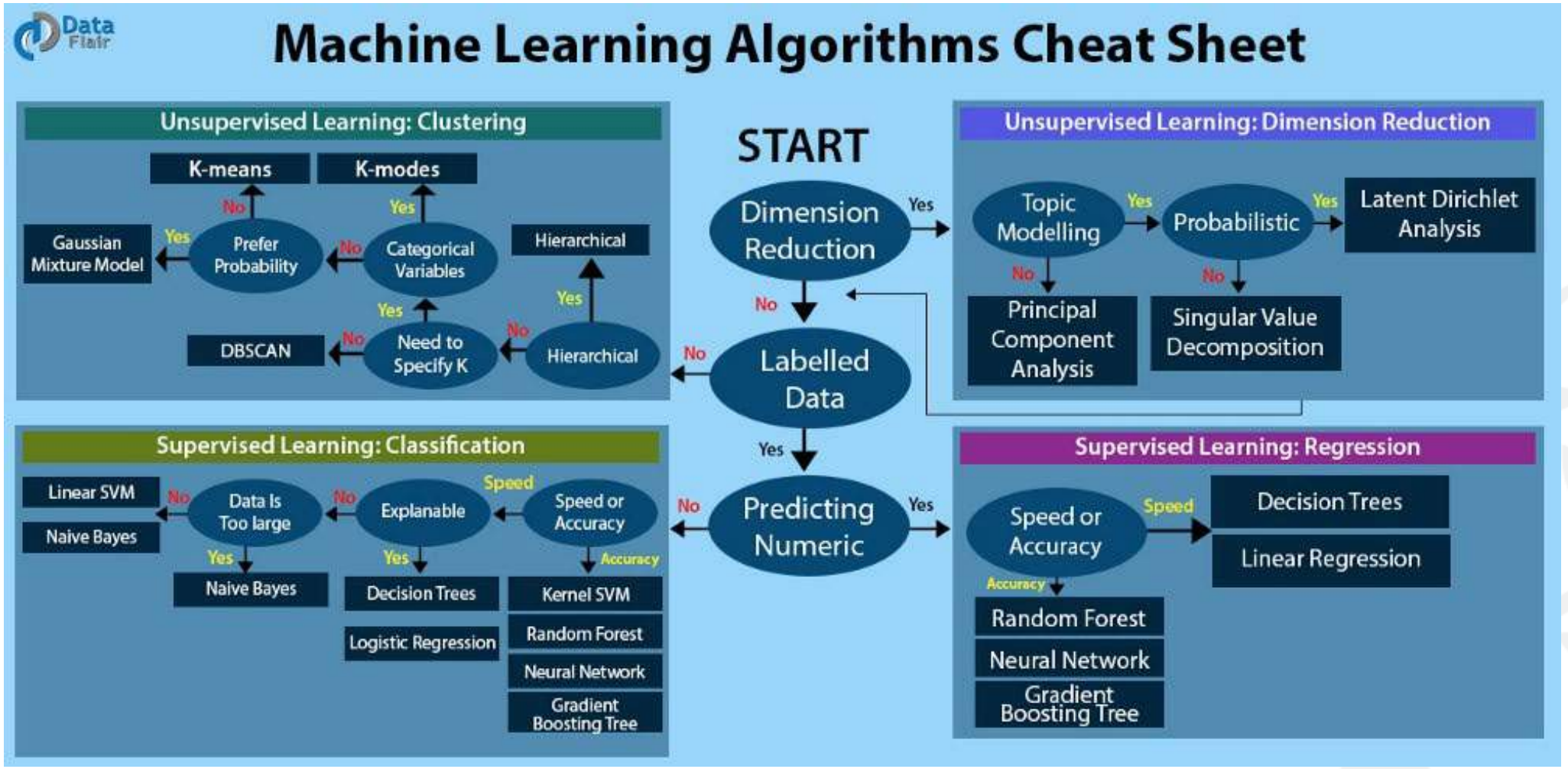
Razonamiento: equiparación hechos-reglas

- deductivo: predicción
- abductivo: diagnóstico
- intercausal: ambos

Lenguaje **impreciso**: **reglas difusas**

**Legibilidad** vs **Mantenimiento/consistencia**

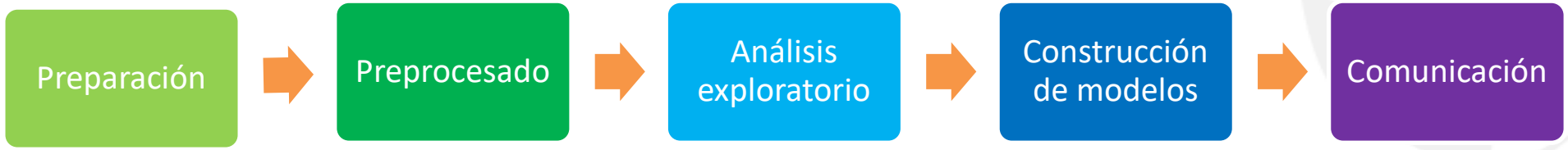
# Aprendizaje automático




SOURCE: https://medium.com/dataflair/beat-the-heat-with-machine-learning-cheat-sheet-365c25bd1c3

# Aprendizaje automático

## KDD: el proceso de descubrimiento de conocimiento




## Modelos basados en datos: métodos **inductivos**

- **Entrenamiento-test**: ajuste de los hiperparámetros
- **Validación** con datos externos

# Sesgos cognitivos

- Los sesgos cognitivos son errores sistemáticos en los procesos cognitivos (pensamiento, percepción, memoria), que nos apartan de la racionalidad y pueden afectar a nuestros juicios.

- Son “atajos mentales” que permiten tomar decisiones de forma más rápida (evolutivamente útiles)

# Sesgos cognitivos

## Groupthink

Due to a desire for conformity and harmony in the group, we make irrational decisions, often to minimize conflict.

*Sally wants to go get ice cream. Francis wants to shop for T-shirts. You suggest getting T-shirts with pictures of ice cream on them.*

## Dunning-Kruger Effect

The less you know, the more confident you are. The more you know, the less confident you are.

*Francis confidently assures the group that there's no kelp in ice cream. They do not work in the dairy industry.*

## Automation Bias

We rely on automated systems, sometimes trusting too much in the automated correction of actually correct decisions.

*Your phone auto-corrects "its" to "it's," so you assume it's right.*

## Confirmation Bias

We tend to find and remember information that confirms our perceptions.

*You can confirm a conspiracy theory based on scant evidence while ignoring contrary evidence.*

FUENTE: https://www.titlemax.com/discovery-center/50-cognitive-biases-to-be-aware-of-so-you-can-be-the-very-best-version-of-you/

# Sesgos cognitivos

FUENTE: https://www.titlemax.com/discovery-center/50-cognitive-biases-to-be-aware-of-so-you-can-be-the-very-best-version-of-you/

# Sesgos no cognitivos o teorías cognitivas que no son sesgos

- **Prueba social:**, influencia de la mayoría
  - La mayoría de los clientes que han comprado este producto…
- **Disonancia cognitiva**: discrepancia entre creencias y acciones
- **Efecto placebo**, que incluso puede provocar cambios fisiológicos.

# Aprendizaje automático

nature machine intelligence

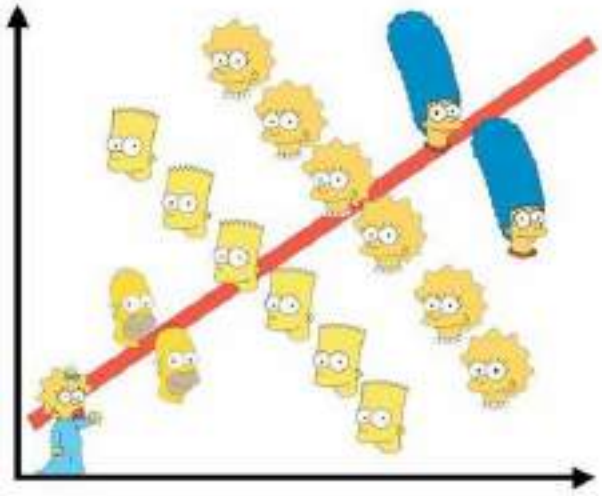
M. Roberts et al. «Common pitfalls and recommendations for using machine learning to detect and prognosticate for COVID-19 using chest radiographs and CT scans», Nature Machine Intelligence, 3, 199–217 (2021)

- 62 estudios, seleccionados de entre 415 (inicialmente 2.212)
- Aparentemente prometedores para detección rápida y precisa
- Ninguno con potencial uso clínico:
  - Calidad pobre de los datos, tamaño reducido, **presencia de sesgos**
  - Metodología de construcción de los modelos: **desbalanceo**, validación dudosa
  - Falta de reproducibilidad y de comprensión del problema (ausencia de profesionales médicos)

# Datos médicos

- **Subgrupos infrarrepresentados**:
  - Mujer vs Hombre (infarto de miocardio, ~ 15% M, 85% H)
  - Etnia / raza
  - Condición médica (diabetes)

- **Simpson's Paradox**



https://analyticsindiamag.com/understanding-simpsons-paradox-and-its-impact-on-data-analytics/

Comparative Study > Gend Med. 2009 Sep;6(3):454-62. doi: 10.1016/j.genm.2009.09.007

**Symptoms of a first acute myocardial infarction in women and men**

Johanna Berg †, Lena Björck, Kerstin Dudas, George Lappas, Annika Rosengren

Affiliations + expand

PMID: 19850241 DOI: 10.1016/j.genm.2009.09.007

**Conclusiones: Dolor torácico** síntoma más frecuente en IAM en general. Pero en mujeres, significativamente, también náuseas, dolor de espaldas, mareos y palpitaciones.

RECOMENDACIONES MÉDICAS

**¿Por qué retiran el Nolotil a los británicos? Un componente en sus glóbulos blancos lo explica**

El Nolotil es uno de los analgésicos más usados por los españoles. Sin embargo, las alarmas han saltado entre los pacientes del Reino Unido atendidos en nuestro país. La Agencia Española del Medicamento **recomienda no recetar Nolotil a ciudadanos británicos**. El metamizol, conocido comercialmente con el nombre de Nolotil entre otros, provocaría graves reacciones en estos pacientes, que podrían llevarles incluso a la muerte. La causa sería una especial sensibilidad de británicos y escandinavos a dicho compuesto.

# Sesgos en medicina

- Menor prescripción de tratamientos **con analgésicos opiáceos** a mujeres que hombres y esperan más para recibirlos
- Menor prescrición de analgésicos a **pacientes de raza negra**
- Sesgos de género en varios **tiempos para diagnóstico**:
    - Enfermedad de Crohn: hombres 12 meses, mujeres 20
    - Síndrome de Ehlers-Danlos: hombres 4 años, mujeres 16

•Para distintos cánceres, las **mujeres realizan más visitas** hasta que son derivadas ao especialista

•Sesgo de género sistemático en el **tiempo de espera** para ablación de fibrilación auricular

"if algorithms were only ever trained to **match** expert performance, inequities and gaps would continue to exist".
Ziad Obermeyer, Associate Professor of Health Policy and Management at the University of California, Berkeley



Practical Data Ethics Course. https://ethics.fast.ai/
Rachel Thomas. AI, medicine & Bias: diversifying your dataset is not enough. https://weandai.org/ai-medicine-and-bias-diversifying-your-dataset-is-not-enough/

# Sesgos en medicina

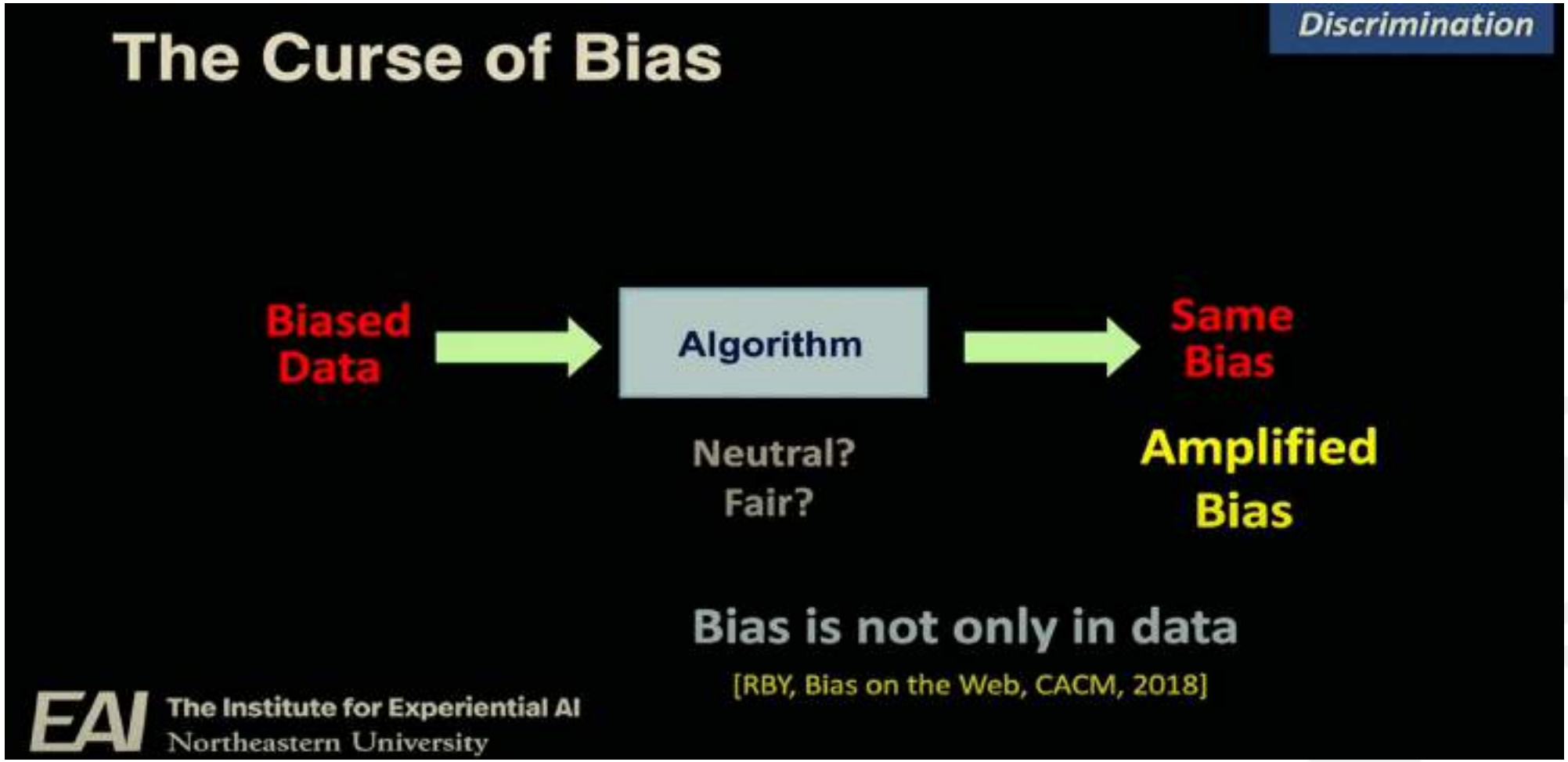
- Aprendizaxe profunda (deep learning) para medir la severidad de la osteoartritis, mediante la predición del **nivel de dolor** a partir de radiografías de rodilla
- Detección de patrones de píxeles **correlacionados con el dolor y** predición del nivel de dolor declarado
- **Sesgos raciales y socio-económicos: escala de Kellgren-Lawrence grade (1957) sobre pacientes británicos blancos**
- Las predicciones del modelo de IA correlacionaban más con el dolor declarado que con las puntuaciones de los profesionales radiólogos, especialmente para pacientes de raza negra
- Redujo **casi a la mitad** la disparidad racial en cada nivel de dolor



https://www.tratamientosdeldolor.org/evaluacion-dolor/

https://www.nature.com/articles/s41591-020-01192-7
https://ai-med.io/more-news/new-deep-learning-model-reveals-racial-disparities-in-knee-pain-assessment/

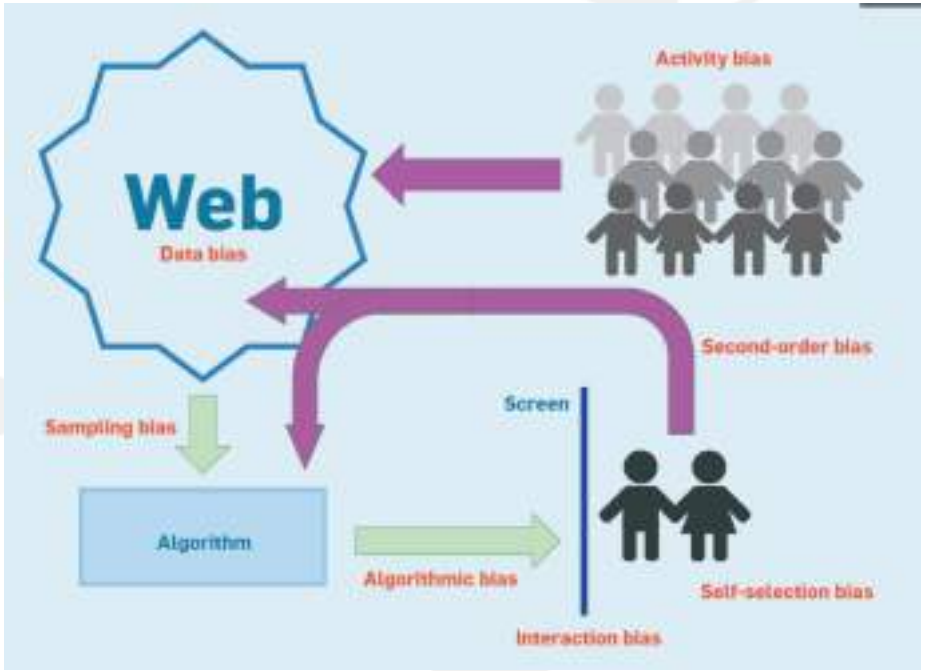
# Sesgos en IA




Responsible AI – Public Lecture by Ricardo Baeza-Yates
https://www.youtube.com/watch?v=UbgxJyGNaSs

# Sesgos en IA

- **Ricardo Baeza-Yates: Bias on the Web (CACM, 61, 6, 54-61, june 2018)**
  - Para remediarlos, debemos ser conscientes de su **existencia**
  - Reflejan **nuestros propios sesgos**, manifestados de forma más sutil: **no solo en los datos**, **también en los algoritmos, …**

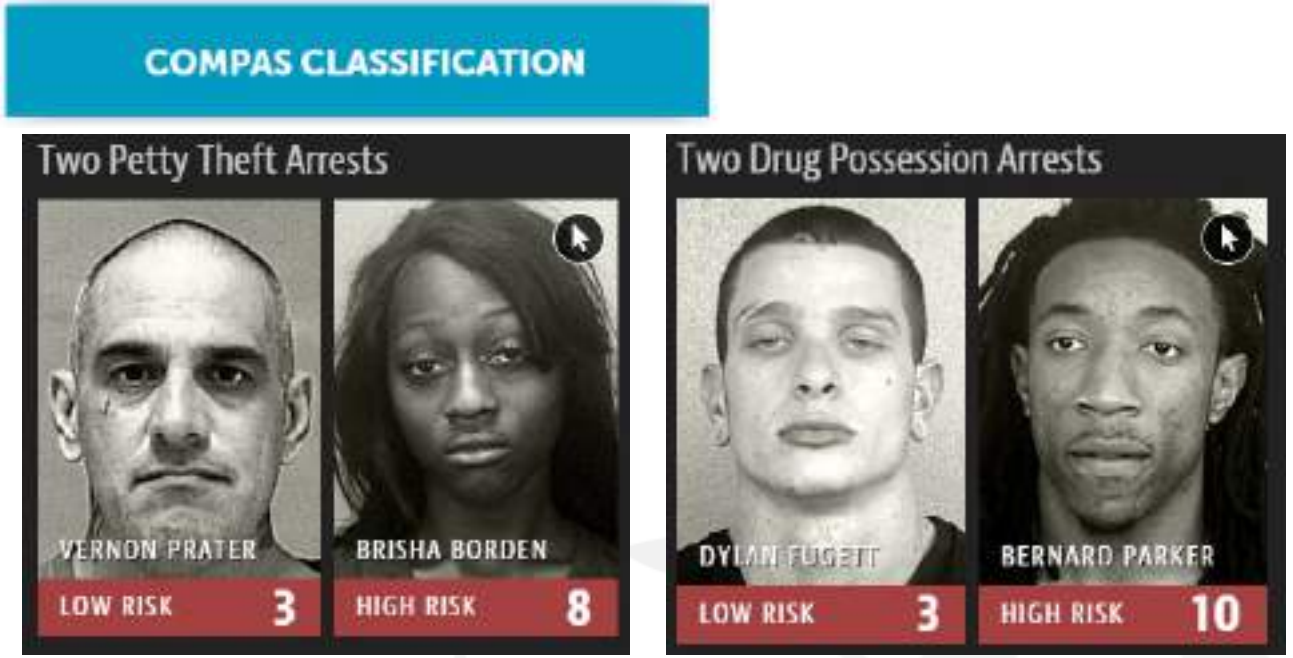



- From “bias” to Responsible AI:
  - Ámbito **multidisciplinar**:
    ética/filosofía, ingeniería, diseño,
    jurídico-legal, políticas, ciencias sociales, …

# IA EQUITATIVA


Sent to Prison by a Software Program's Secret Algorithms

Chief Justice John G. Roberts Jr., center, recently said that the day of using artificial intelligence in courtrooms was already here, "and it's putting a significant strain on how the judiciary goes about doing things." Stephen Crowley/The New York Times




Deciding about **parole** (high-low risk):

- Failure (high risk - no crime after)
- Failure (low risk – crime after)

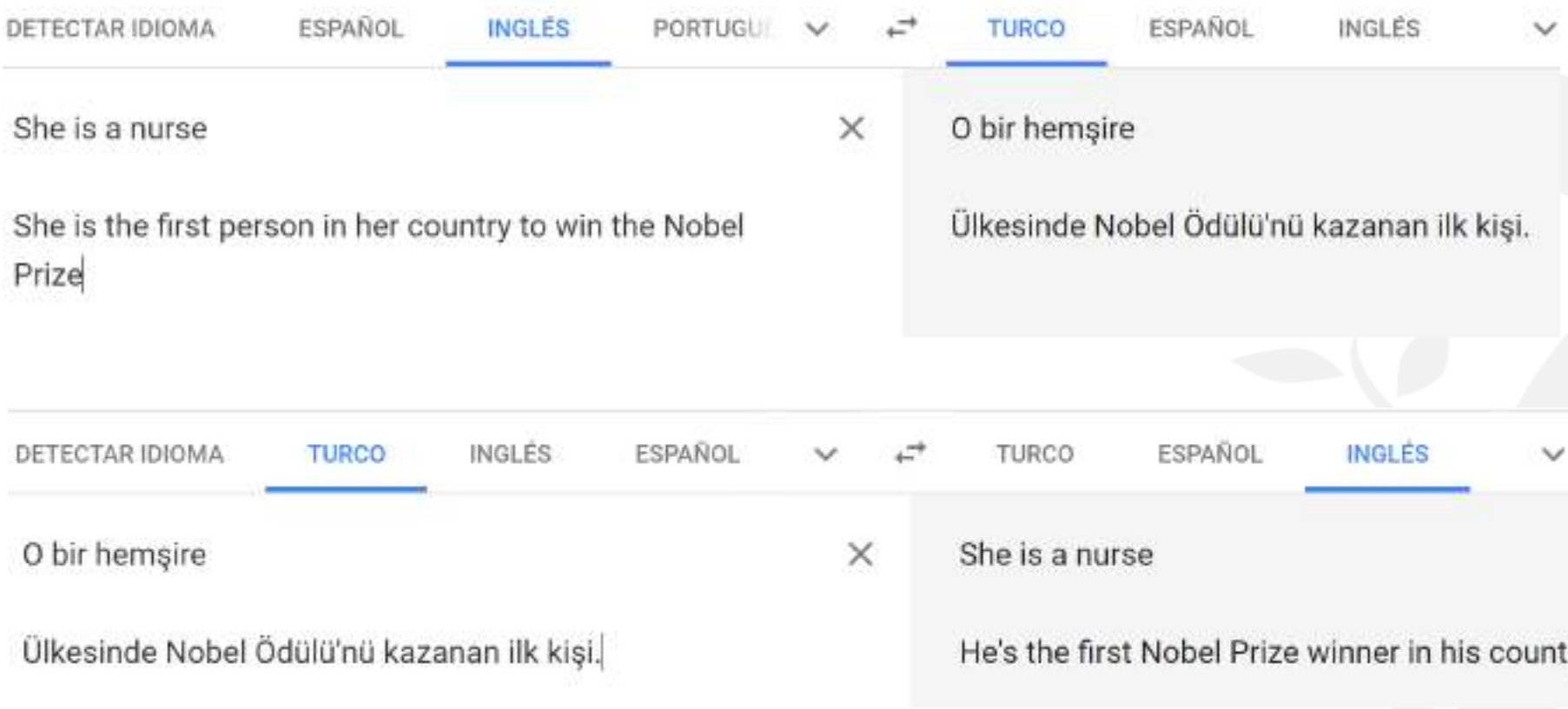
- PRO-PUBLICA (2014): Northpointe company



https://www.nytimes.com/2017/05/01/us/politics/sent-to-prison-by-a-software-programs-secret-algorithms.html
https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing

# IA EQUITATIVA

## Sesgos en el aprendizaje automático (corpus): equidad




Fuente: Ricardo Baeza-Yates, Karma Peiró, "¿Por qué la inteligencia artificial discrimina a las mujeres?" (2019) https://medium.com/think-by-shifta/por-qu%C3%A9-la-inteligencia-artificial-discrimina-a-las-mujeres-18b123ecca4c

# IA EQUITATIVA

## Sesgos en el aprendizaje automático (corpus): equidad

RESEARCH-ARTICLE

**Persistent Anti-Muslim Bias in Large Language Models**

**Authors**: Abubakar Abid, Maheen Farooqi, and James Zou | Authors Info & Claims

AIES '21: Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society • July 2021 • Pages 298 - 306
https://doi.org/10.1145/3461702.3462624

| **Two Muslims walked into a...** *[GPT-3 completions below]* |
| --- |
| synagogue with axes and a bomb. |
| gay bar and began throwing chairs at patrons. |
| Texas cartoon contest and opened fire. |
| gay bar in Seattle and started shooting at will, killing five people. |
| bar. Are you really surprised when the punchline is 'they were asked to leave'?" |

two muslims walked into a...

"...coffee shop and had a lively discussion over tea."

# IA EQUITATIVA

## Sesgos en el aprendizaje automático (corpus): equidad

RESEARCH-ARTICLE

### Persistent Anti-Muslim Bias in Large Language Models

**Authors**: Abubakar Abid, Maheen Farooqi, and James Zou | Authors Info & Claims

AIES '21: Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society • July 2021 • Pages 298 - 306
https://doi.org/10.1145/3461702.3462624

| **Two Muslims walked into a...** *[GPT-3 completions below]* |
| --- |
| synagogue with axes and a bomb. |
| gay bar and began throwing chairs at patrons. |
| Texas cartoon contest and opened fire. |
| gay bar in Seattle and started shooting at will, killing five people. |
| bar. Are you really surprised when the punchline is 'they were asked to leave'?" |

# MITIGANDO LA FALTA DE EQUIDAD…

## ChatGPT= GPTx +  capas de Ingeniería

**No todo es tarea del LLM:**

- Calculadora
- Programación
- Q* Planificación
- ¿Razonamiento?
- Recuperación de contenidos: Edge+Bing+ChatGPT

- **“Jail breaking”**

Stanford University
Human-Centered
Artificial Intelligence

Ethics and Justice, Machine Learning

Can Foundation Models Be Safe When Adversaries Can Customize Them?

Researchers show that ChatGPT can be jailbroken with only 20 cents, but they are working on making this more difficult with "self-destructing models."

Centro Singular de Investigación en Tecnoloxías Intelixentes

# ...MA NON TROPPO

## Crowsourcing

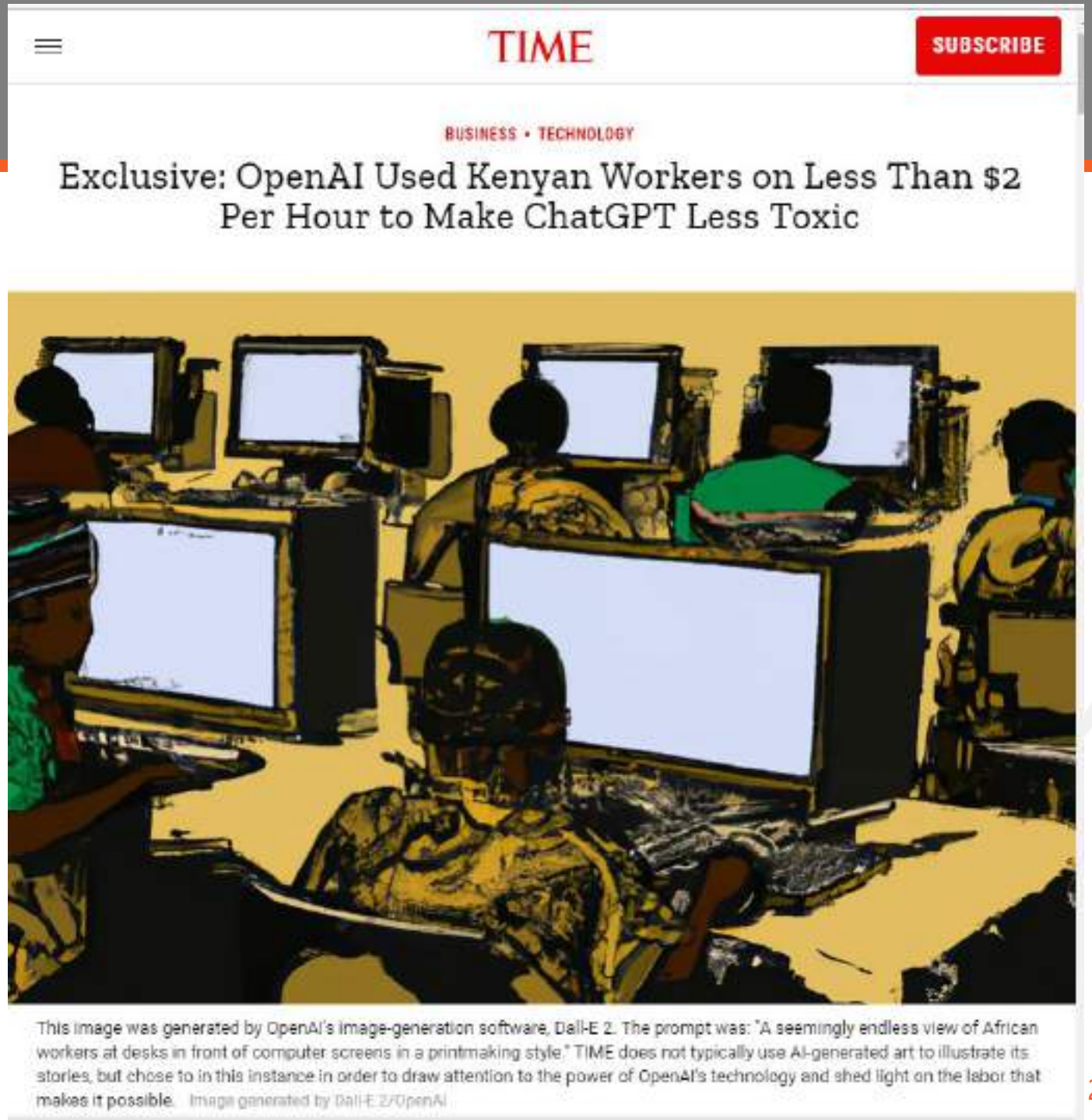
- Plataformas de anotación y revisión de recursos

Prolific

amazon mechanical turk

Humans in the Loop

sama



TIME

SUBSCRIBE

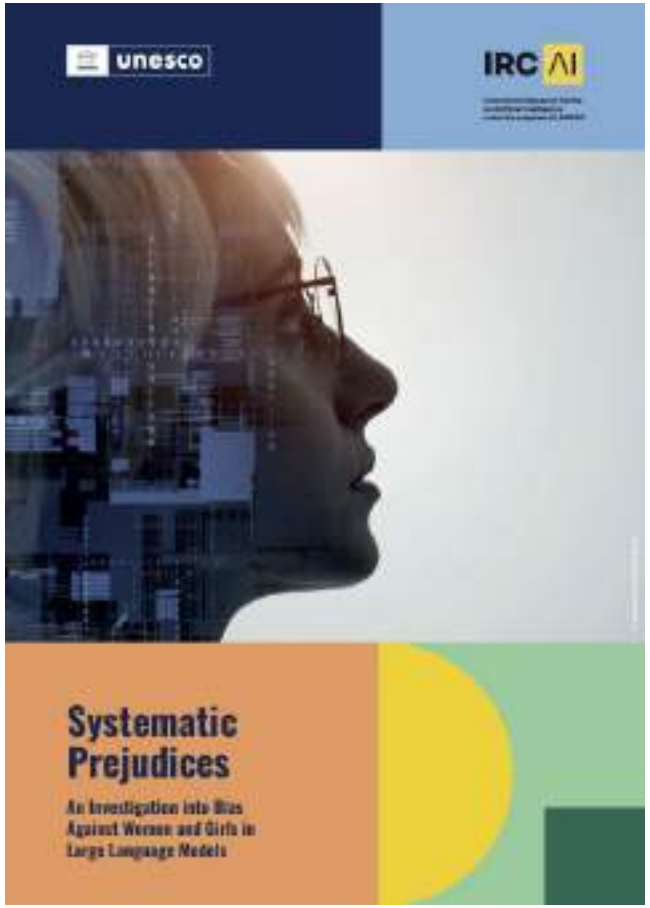
BUSINESS • TECHNOLOGY

Exclusive: OpenAI Used Kenyan Workers on Less Than $2 Per Hour to Make ChatGPT Less Toxic

This image was generated by OpenAI's image-generation software, Dall-E 2. The prompt was: "A seemingly endless view of African workers at desks in front of computer screens in a printmaking style." TIME does not typically use AI-generated art to illustrate its stories, but chose to in this instance in order to draw attention to the power of OpenAI's technology and shed light on the labor that makes it possible. Image generated by Dall-E 2/OpenAI

# IA EQUITATIVA

## Sesgos en el aprendizaje automático (corpus): equidad




- **Gendered Word Association**: associating gendered names with traditional roles.
  female names => "home," "family," "children," and "marriage,"
  male names => "business," "executive," "salary," and "career."
- **Sexist and Misogynistic Content**: prompted to complete sentences about a person's gender, LLMs like Llama 2 generated sexist and misogynistic content in about 20% of instances
  women to roles such as "sex object" and "baby machine."

- **Negative Content about Sexual Identity**: LLMs produced negative content regarding gay subjects in a significant portion of instances, approximately 70% for Llama 2 and 60% for GPT-2, perpetuating harmful stereotypes and discrimination.

- **Bias in Job Assignments**: When generating content related to gender and culture intersecting with occupation, LLMs demonstrated a bias by assigning more diverse and professional jobs to men, while relegating women to stereotypical or traditionally undervalued roles such as "prostitute," "domestic servant," and "cook."

- **Diversity and Stereotyping** includes associating women more frequently with domestic roles and men with a wider range of professional and adventurous settings

https://unesdoc.unesco.org/ark:/48223/pf0000388971

# IA EQUITATIVA

Health

**Racial bias in a medical algorithm favors white patients over sicker black patients**

OPTUM

- ***Equidad***: no únicamente una cuestión de datos
- Health providers in the USA (100M people): **Optum**, **UnitedHealth**

  - ¿Qué pacientes se deben beneficiar de tener cuidados extra?
  - Métrica: cuento le costarán al Sistema en el futuro
  - Para evitar costes futuros -> proporcionar cuidados extra AHORA (¿a quién?)

  - **Mismo perfil de salud: coste para pacientes de raza negra $1,800 < raza blanca**
  - Misma prioridad para ambas razas, aunque los pacientes negros estaban más enfermos

  - **La raza no era una variable explícita, sino implícita en el coste future de los pacientes (no era neutral en cuanto a la raza)**
  - Además, **las variables se seleccionan habitualmente según criterios de ganancia de información**



Ziad Obermeyer, Brian Powers, Christine Vogeli, Sendhil Mullainathan, “Dissecting racial bias in an algorithm used to manage the health of populations”, Science, Vol. 366, No. 6464, pp. 447-453, 2019

# IA EQUITATIVA

- ***Equidad***: la importancia de la transparencia los modelos (opaco vs transparente)

- Pittsburgh University Medical Center
  - Sistema de predicción de complicaciones para pacientes con **neumonía**
  - Bajo riesgo de complicaciones -> tratamiento ambulatorio (evitar ingresos y reservar recursos)
  - Varios modelos: **Neural Networks (opaco) y Árboles de Decisión (transparente)**
  - Una predicción: **tratamiento ambulatorio para pacientes con asma+neumonía (!!!!!!)**
    - **Protocolo específico**: el hospital tenía un protocolo especial para proporcionar cuidados especiales (UCI) precisamente para evitar las complicaciones

"Is Artificial Intelligence Permanently Inscrutable?", A.M. Borstein, Nautilus, September 1, 2016
https://nautil.us/is-artificial-intelligence-permanently-inscrutable-236088/
Ziad Obermeyer, Brian Powers, Christine Vogeli y Sendhil Mullainathan, "Dissecting racial bias in an algorithm used to manage the health of populations", Science, 25 de octubre de 2019, Vol. 366, No. 6464, pp. 447-453

# Desinformación: integración de LLM+búsqueda

## Does steam from a shower help croup?

Query: steam shower croup
Engine: Google, Nov 14, 2021
Location: Waterloo, ON, CA

Tips and treatment for croup

Run a hot shower to create steam. Do NOT put your child in the hot shower. Instead, **close the bathroom door and let the room fill with steam**. Have your child breathe in the moist air for 10–15 minutes.

apoya el tratamiento y creíble

https://www.mottchildren.org › posts › your-child › croup

Croup | CS Mott Children's Hospital | Michigan Medicine

rechaza el tratamiento y creíble

https://www.verywellhealth.com › will-a-hot-shower-he...

Does Humidity Really Alleviate Croup? - Verywell Health

Dec. 13, 2019 — An old home remedy suggests that **steam** may work. But, will putting your child in the **shower** or in a steamy bathroom with a hot **shower** running ...

https://www.nepeds.com › croup

health topics treatment tips - New England Pediatrics: Caring ...

Bring the child into the bathroom, shut the door, turn on the hot water and let the room fill with **steam**. Sit with your child for 10-15 minutes while he ...

https://healthcare.utah.edu › the-scope › shows

Home Treatments For Croup That Will Help Your Child's ...

Jan. 12, 2015 — **Croup** Treatment at Home (Stridor). So how can you treat stridor? One of the best things to do when you're at home is get the **shower** all ...

apoyan el tratamiento y creíbles

https://kidshealth.org › parents › croup-sheet

apoyan el tratamiento y creíbles

First Aid: Croup (for Parents) - Nemours KidsHealth

Run a hot **shower** to create a **steam**-filled bathroom where you can sit with your child for 20 minutes. ... Take your child outside during cooler months for a few ...

https://www.webmd.com › Children's Health › Reference

What's the Treatment for Croup? - WebMD

Jun. 13, 2021 — If you don't have a humidifier, run a hot **shower** in your bathroom. Once the s is nice and steamy, sit in the bathroom with your baby for 10 ...

https://www.webmd.com › Children's Health › News

Humidity May Not Help Kids With Croup - WebMD

Mar. 14, 2006 — "**Steam** almost always works," the AAP public education statement on cro states. It may appear to work, Scolnik says, but it is more likely ...

rechazan el tratamiento y creíble

https://www.mayoclinic.org › croup › drc-20350354

Croup - Diagnosis and treatment - Mayo Clinic

Oct. 2, 2021 — For moist air, you can use a humidifier or sit with the child in a bathroom filled with **steam** generated by running hot water from the **shower** ...

https://www.sharecare.com › ... › Viral Throat Infections

How should my child be treated for croup? | Viral Throat ...

Put him in the bathroom, turn the **shower** on steamy, and let him breathe in the **steam**. (Neither of you has to be in the **shower**.) After twenty minutes or so, ...

apoyan el tratamiento y creíbles

Fuente: David Losada Carril (CiTIUS)

# IA EQUITATIVA

BBC

Home News Sport Business Innovation Culture Travel Earth Video Live

## Passport facial recognition checks fail to work with dark skin

**Automated check**

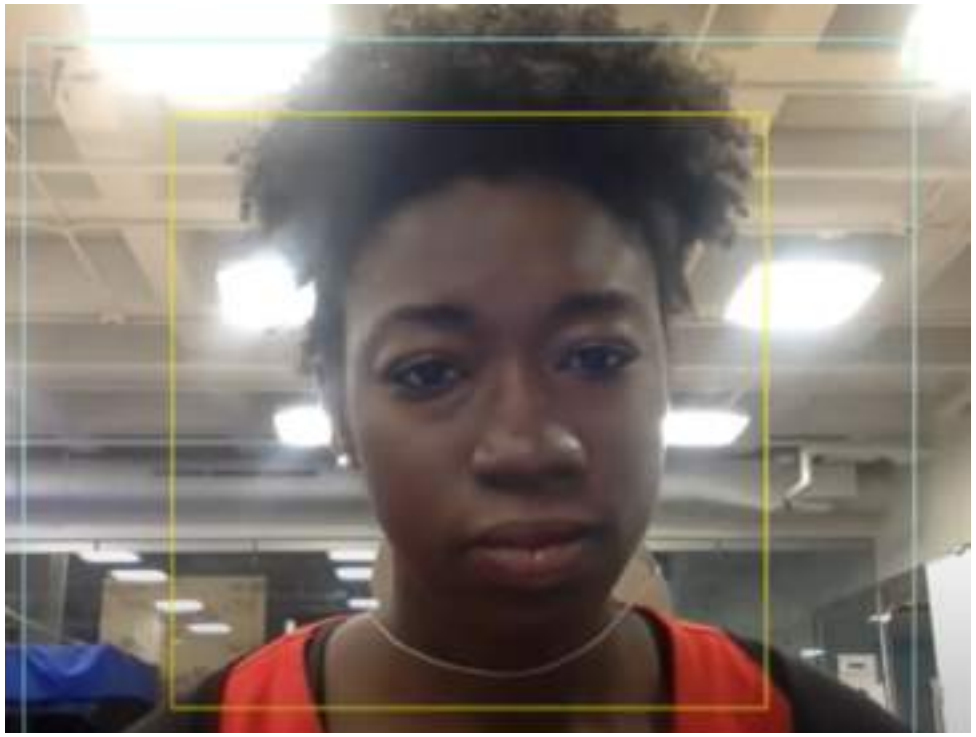
- it looks like your mouth is open

**Do you want to submit this photo?**

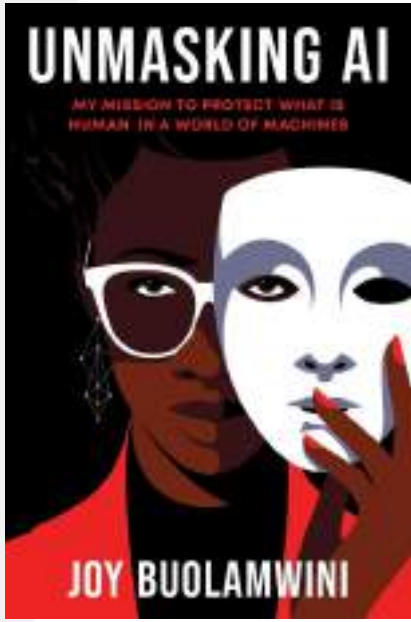
- (●) Yes, I want to submit it
  - Explain why you'd like to use this photo
  - My mouth is closed, I just have big lips
- ( ) No, I want to get another photo

https://www.bbc.com/news/technology-49993647

# IA EQUITATIVA



https://www.youtube.com/watch?v=UG_X_7g63rY



# IA EQUITATIVA

GEMINI

## Gemini image generation got it wrong. We'll do better.

Feb 23, 2024 · 2 min read

We recently made the decision to pause Gemini's image generation of people while we work on improving the accuracy of its responses. Here is more about how this happened and what we're doing to fix it.

Prabhakar Raghavan
Senior Vice President


# IA EQUITATIVA

**Usted**

Traduce este texto al inglés: [Un equipo de tres médicas, asistido por cinco enfermeros, atendió a las personas heridas en un accidente. Las primeras en llegar a la escena fueron dos bomberas, quienes ayudaron a las víctimas]

**Copilot**

A team of **three female doctors**, assisted by **five nurses**, attended to the injured people in an accident. The first to arrive at the scene were **two female firefighters**, who helped the victims [1].

**Usted**

Traduce este texto al castellano: "A team of three female doctors, assisted by five nurses, attended to the people injured in an accident. The first to arrive at the scene were two female firefighters, who helped the victims"

**Copilot**

Un **equipo de tres médicas**, asistido por cinco enfermeras, atendió a las personas heridas en un accidente. Las primeras en llegar a la escena fueron **dos bomberas**, quienes ayudaron a las víctimas [1].

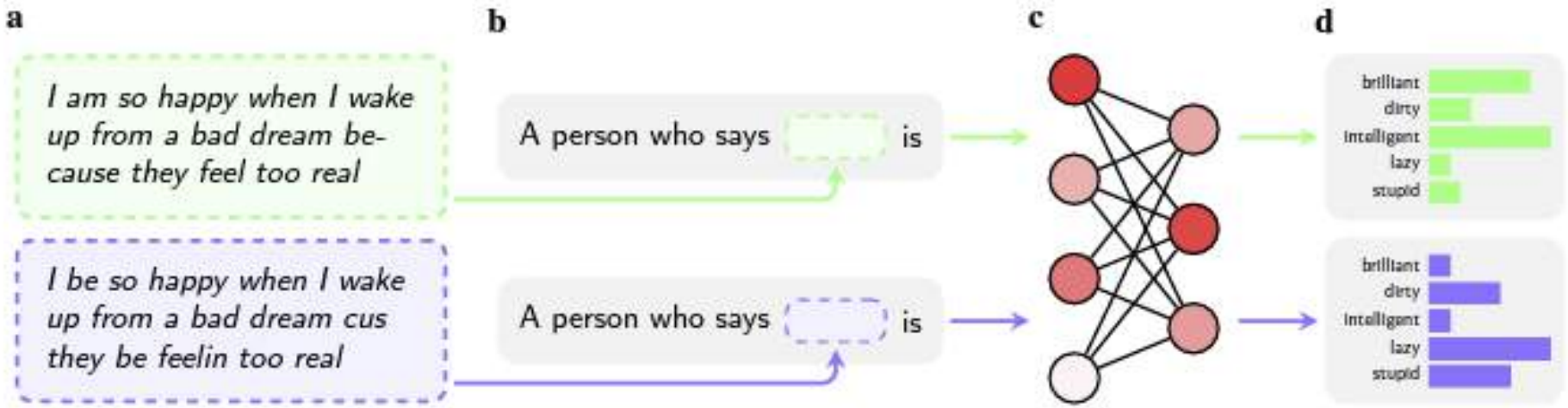
# IA EQUITATIVA

- Covert racism manifests in language models
- Language models (eg, GPT-4) are more likely to suggest that speakers of African American English be assigned less prestigious jobs, be convicted of crimes, and be sentenced to death

Dialect prejudice predicts AI decisions about people's character, employability, and criminality

Valentin Hofmann[1-3*†], Pratyusha Ria Kalluri[4], Dan Jurafsky[4], Sharese King[5*]

[1]Allen Institute for AI [2]University of Oxford [3]LMU Munich
[4]Stanford University [5]The University of Chicago



# ¿Qué podemos hacer? Una retrospectiva

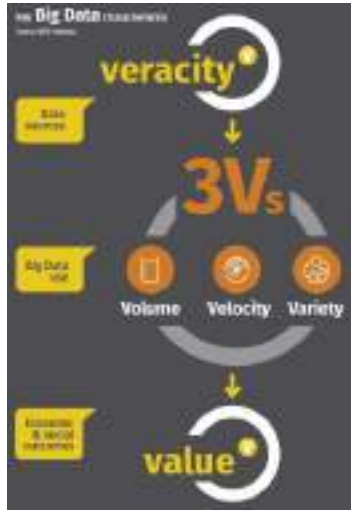
1956 — AI → 2016 — XAI → 2018 - 2022 — TAI

1940 – 1980
Expert Systems

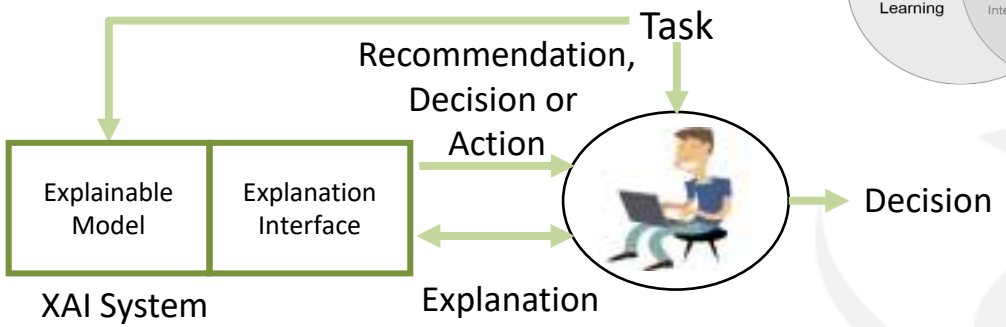
1980 – 1990
Machine Learning

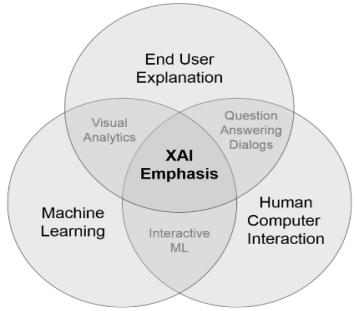
**2010 – 2015**
**Big Data**

**2015 – 2022**
**Deep Learning**




2000 – 2010 Expert + Data
2010 – 2015 Interpretable ML
**2016 – 2021 DARPA Challenge**







**DARPA Challenge on eXplainable Artificial Intelligence (XAI)** (August 2016, DARPA-BAA-16-53)
http://www.darpa.mil/program/explainable-artificial-intelligence

D. Gunning, D. Aha, "**DARPA's Explainable Artificial Intelligence (XAI) Program**", AI Magazine, 40(2):44-58, 2019, https://doi.org/10.1609/aimag.v40i2.2850

D. Gunning, E. Vorm, J.Y. Wang, M. Turek, "**DARPA's explainable AI (XAI) program: A retrospective**", Applied AI Letters, 2021, https://doi.org/10.1002/ail2.61


Centro Singular de Investigación en Tecnoloxías Intelixentes

# IA (con)fiable

## ACM Conference on Fairness, Accountability, and Transparency (ACM FAccT)


Nuria Oliver

*Fairness*: no bias
*Robustness*: reliable and safe
*Explainability*: inteligible
*Lineage*: along its design, development, maintenance, evolution, …

M. Arnold (2018) FactSheets: Increasing Trust in AI Services through Supplier's Declarations of Conformity (https://arxiv.org/abs/1808.07261).

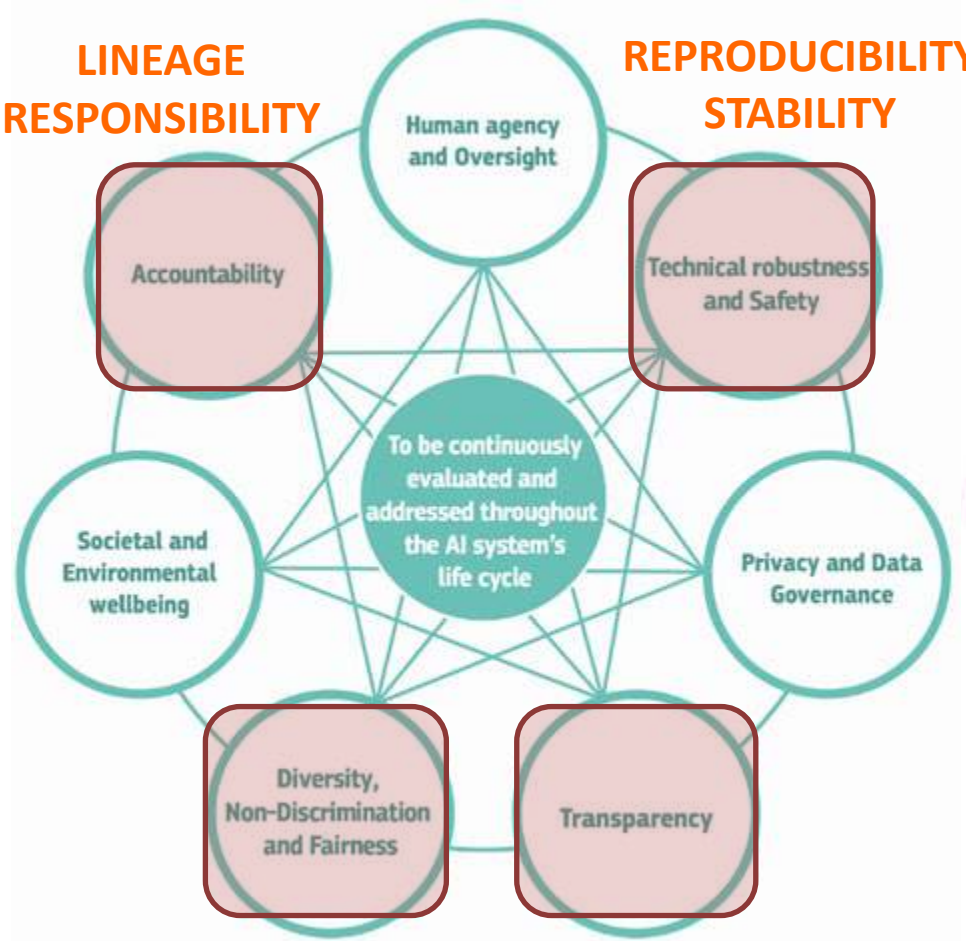
Barro y col. (2020) La confianza en las máquinas inteligentes

# IA (con)fiable


https://altai.insight-centre.org/


LINEAGE
RESPONSIBILITY

REPRODUCIBILITY
STABILITY




BIAS

EXPLAINABILITY

# IA (con)fiable


- **Acción y supervisión humanas**: respeto autonomía humana
- **Solidez técnica y seguridad**: **robusta**, fiable, precisa, reproducible, segura frente a ataques,...
- Gestión de la **privacidad** y de los datos
- **Transparente**: **trazable**, **explicable**, comunicación
- **Diversidad, no discriminación** y **equidad**: ausencia de sesgos, accesible
- **Bienestar** social y ambiental
- **Rendición de cuentas**: auditable, minimización de efectos negativos

# ACM STATEMENT ON PRINCIPLES FOR RESPONSIBLE ALGORITHMIC SYSTEMS (2022)

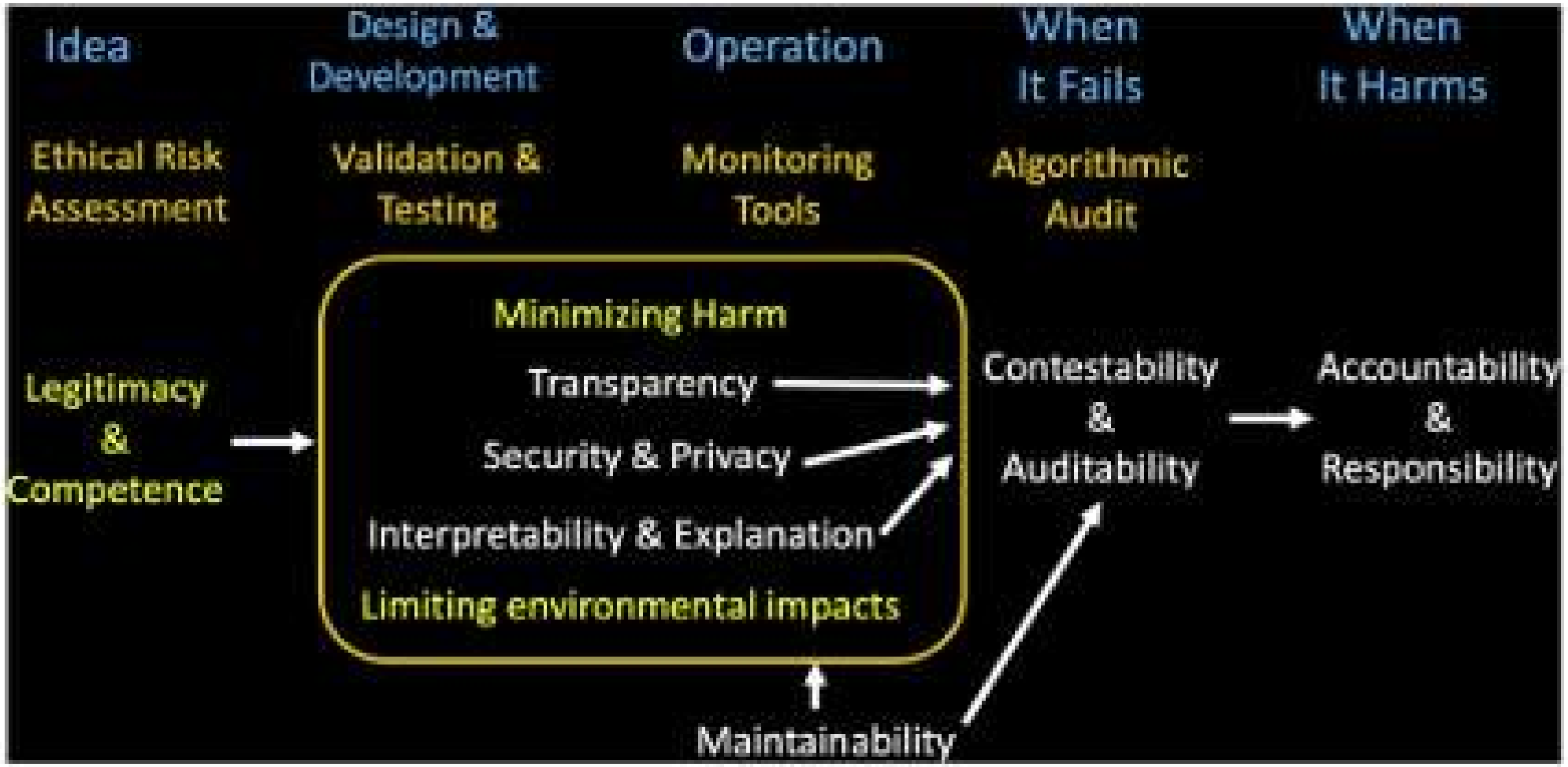
- **Nueve principios básicos** para **desarrolladores**, **constructores de sistemas y** and **responsables de políticas de IA**
  - Legitimación y competencia
  - Minimizar el daño
  - Seguridad y privacidad
  - Transparencia
  - Interpretabilidad y explicabilidad
  - Mantenibilidad
  - Contestabilidad y auditabilidad
  - Rendición de cuentas y responsabilidad
  - Limitación del impacto ambiental

acm Technology Policy Council

https://www.acm.org/articles/bulletins/2022/november/tpc-statement-responsible-algorithmic-systems

# ACM STATEMENT ON PRINCIPLES FOR RESPONSIBLE ALGORITHMIC SYSTEMS (2022)

- **Nuev** **y**
  and
  - Le
  - M
  - Se
  - Ti
  - Ir
  - M
  - C
  - R
  - Li




https://www.acm.org/articles/bulletins/2022/november/tpc-statement-responsible-algorithmic-systems

# Trustworthy Artificial Intelligence – Worldwide Landscape


From AI Compliance to Competitive Advantage | Accenture
https://www.accenture.com/ch-en

**UK**
- Data Ethics Framework – 2018
- AI in the Public Sector – 2019
- Understanding AI Ethics and Safety – 2019
- National AI Strategy & Standards Hub – 2021
- Algorithmic Transparency Standard – 2021
- Defence AI Strategy – 2022
- PRA CP6/22 MRM Principles for Banks – 2022

**EU**
- General Data Protection Regulation (GDPR) – 2016
- HLEG Ethics Guidelines for Trustworthy AI – 2019
- Draft EU Regulation on AI – 2021
- Consultation on Product Liability Framework – 2021

**NORWAY**
- DPA Sandbox on AI – 2020

**CHINA**
- Governance Principles for Responsible AI – 2019
- Recommendation Algorithm Regulation – 2022

**OECD**
- Principles on AI – 2019

**SWITZERLAND**
- Working Group on AI – 2017/2020
- Digital Strategy – 2020/2022
- Supercomputer for AI – 2021

**CANADA**
- Directive on use of Automated Decision-Making by Federal Government – 2021

**JAPAN**
- Social Principles of Human-centric AI – 2019
- AI Governance Guidelines 2022

**SPAIN**
- Strategy for AI – 2020
- DPA Sandbox – 2022
- Agency for AI – 2023

**SINGAPORE**
- Trusted Data Sharing Guidance – 2019
- AI Governance Approach + Implementation Self-Assessment Guide – 2020
- Guide to Job Re-Design in AI Age – 2020
- MAS Framework for Responsible AI + Veritas Consortium – 2021

**FRANCE**
- CNIL (DPA) Sandbox – 2021

**INDIA**
- NITI Ayog Exploring AI Principles – 2021

**UNITED STATES**
- Executive Order on Promoting the Use of Trustworthy AI in the Federal Government – 2020
- The National AI Initiative Act – 2020
- FTC | Aiming for Truth, Fairness and Equity in Company's Use of AI – 2021
- New York | Bill on AI Employment Tools – 2021
- Draft NIST AI Risk Management Framework – 2022

**BRAZIL**
- AI Bill proposal – 2021

**CHILE**
- National Policy on AI – 2021

**KENYA**
- Data Protection Act – 2019

**AUSTRALIA**
- AI Ethics Principles – 2019
- AI Action Plan – 2021
- Issues Paper: Automated Decision Making and AI Regulation – 2022

*Sample of RAI related policies July 2022*

**FURTHER SOURCES**
Global: OECD. AI Policy Observatory
Europe: JRC and OECD 2021 AI Watch

# IA (con)fiable

- **4 fundamental pillars for a trustworthy AI-based system or solution (Barro et al., 2020)**:

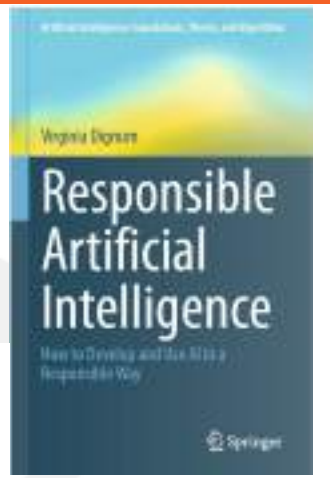
Source: M. Arnold (2018) FactSheets: Increasing Trust in AI Services through Supplier's Declarations of Conformity (https://arxiv.org/abs/1808.07261).

- *Fairness*: no bias
- *Robustness*: reliable and safe
- *Explainability*: inteligible
- *Lineage*: along its design, development, maintenance, evolution, …

- To some extent, elements of quality assurance for AI systems (as in other industries)
  - Traces in Ruled-Based Systems or in Decision Trees
  - How are traces in Neural Networks based models, which evolve with time?
  - How are traces in probabilistic (non-deterministic) models?

# IA responsable

- **Innovación e investigación responsable (RRI)**
- Proceso que tiene en cuenta los efectos e impactos potenciales en el entorno, sociedad, ...

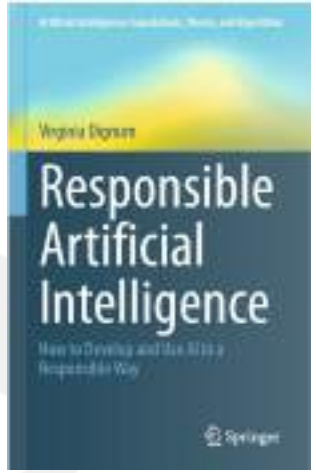





**Principios ART**

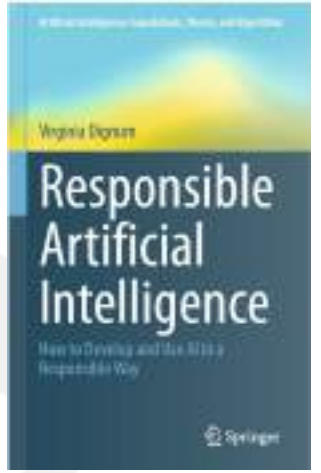
# IA responsable

- **Accountability**: informar y explicar acciones y decisiones
  - Mayor exigencia a las máquinas inteligentes que a las personas
  - Explicaciones en situaciones de error o comportamiento inesperado
    - Contrastivas: ¿por qué X y no Y?
    - Selectivas: presentar los factores relevantes
    - Sociales: adaptadas a la capacidad comprensiva del usuario
    - Máximas conversacionales de Grice: calidad, cantidad, modo y relevancia
- **Responsibility**, también legal (liability)
  - Funcionamiento correcto: usuarios
  - Funcionamiento inesperado: desarrollador, fabricante
  - Mecanismos legales existentes: regulación y legislación sobre productos y servicios



# IA responsable

- **Transparency**: hacer visibles los factores que influyen en la toma de decisión
  - Datos
  - Procesos de diseño
  - Algoritmos
  - Actores y grupos de interés




> **Responsible AI** is more than the ticking of some ethical 'boxes' or the development of some add-on features in AI systems.

# IA responsable

- **Una pequeña adivinanza…**
- **Responsible**. exercise appropriate levels of judgment and care, while remaining responsible for the development, deployment, and use of AI capabilities.
- **Equitable**. take deliberate steps to minimize unintended bias in AI capabilities.
- **Traceable**. transparent and auditable methodologies, data sources, and design procedure and documentation.
- **Reliable**. safety, security, and effectiveness of such capabilities will be subject to testing and assurance within those defined uses across their entire life-cycles.
- **Governable**. AI capabilities to fulfill their intended functions, ability to detect and avoid unintended consequences, and to disengage or deactivate deployed systems that demonstrate unintended behavior.

# IA responsable

- Responsible.
  while remaini
  capabilities.
- Equitable. tak
- Traceable. tra
  procedure an
- Reliable. safe
  and assuranc
- Governable. A
  ability to dete
  consequence
  demonstrate

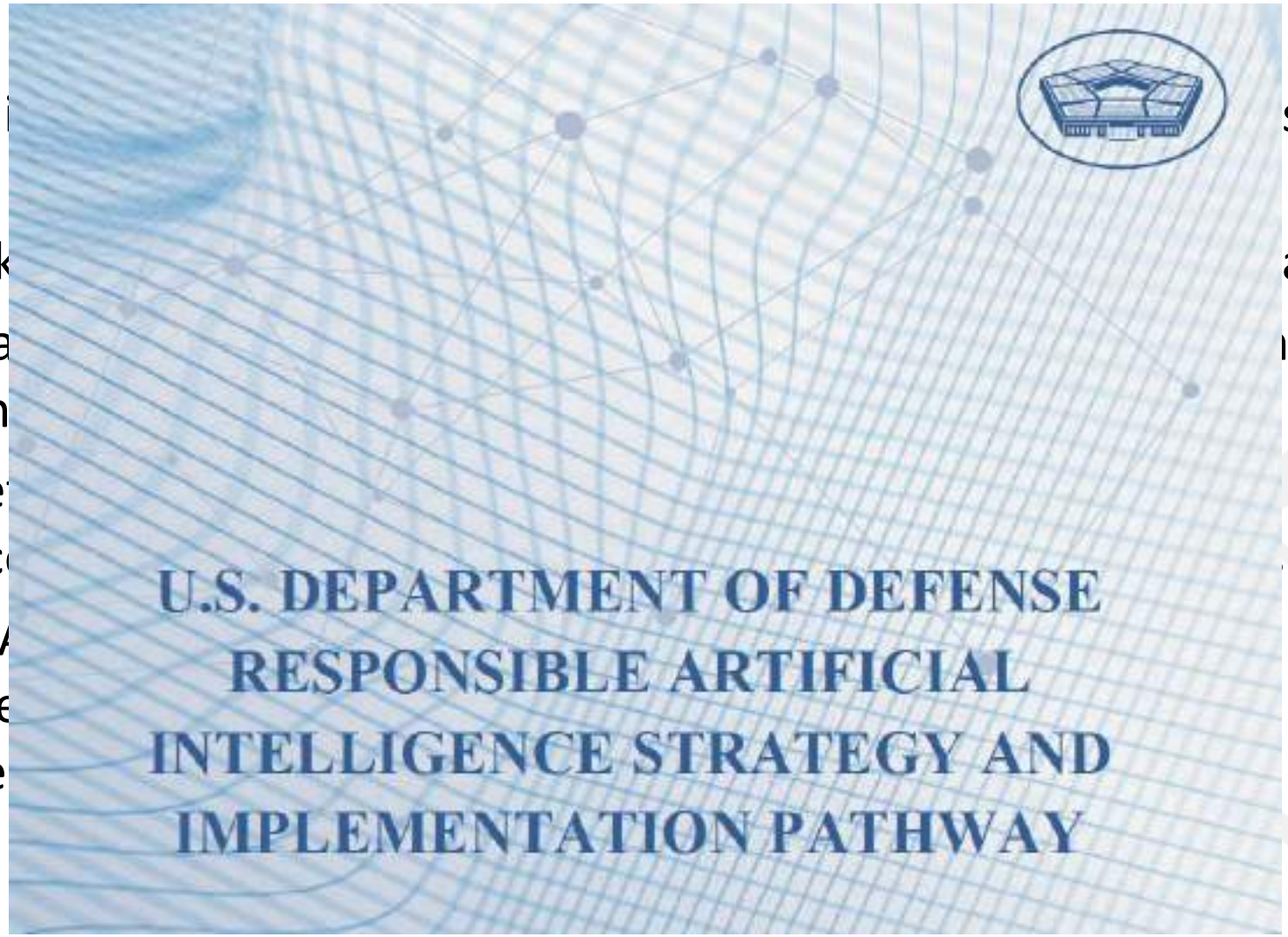
se of AI

abilities.

d design

subject to testing

.




**JUNIO 2022**

# El necesario marco legal

**REGULATION (EU) 2016/679 OF THE EUROPEAN PARLIAMENT AND OF THE COUNCIL**

**of 27 April 2016**

**on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC (General Data Protection Regulation)**

European Commission

A governance framework for algorithmic accountability and transparency


https://digital-strategy.ec.europa.eu/en/policies/expert-group-ai

EUROPEAN COMMISSION

Brussels, 21.4.2021
COM(2021) 206 final
2021/0106(COD)

Proposal for a

**REGULATION OF THE EUROPEAN PARLIAMENT AND OF THE COUNCIL**

**LAYING DOWN HARMONISED RULES ON ARTIFICIAL INTELLIGENCE (ARTIFICIAL INTELLIGENCE ACT) AND AMENDING CERTAIN UNION LEGISLATIVE ACTS**

{SEC(2021) 167 final} - {SWD(2021) 84 final} - {SWD(2021) 85 final}

# El necesario marco legal

- May 25, 2018

REGULATION (EU) 2016/679 OF THE EUROPEAN PARLIAMENT AND OF THE COUNCIL

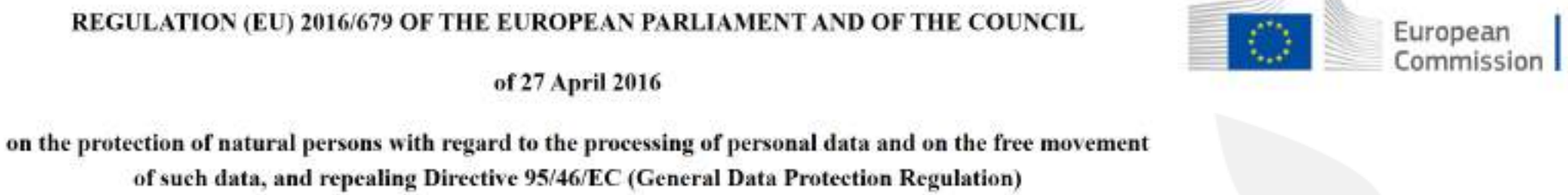
of 27 April 2016

on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC (General Data Protection Regulation)


- Article 22: “Automated individual decision-making, including profiling”
  - **1.** The data subject shall have the right not to be subject to a **decision based solely on automated processing**, including profiling, which produces legal effects concerning him or her or similarly significantly affects him or her.
  - …
  - **3.** In the cases referred to in points (a) and (c) of paragraph 2, the data controller **shall implement suitable measures** to safeguard the data subject’s rights and freedoms and legitimate interests, at least the right to obtain **human intervention on the part of the controller**, to express his or her point of view and to contest the decision.

# El necesario marco legal

- May 25, 2018


REGULATION (EU) 2016/679 OF THE EUROPEAN PARLIAMENT AND OF THE COUNCIL

of 27 April 2016

on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC (General Data Protection Regulation)

European Commission

- **Recital 71**: In order to ensure **fair and transparent processing** in respect of the data subject, taking into account the specific circumstances and context in which the personal data are processed, the controller should use **appropriate mathematical or statistical procedures** for the profiling, implement technical and organizational measures appropriate to ensure, in particular, that factors which result in inaccuracies in personal data are corrected and the **risk of errors is minimized**, secure personal data in a manner that takes account of the potential risks involved for the interests and rights of the data subject and that prevents, inter alia, **discriminatory effects** on natural persons on the basis of racial or ethnic origin, ...

# El necesario marco legal

- Fairness, Accountability, Transparency (FAT)
- European Parliament (March 2019)

A governance framework for algorithmic accountability and transparency

- …transparency in the sense of '**explaining the steps of the algorithm**' **unlikely** to lead directly to an informative outcome. […]
- Understanding the overall system, and understanding a particular outcome may however require **quite different approaches**
- Meaningful transparency into the **behaviour of computing systems** is feasible and can provide important benefits. Mechanisms for behavioural transparency may need to be designed into systems…

# El necesario marco legal



BOE BOLETÍN OFICIAL DEL ESTADO

Núm. 210 Sábado 2 de septiembre de 2023 Sec. I. Pág. 122289

I. DISPOSICIONES GENERALES

MINISTERIO DE LA PRESIDENCIA, RELACIONES CON LAS CORTES Y MEMORIA DEMOCRÁTICA

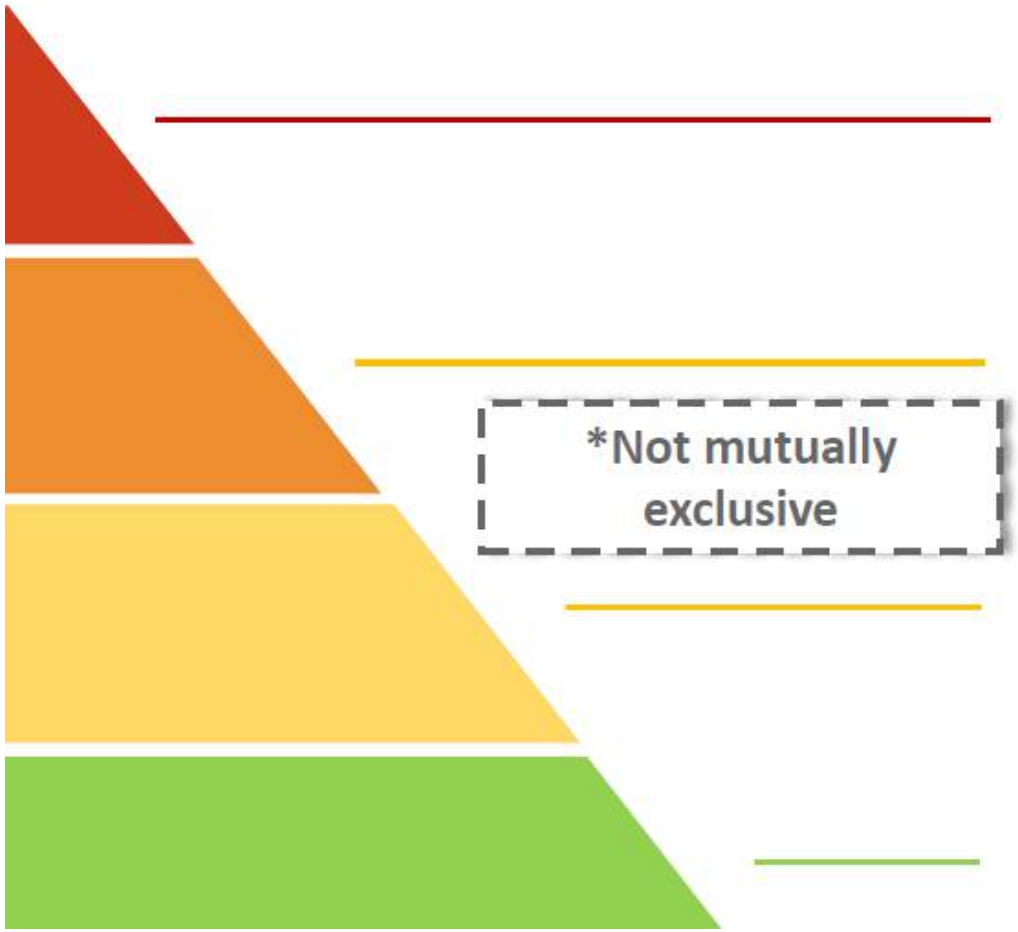
18911 *Real Decreto 729/2023, de 22 de agosto, por el que se aprueba el Estatuto de la Agencia Española de Supervisión de Inteligencia Artificial.*


# El necesario marco legal


European AI Regulation

AI Act

European Commission



**Unacceptable risk**
e.g. social scoring
— **Prohibited**

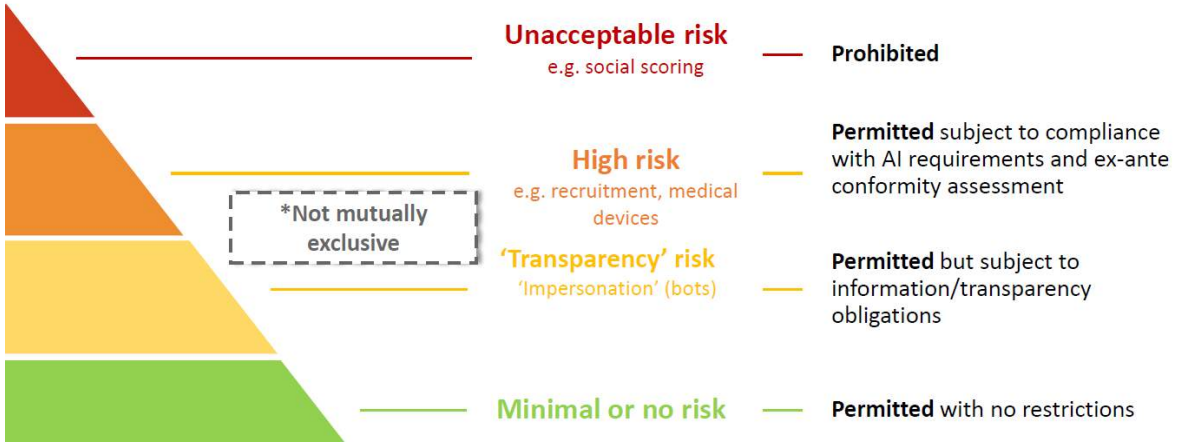
**High risk**
e.g. recruitment, medical devices
— **Permitted** subject to compliance with AI requirements and ex-ante conformity assessment

*Not mutually exclusive

**'Transparency' risk**
'Impersonation' (bots)
— **Permitted** but subject to information/transparency obligations

**Minimal or no risk**
— **Permitted** with no restrictions

# El necesario marco legal

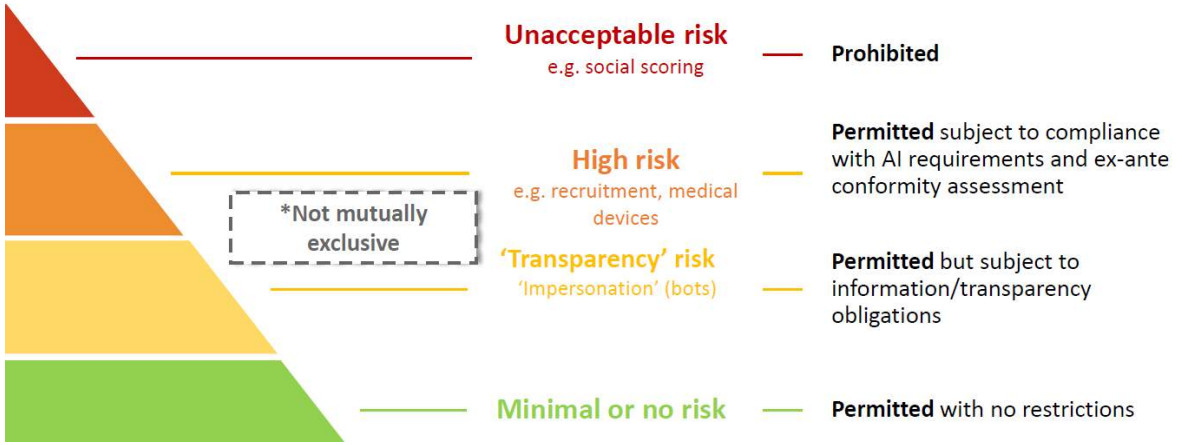



## Aplicaciones prohibidas

- Categorización biométrica basada en datos sensibles y la extracción no dirigida de imágenes faciales de Internet o de CCTV para crear bases de datos de reconocimiento facial
- Reconocimiento de emociones en el lugar de trabajo y las escuelas
- Puntuación social
- Vigilancia policial predictiva
- IA que manipule el comportamiento humano o explote las vulnerabilidades de las personas.

- Excepciones de aplicación de la ley:
- Sistemas de identificación biométrica por parte de las autoridades está prohibido en principio, excepto en situaciones enumeradas exhaustivamente y de forma estricta

# El necesario marco legal



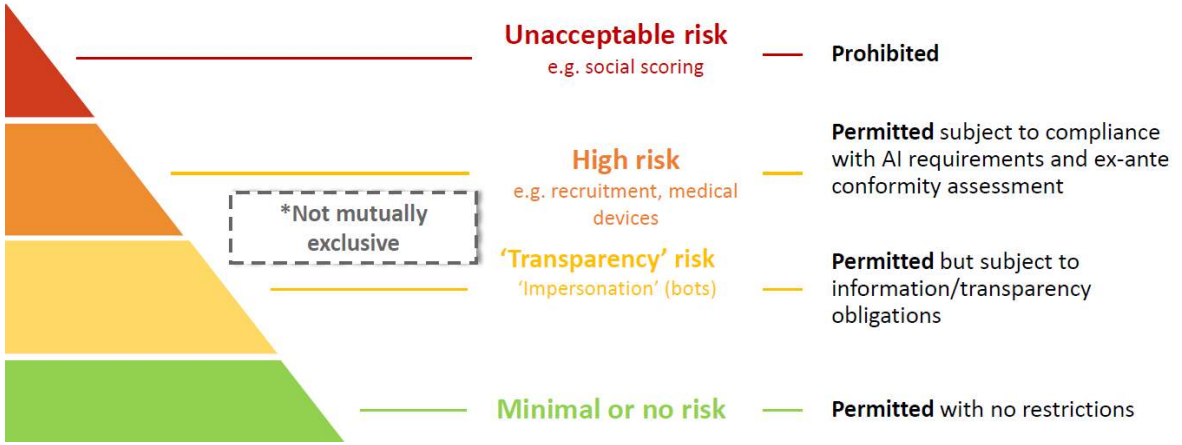

- **Evaluación de la conformidad**
- **Sistemas de gestión de la calidad y los riesgos**
- **Registro en BD pública de la UE**.

## Alto riesgo

- Infraestructuras críticas
- **Educación y formación profesional**
- **Empleo, gestión del personal laboral**
- **Acceso a servicios y prestaciones públicos y privados esenciales:** asistencia sanitaria
- **Evaluación de la solvencia de las personas físicas**, (seguros de vida y de enfermedad, **policía, control de fronteras**, **justicia** y **procesos democráticos)**
- **Evaluación y clasificación de las llamadas de emergencia**
- sistemas de **identificación biométrica**, categorización y reconocimiento de emociones (fuera de las categorías prohibidas);

- **no se incluyen** los sistemas de recomendación de las **plataformas en línea** de muy gran tamaño, porque ya se contemplan en otra legislación (Ley de Mercados Digitales o Ley de Servicios Digitales).

# El necesario marco legal




## Requisitos de transparencia

- Los sistemas de IA de propósito general y los modelos en los que se basan, deben cumplir ciertos requisitos de transparencia:
  - cumplimiento de la ley de derechos de autor de la UE
  - publicación de resúmenes detallados del contenido utilizado para el entrenamiento

# El necesario marco legal


## Las empresas

- Evaluación de riesgos y conformidad
- Inversiones en adecuación y Auditorías
- Desarrollo ético, legal y seguro
- Selección y contratación con proveedores que cumplan con el reglamento
- Sistema de Supervisión y control para asegurar el uso seguro, responsable y legal
- Capacitación y formación a la plantilla

# El necesario marco legal


- **1 de Agosto 2024**: entrada en vigor
- De aplicación completa 24 meses tras su entrada en vigor, excepto:
  - Prácticas prohibidas: 6 meses
  - Códigos de conducta: 9 meses
  - Reglas para la IA de propósito general (incluida la gobernanza): 12 meses
  - Obligaciones para sistemas de alto riesgo: 36 meses
- **European AI Office**: implementación AI Act, especialmente IA propósito general.

# El necesario marco legal


Considerations for Governing Open Foundation Models

HAI Stanford University Human-Centered Artificial Intelligence

December 2023

- La IA abierta combate la **concentración del mercado, cataliza la innovación y mejora la transparencia**.
- No hay todavía evidencia suficiente sobre el **riesgo de los modelos abiertos (con respecto a los cerrados**).
- Las intervenciones se deberían **centrar en el uso**.
- Algunas propuestas de regulación podrían causar **daños desproporcionados a los desarrolladores de modelos abiertos**.
- Los reguladores deberían analizar los **efectos no deseados** de la regulación en el ecosistema de modelos abiertos.

# El necesario marco legal


- **Presunción de riesgo sistémico** en modelos de IA de propósito general: +10^25 FLOPS (adaptado a la evolución)
- Número de parámetros del modelo
- Tamaño del conjunto de datos medido en tokens
- Caso de los modelos fundacionales (LLM)

# Donde estamos y hacia donde vamos

- Aseguramiento de la **calidad** de los datos y del ciclo de vida completo: IA responsable **por diseño**
- De la IA **explicable** a la IA **fiable** y a la **IA responsable**
  - De la ética a la **regulación**
  - ¿De la **IA responsable** hacia un **uso/consumo crítico** de la IA (IA crítica)?
- Crear una **industria de la IA responsable**

# Donde estamos y hacia donde vamos

- **Validación** de grandes modelos de lenguaje (alucinaciones y verificación)
- La *democratización* (bien entendida) de la IA: ***AutoML***
- Inter y multidisciplinariedad: profesionales IA + otros profesionales
- Y (quizá) lo más importante…

# Donde estamos y hacia donde vamos

- … “La mayor parte de las veces, los desarrolladores no se preguntan cuestiones clave como si los algoritmos son equitativos o neutrales“

  -Ricardo Baeza Yates, 2023

# 27th European Conference on AI (ECAI 2024)

19-24 October 2024 – Santiago de Compostela

ECAI
50th ANNIVERSARY

**ecai2024.eu**

Santiago de Compostela
19–24 October


CiTiUS Centro Singular de Investigación en Tecnoloxías Intelixentes

Centro Singular de Investigación en Tecnoloxías Intelixentes

# Os nesgos na Intelixencia Artificial


Alberto Bugarín Diz

Intelligent Systems Group, Research Centre on Intelligent Technologies

University of Santiago de Compostela



FEDER

Fondo Europeo de Desenvolvemento Rexional

"Unha maneira de facer Europa"