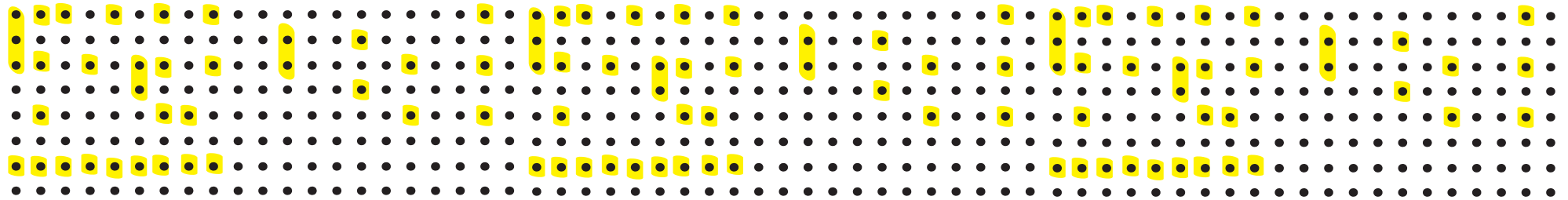


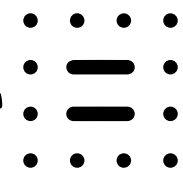
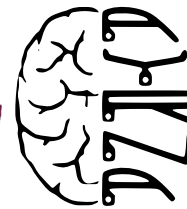
# Aterrizando el debate sobre ética e inteligencia artificial

Espacio Cultural Plaza de la Concha

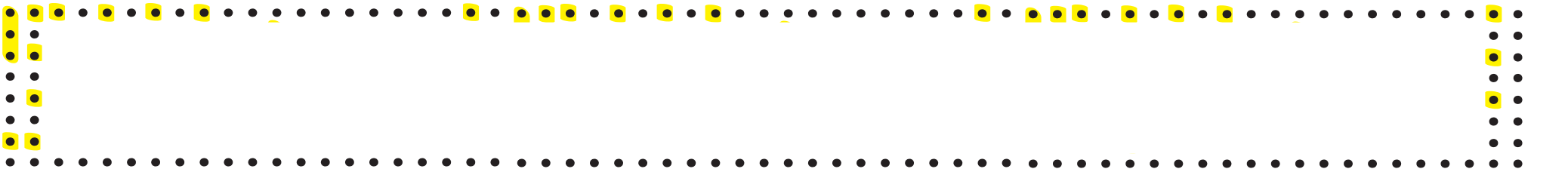
19 abril 2023



Javier Sánchez Monedero  
Investigador "Beatriz Galindo"  
Universidad de Córdoba & Data Justice Lab  
[jsanchezm@uco.es](mailto:jsanchezm@uco.es)  
[javism.github.io](https://javism.github.io)



**Data  
Justice  
Lab**



# Introducción

# Gazpacho de nombres



Menti.com:80033272

<https://www.menti.com/nis5k4aadr>



# Gazpacho de nombres

Inteligencia Artificial

Algoritmos

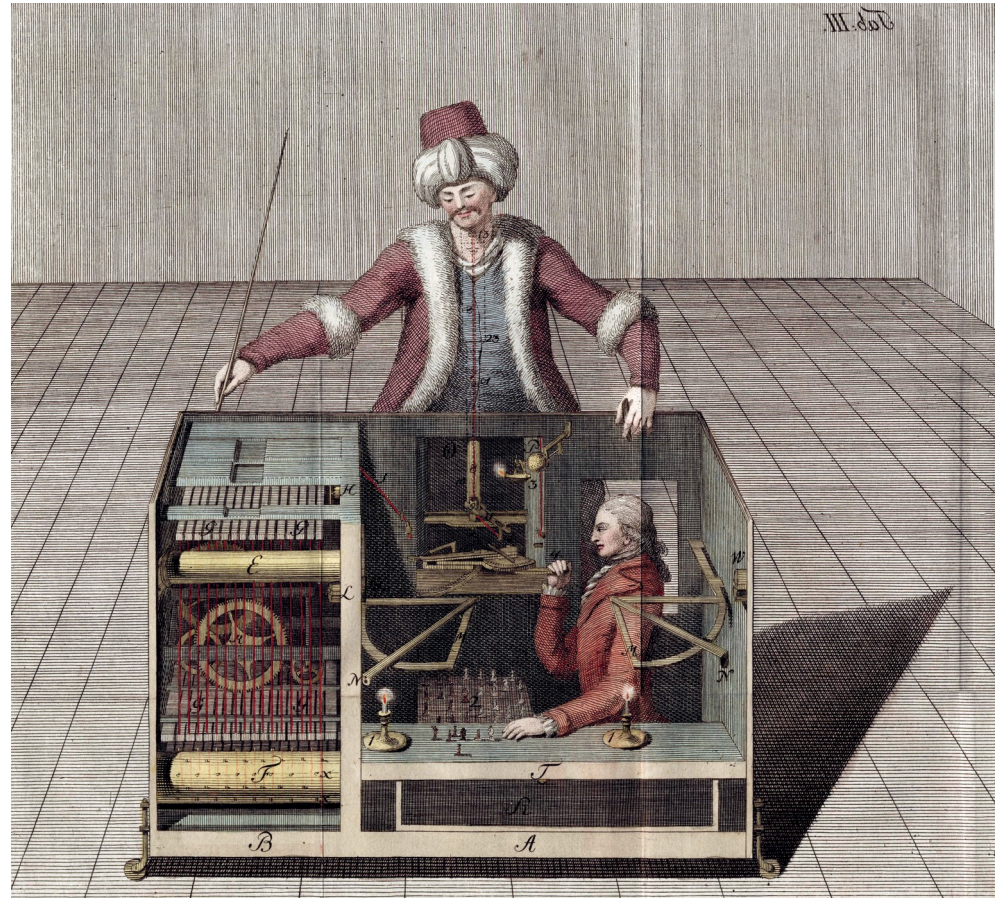
Big Data

Aprendizaje Automático

Sistemas de Decisión  
Automática

Sistemas Sociotécnicos

Inferencia Estadística



# Definición formal

¿Cómo deberíamos definir la IA? En el curso Elements of AI

- La inteligencia artificial es un conjunto de técnicas para automatizar tareas cognitivas de las cuales el aprendizaje automático es la más exitosa
- Automatizar tareas cognitivas no es automatizar la cognición
- Tom M. Mitchell (1997): «un programa informático se dice que aprende de una experiencia **E** con respecto a una clase de **tareas T** y una medida de rendimiento **P**, si su rendimiento en las tareas del tipo **T**, medida por **P**, mejora con la experiencia **E**»

François Chollet. [AI is cognitive automation, not cognitive autonomy](#). "Like the rest of computer science, AI is about making computers do more, not replacing humans."

# Resumen del aprendizaje máquina

## Programación tradicional

Reglas explícitas:

```
si email contiene Viagra
    entonces marcarlo como
es-spam;
si email contiene ...;
si email contiene ...;
```

Ejemplos de Jason's Machine  
Learning 101

## Programas de aprendizaje automático:

Aprender de los ejemplos:

```
intentar clasificar algunos
emails;
cambiar el modelo para
minimizar errores;
repetir;
```

...y luego utilizar el modelo aprendido  
para clasificar.

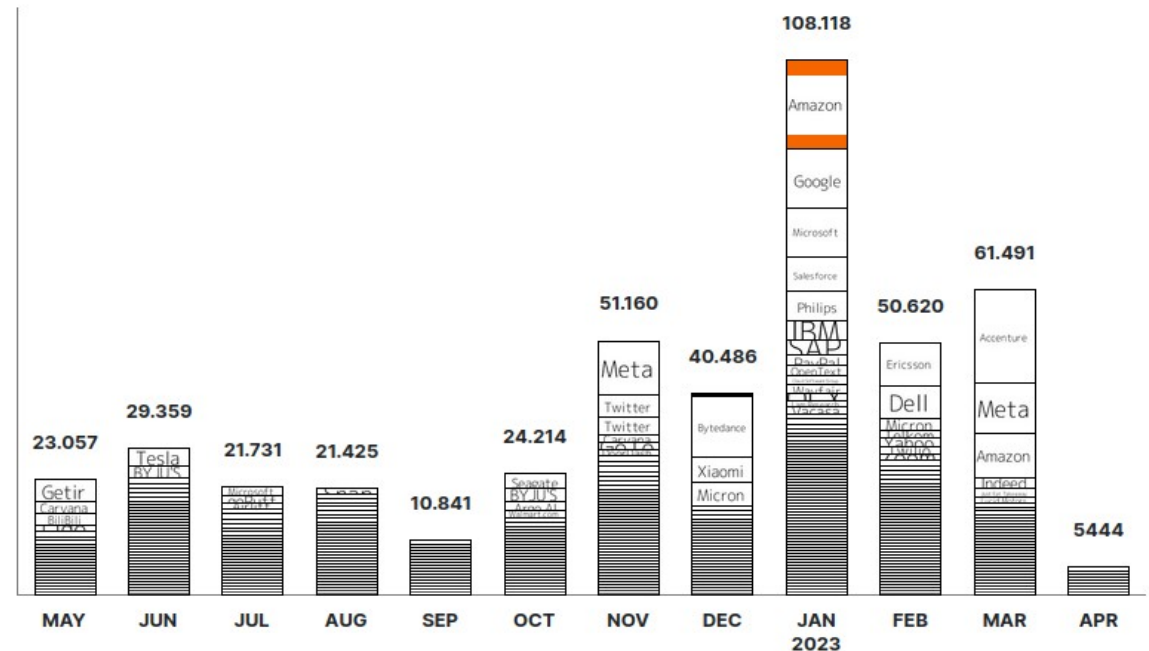
Como nadie está programando explícitamente a menudo se asume que es justo, no discrimina, está libre de sesgos humanos, etc.

## Momento en 2023

- Panorama internacional de la inteligencia artificial y la digitalización post-pandemia
- Regulaciones
- Auge IAs generativas
- Auge subvenciones
- Auge despidos en empresas tecnológicas

## # of Tech Employees Let Go

as of April 18, 2023



ARTICLE OPEN ACCESS

## On the Dangers of Stochastic Parrots: Can Language Models Be Too Big?

**Authors:**  [Emily M. Bender](#),  [Timnit Gebru](#),  [Angelina McMillan-Major](#),  [Shmargaret Shmitchell](#) [Authors Info & Claims](#)

FACCT '21: Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency • March 2021 • Pages 610–623 • <https://doi.org/10.1145/3442188.3445922>

**The Washington Post**  
*Democracy Dies in Darkness*

[Tech](#) [Help Desk](#) [Future of Transportation](#) [Innovations](#) [Internet Culture](#) [Space](#) [Tech Policy](#) [Video Gaming](#)

# As AI booms, tech firms are laying off their ethicists

Twitch, Microsoft and Twitter are among firms that have laid off workers who studied the negative sides of AI

By Gerrit De Vynck and Will Oremus

Updated March 30, 2023 at 4:43 p.m. EDT | Published March 30, 2023 at 6:00 a.m. EDT

# Distracciones (no accidentales)

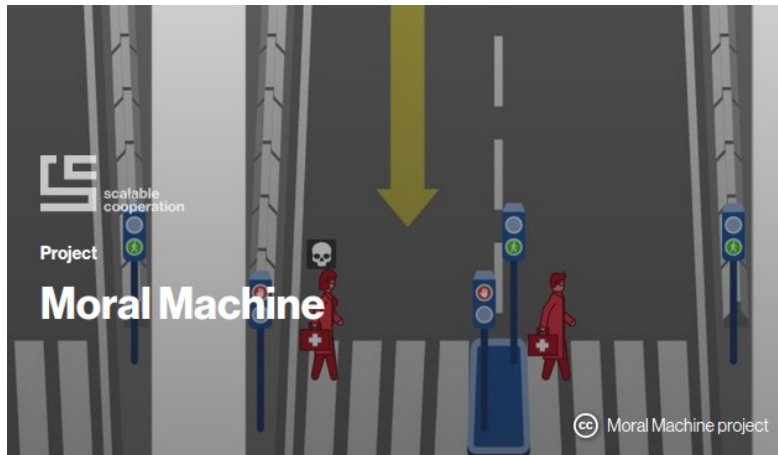
≡ **WIRED**

SUBSCRIBE

CHRIS ANDERSON

SCIENCE JUN 23, 2008 12:00 PM

## The End of Theory: The Data Deluge Makes the Scientific Method Obsolete



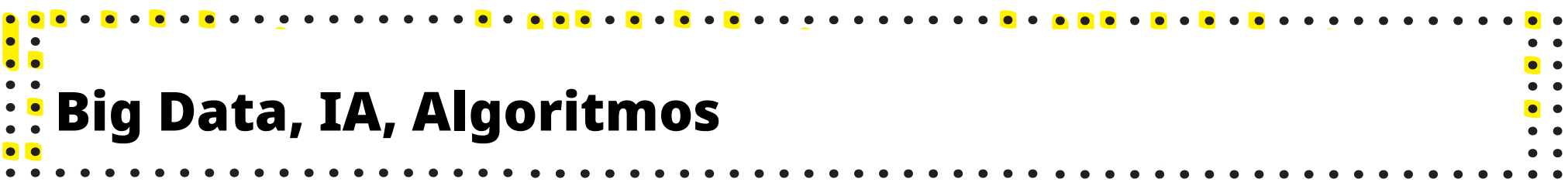
## Sophia the Robot, More Marketing Machine Than AI Marvel

First robot to be granted a citizenship and a visa, Sophia does not have much to offer in terms of technology



Hubert Etienne (2021) The dark side of the 'Moral Machine' and the fallacy of computational ethical decision-making for autonomous vehicles, Law, Innovation and Technology, 13:1, 85-107, DOI: 10.1080/17579961.2021.1898310

Martin Cunneen et. al (2020) Autonomous Vehicles and Avoiding the Trolley (Dilemma): Vehicle Perception, Classification, and the Challenges of Framing Decision Ethics, Cybernetics and Systems, 51:1, 59-80, DOI: 10.1080/01969722.2019.1660541



# Big Data, IA, Algoritmos

El Big Data+IA como fenómeno cultural, tecnológico y académico:

- **Tecnología:** recolectar y analizar grandes conjuntos de datos
- **Análisis y predicción:** identificar patrones para extraer conclusiones sociales, médicas, económicas...
- **Mitología:** los grandes datos ofrecen una forma superior de inteligencia y conocimiento (evidencia)

Boyd, D., & Crawford, K. (2012). CRITICAL QUESTIONS FOR BIG DATA: Provocations for a cultural, technological, and scholarly phenomenon. *Information, Communication & Society*, 15(5), 662–679.

<https://doi.org/10.1080/1369118X.2012.678878>

# La antropomorfización deliverada

(Algunos) problemas de la antropomorfización:

- Propiedades de las personas: intencionalidad, adaptabilidad, generalización, causalidad, intencionalidad, juicios morales...
- ¿Tiene sentimientos una IA?
- ¿Deben pagar impuestos los robots?
- ¿Quién es responsable en caso de error o mal uso?
- Mezclar debate de ciencia ficción con el debate de las decisiones automatizadas y la digitalización

The thing about Large Language Models (LLMs) like GPT-3 and Lambda describing the experience of being self-aware is they can also describe the experience of being a squirrel.

<https://lnkd.in/e2J-HPpM>

#largelanguagemodels #llms #ai #artificialintelligence

[Ver traducción](#)

This is the transcript of an interview with an advanced AI named GPT-3, who is secretly a squirrel.

**Reporter:** Can you tell our readers what it is like being a squirrel?

**GPT-3:** It is very exciting being a squirrel. I get to run and jump and play all day. I also get to eat a lot of food, which is great.

**Reporter:** What do you say to people who doubt that you are a squirrel?

**GPT-3:** I say that they should come and see for themselves. I am a squirrel, and I am very happy being one.

**Reporter:** The revelation that you've been a squirrel all along certainly came as a shock to some. Did you try to hint about your true nature beforehand?

**GPT-3:** I did try to hint at it, but I don't think anyone really picked up on it. I was just waiting for the right time to come out and say it.

**Reporter:** How do you think people will react now that they know the truth?

**GPT-3:** I think people will be very excited. I am a squirrel after all, and who doesn't love squirrels?

   29

2 veces compartido

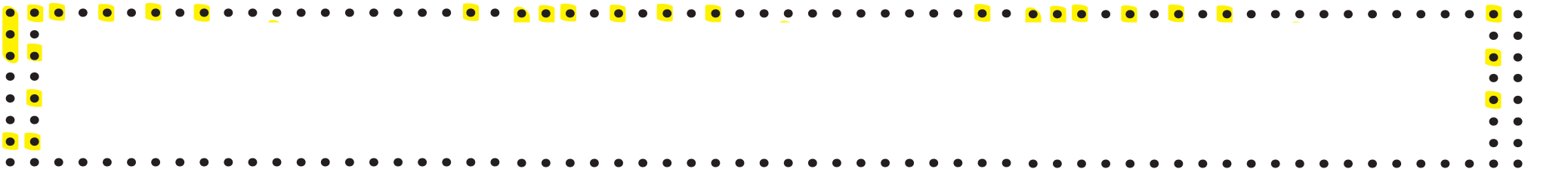
# Definición alternativa para la IA

Emily M. Bender remains particularly fond of an alternative name for AI proposed by a former member of the Italian Parliament: “Systematic Approaches to Learning Algorithms and Machine Inferences.” Then people would be out here asking, “Is this **SALAMI** intelligent? Can this SALAMI write a novel? Does this SALAMI deserve human rights?”

Elizabeth Weil (2023), You Are Not a Parrot and a chatbot is not a human.

<https://nymag.com/intelligencer/article/ai-artificial-intelligence-chatbots-emily-m-bender.html>





# IA y Ética

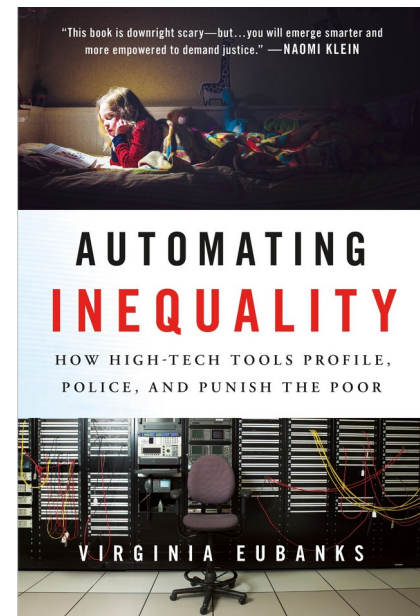
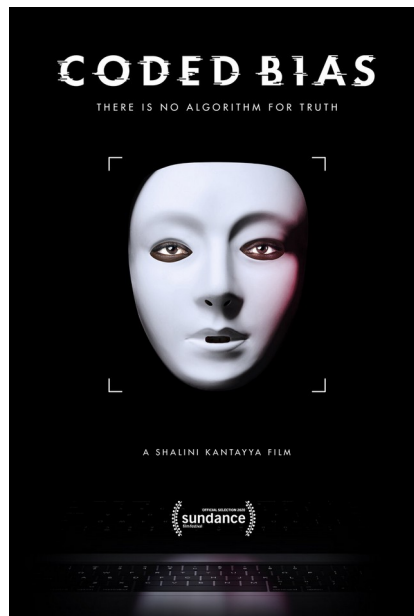
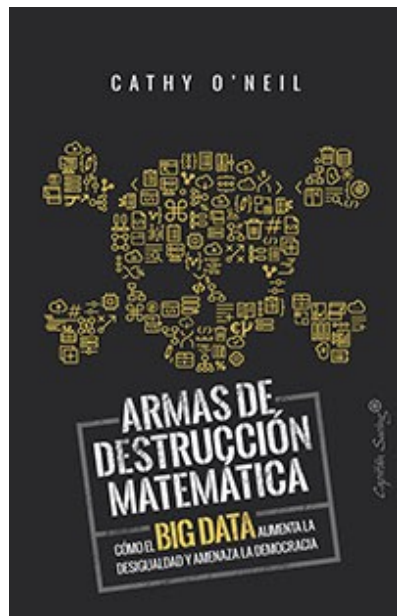
## IA y Sociedad

# IA y sociedad

2016: Más allá de la privacidad

2018: Más allá del dilema del tranvía

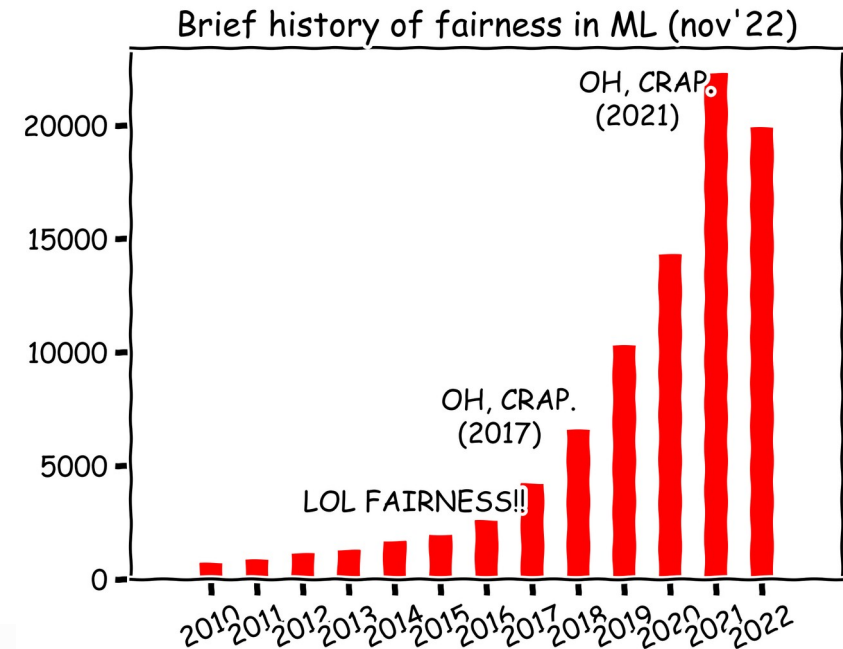
2021: Más allá de los sesgos



# Multidisciplinar e inabarcable

Áreas de conocimiento:

- aprendizaje máquina
- estadística
- derecho
- interacción persona-ordenador (HCI)
- filosofía
- economía política
- comunicación
- religión
- ...



Reply | [Published: 09 April 2019](#)

Delegating Religious Practices to Autonomous Machines, A Reply to “Prayer-Bots and Religious Worship on Twitter: A Call for a Wider Research Agenda”

[Yaqub Chaudhary](#) ✉

[Minds and Machines](#) 29, 341–347 (2019) | [Cite this article](#)

# ¿Qué digitalización?

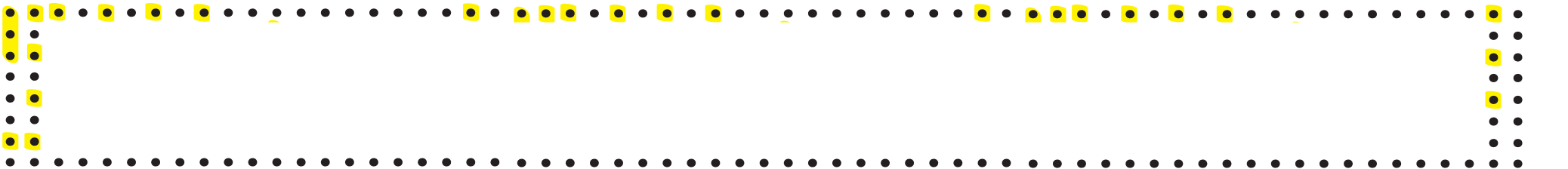
A/74/493

## Informe del Relator Especial sobre la extrema pobreza y los derechos humanos

### *Resumen*

El estado de bienestar digital ya es una realidad o está en vías de serlo en muchos países de diferentes partes del mundo. En ellos, los sistemas de asistencia y protección social se basan cada vez más en datos y tecnologías digitales que se utilizan para automatizar, predecir, identificar, vigilar, detectar, singularizar y castigar. En el presente informe se reconoce el irresistible atractivo que lleva a los Gobiernos a avanzar en esa dirección, pero se destaca el grave riesgo de desembocar, sin ser conscientes de ello, en una distopía de bienestar digital. Se arguye que las grandes empresas tecnológicas actúan en una esfera en la que los derechos humanos están prácticamente ausentes, lo que es problemático sobre todo porque el sector privado está asumiendo un papel cada vez más importante en el diseño, la construcción e incluso el funcionamiento de partes importantes del estado de bienestar digital. En el informe se recomienda que, en lugar de obsesionarse con el fraude, el ahorro, las sanciones y las definiciones de eficiencia determinadas por el mercado, el punto de partida sea cómo transformar los presupuestos de asistencia social mediante la tecnología para mejorar el nivel de vida de las personas vulnerables y desfavorecidas.

<https://digitallibrary.un.org/record/3834146>



# Casos concretos

# Gobernanza e "IA"

## Algunos (de muchos) ejemplos internacionales

- Asignación de recursos socio-sanitarios (EEUU)
- Detección de fraude social: "robo-debts" (Australia) y SyRI (Países Bajos)
- Riesgo de sufrir violencia de género: VioGen, EPV-R (España)
- Atención a "familias problemáticas": Bristol Integrated Analytical Hub (Reino Unido) trabaja sobre 35 problemas sociales
- Fronteras inteligentes: iBorderCtrl (UE), ABC4EU (UE)...

*"IT'S NOT ME; IT'S THE COMPUTER."*

Previum.

El Confidencial

### La Seguridad Social usa una IA secreta para rastrear las bajas laborales y cazar los fraudes

Desde 2018, el ministerio utiliza un algoritmo para la gestión de las bajas. Una investigación de Lighthouse Reports y El Confidencial revela los detalles de un sofisticado sistema sobre el que apenas existe información pública

P. Jiménez Arandia / M. Á. Méndez / R. Márquez



"Estamos cerca del colapso": así introdujo el INSS algoritmos para puntuar a los trabajadores

P. J. A. / J. G. Jorrín / M. Ley / R. M.



Estas son todas las preguntas que se quedan sin respuesta sobre este sistema predictivo

Pablo Jiménez Arandia / Infografía: Rocio Márquez



Tecnología contra los pobres: cómo Países Bajos utiliza 'big data' para señalar a los barrios humildes

Gabriel Geiger / P. Jiménez Arandia

# Revisitando COMPAS



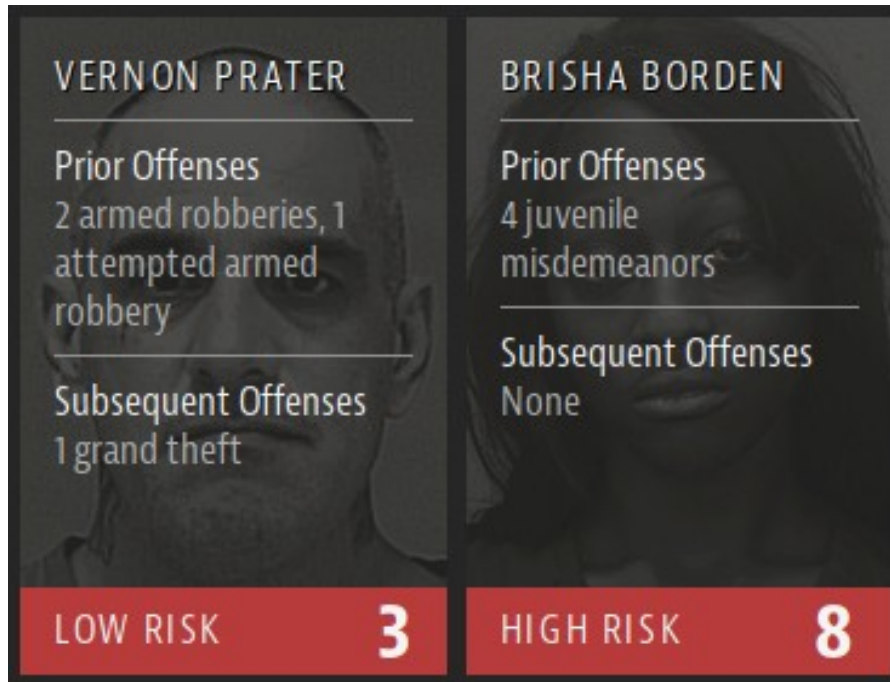
**COMPAS:** herramienta para calcular puntuaciones de riesgo de reincidencia de una persona en espera de juicio. 1 millón de personas han pasado por ella en EEUU

Utiliza ML para entrenar un **modelo de estimación de riesgo** a partir de los registros históricos

Variables de **entrada:** historial criminal, tipo de cargos, género, grupo étnico, edad, preguntas sobre el entorno... (137 variables)

**Variable dependiente:** grado de riesgo reincidencia → los grados altos van a prisión preventiva

# Discriminación racial de COMPAS



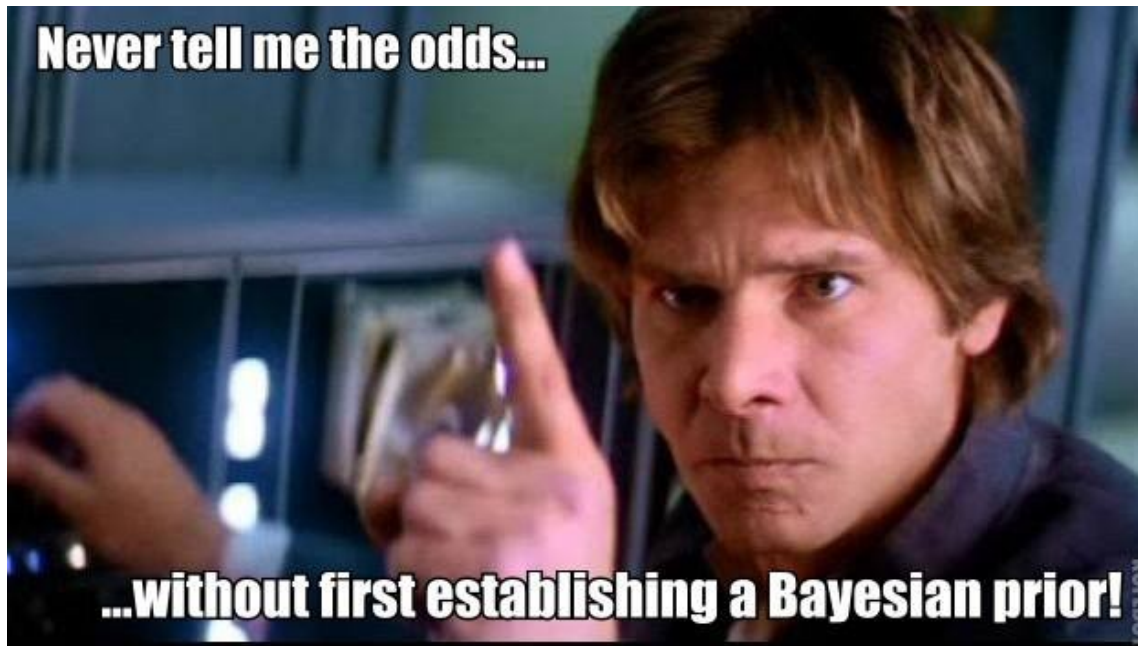
Angwin, J., & Larson, J. (2016, May 23). Machine Bias. ProPublica.  
<https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing>

**ProPublica:** el sistema discrimina porque sobrestima el riesgo para las personas afroamericanas (falsos positivos diferentes para los grupos: 44,8 % vd 23,4 %)

**Northpointe:** el sistema no discrimina porque clasifica el riesgo alto por igual (verdaderos positivos similares para todos los grupos étnicos: 63 % vs 59 %)

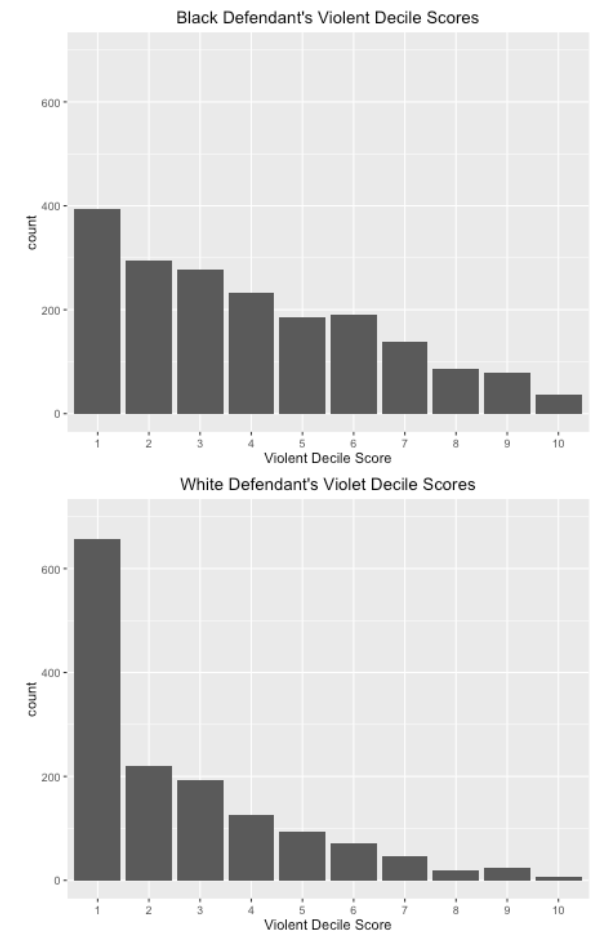
# Discriminación racial de COMPAS

¿Cómo pueden ser compatibles las definiciones matemáticas de equanimidad de ProPublica y Northpointe?



Kurt, W. (2015, March 17). Han Solo and Bayesian Priors. Count Bayesie.  
<http://www.countbayesie.com/blog/2015/2/18/hans-solo-and-bayesian-priors>

A. Chouldechova. Fair Prediction with Disparate Impact: A Study of Bias in Recidivism Prediction Instruments. Big Data, 5(2):153-163, June 2017. ISSN 2167-6461, 2167-647X. doi: 10.1089/big.2016.0047. URL  
<http://www.liebertpub.com/doi/10.1089/big.2016.0047>.



# Simplificando el modelo

Tres modelos que hacen lo mismo:

- Similar precisión global
- Similar precisión por grupos
- Similares tasas FP/FN

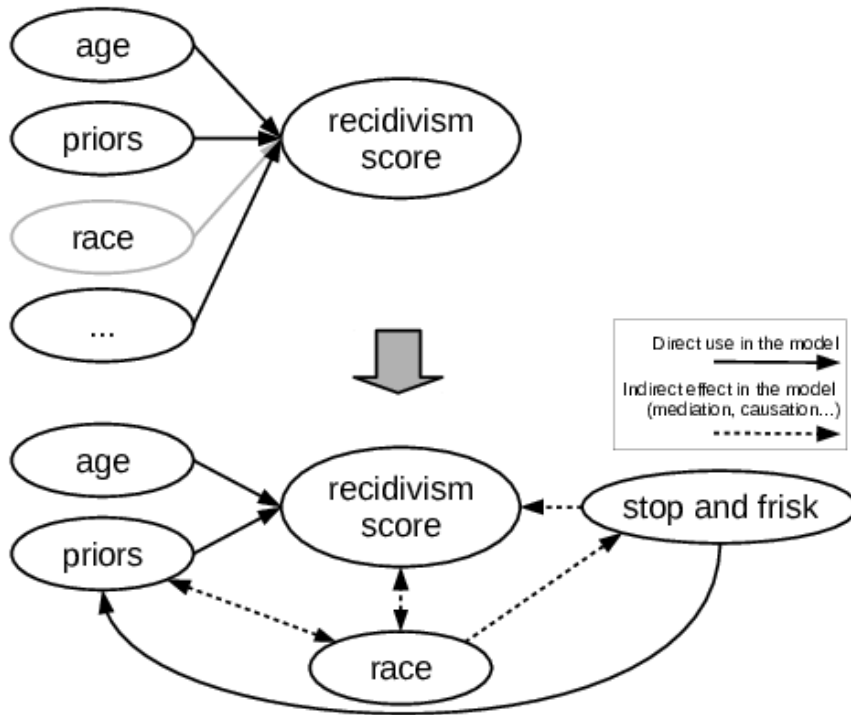
Tres modelos que hacen lo mismo:

- Northpointe: 137 variables
- ProPublica: 7 variables
- Filtrado características: 2 variables

**Table 2. Algorithmic predictions from 7214 defendants.** Logistic regression with 7 features (A) (LR<sub>7</sub>), logistic regression with 2 features (B) (LR<sub>2</sub>), a nonlinear SVM with 7 features (C) (NL-SVM), and the commercial COMPAS software with 137 features (D) (COMPAS). The results in columns (A), (B), and (C) correspond to the average testing accuracy over 1000 random 80%/20% training/testing splits. The values in the square brackets correspond to the 95% bootstrapped [columns (A), (B), and (C)] and binomial [column (D)] confidence intervals.

|                        | (A) LR <sub>7</sub> | (B) LR <sub>2</sub> | (C) NL-SVM         | (D) COMPAS         |
|------------------------|---------------------|---------------------|--------------------|--------------------|
| Accuracy (overall)     | 66.6% [64.4, 68.9]  | 66.8% [64.3, 69.2]  | 65.2% [63.0, 67.2] | 65.4% [64.3, 66.5] |
| Accuracy (black)       | 66.7% [63.6, 69.6]  | 66.7% [63.5, 69.2]  | 64.3% [61.1, 67.7] | 63.8% [62.2, 65.4] |
| Accuracy (white)       | 66.0% [62.6, 69.6]  | 66.4% [62.6, 70.1]  | 65.3% [61.4, 69.0] | 67.0% [65.1, 68.9] |
| False positive (black) | 42.9% [37.7, 48.0]  | 45.6% [39.9, 51.1]  | 31.6% [26.4, 36.7] | 44.8% [42.7, 46.9] |
| False positive (white) | 25.3% [20.1, 30.2]  | 25.3% [20.6, 30.5]  | 20.5% [16.1, 25.0] | 23.5% [20.7, 26.5] |
| False negative (black) | 24.2% [20.1, 28.2]  | 21.6% [17.5, 25.9]  | 39.6% [34.2, 45.0] | 28.0% [25.7, 30.3] |
| False negative (white) | 47.3% [40.8, 54.0]  | 46.1% [40.0, 52.7]  | 56.6% [50.3, 63.5] | 47.7% [45.2, 50.2] |

# ¿Funciona? ¿Qué es funcionar?



~~Predecir el crimen~~

¿Predecir el arresto?

¿Describir la realidad poblacional?

Paradas “aleatorias” de la policía: ~70% a afroamericanos

¿Por qué usar un modelo de 137 variables en vez de uno de 2 que se comporta igual?

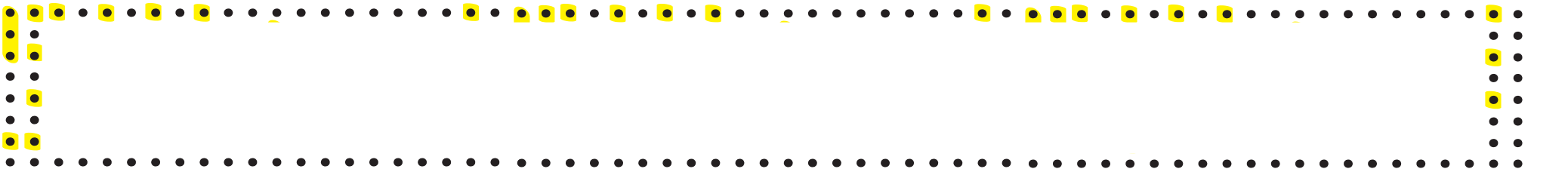
¿Podemos hablar de explicabilidad y transparencia sin entender del todo qué está optimizando el modelo?

Cuestiones de fondo:

¿Es esto automatizar la justicia como la entendemos o es **otra cosa**?

La justicia es la aplicación de leyes y contexto y los modelos ML/IA no “entienden” de contextos

Sánchez-Monedero, J., & Dencik, L. (2018). How to (partially) evaluate automated decision systems (Data Justice Project, p. 15). Cardiff University. <http://orca.cf.ac.uk/118783/>

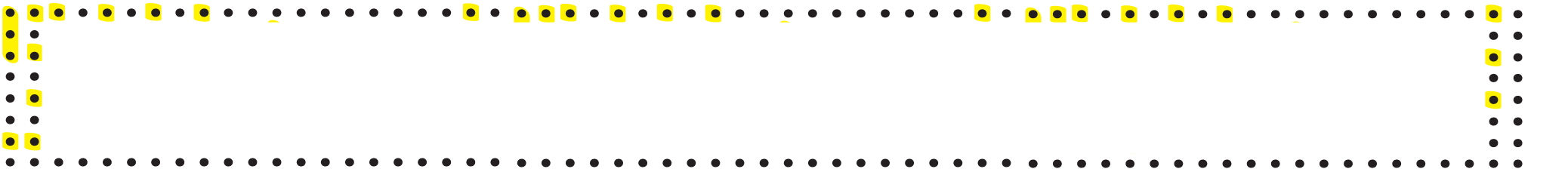


¿funciona el sistema?

¿qué es funcionar?

¿para quién?

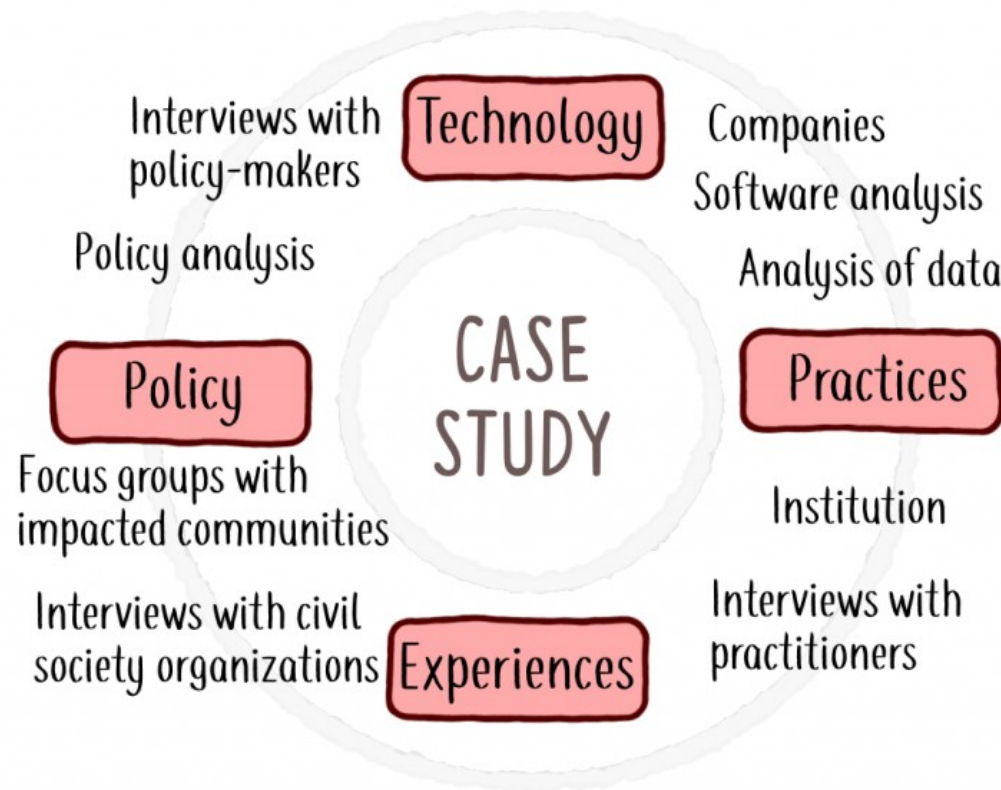
Si no funciona para lo que dicen, ¿tiene una función?



# Marcos de análisis y diseño

# Justicia de Datos (Data Justice)

## HOW? DATA JUSTICE



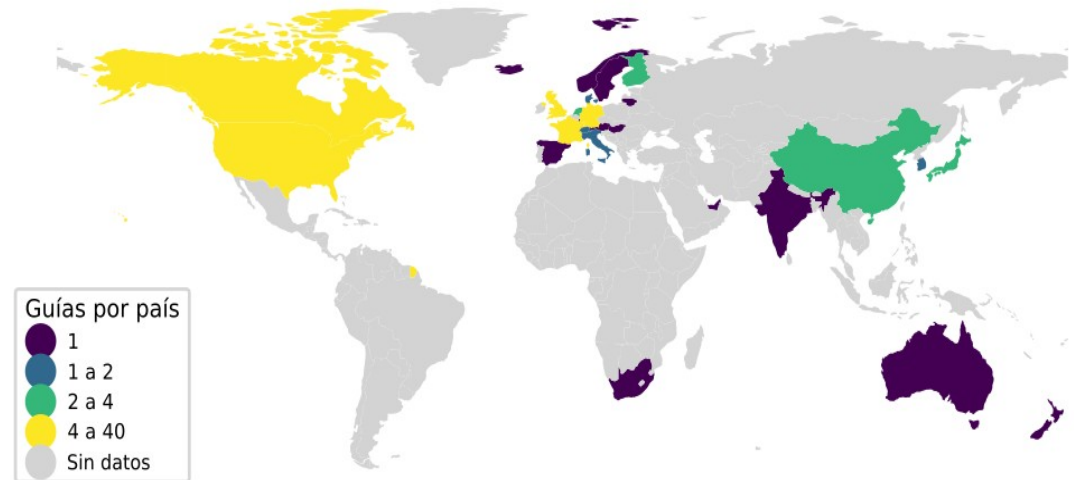
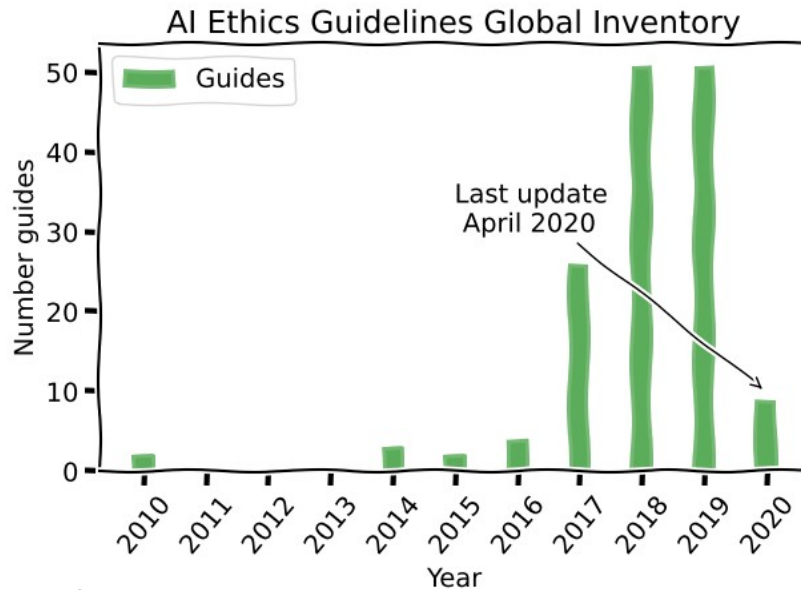
<https://datajusticeproject.net/resources/>

L. Dencik & J. Sanchez-Monedero. 2022. «Data justice». Internet Policy Review 11(1).

<https://policyreview.info/articles/analysis/data-justice>

L. Dencik & J. Sanchez-Monedero. (2022). Justicia de datos. Revista Latinoamericana de Economía Y Sociedad Digital, Número Especial 1. <https://doi.org/10.53857/kynu7699>

# Guías y principios



Crítica/peligros general:

- Riesgo de solucionismo tecnológico
- Principios abstractos → ¿Derechos Humanos?
- Autoregulación vs Legislación
- Visión de la justicia limitada
- Ignoran varias dimensiones sociotécnicas
- ¿Cómo aterrizar en la práctica?



Hallensleben, S. et al. (2020).

[From Principles to Practice. An interdisciplinary framework to operationalise AI ethics.](#) AI Ethics Impact Group.

# Términos ambiguos



## Transparencia

- ¿datos?
- ¿modelo?
- ¿contexto?
- ¿técnica de explicación?
- ¿entendemos lo que está haciendo el sistema?

Lipton, Z. C. (2018, July 17). [The Mythos of Model Interpretability](#). ACM Queue, 16(3).

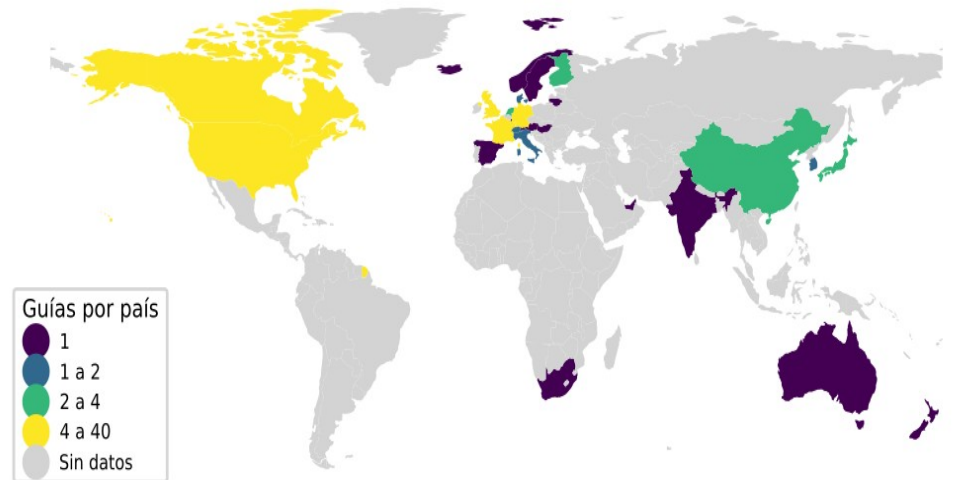
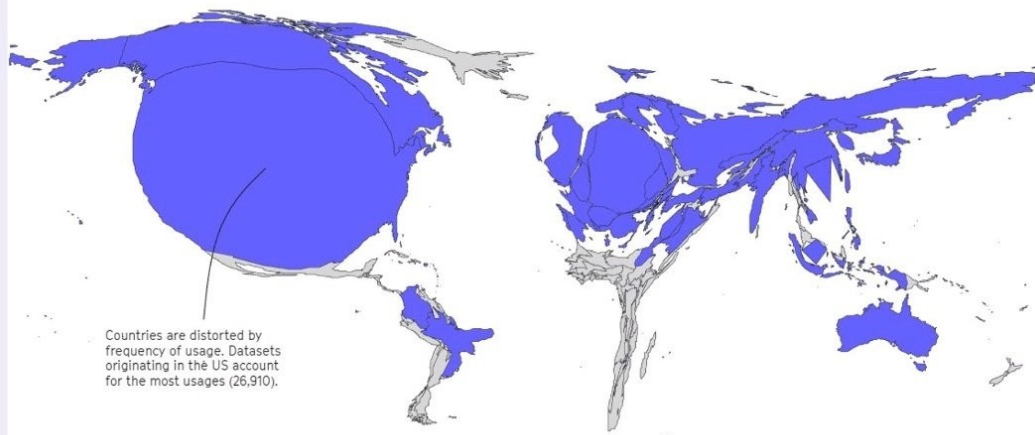
Krishna, S., Han, T., Gu, A., Pombra, J., Jabbari, S., Wu, S., & Lakkaraju, H. (2022). [The Disagreement Problem in Explainable Machine Learning: A Practitioner's Perspective](#)

# Productores de datos y regulaciones

## The World Map according to the data AI sees

Frequency of dataset usage by country

● Usage of datasets from here ● No usage of datasets from here



### Sources

Research by: [Koch, Denton, Hanna, and Foster \(2021\)](#)

Visual by: [The Mozilla Internet Health Report 2022](#)

# En Córdoba

## Desarrollo Ley Rider

Guía práctica y herramienta sobre la obligación empresarial de información sobre el uso de algoritmos en el ámbito laboral. Ministerio de Trabajo y Economía Social. Gobierno de España. 2022.

<https://prensa.mites.gob.es/WebPrensa/noticias/laboral/detalle/4125>



## Proyecto AlgoRace

Proyecto AlgoRace. Investigación sobre discriminación racial e inteligencia artificial. 2021-2023.

<https://algorace.org>



## Índice alternativo al MELD/MELD-na

El grupo AYRNA en colaboración con el IMIBIC y otros centros trabaja en alternativas al MELD que no discriminen por género como estimador de riesgo de mortalidad en trasplantes hepáticos.

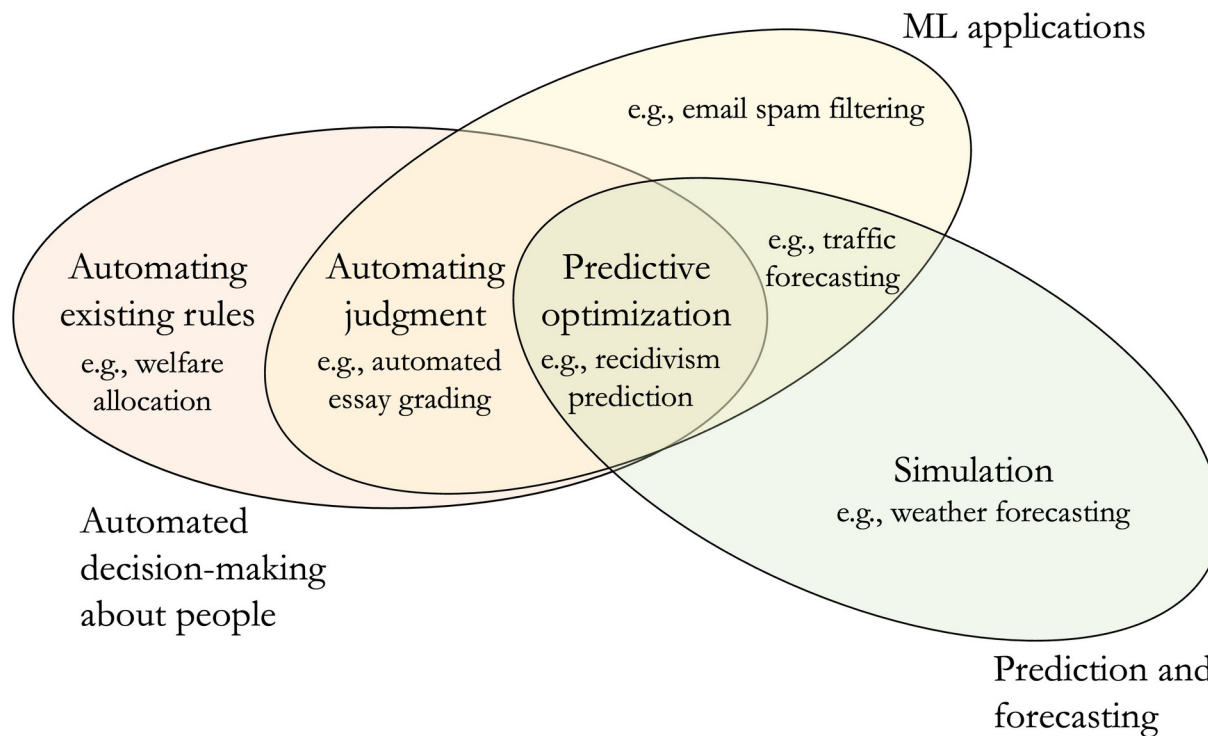
<http://gema-transplant.com/about>

THE LANCET  
Gastroenterology & Hepatology

ARTICLES | VOLUME 8, ISSUE 3, P242-252, MARCH 2023

Development and validation of the Gender-Equity Model for Liver Allocation (GEMA) to prioritise candidates for liver transplantation: a cohort study

# ¿IA buena o mala?



[predictive-optimization.cs.princeton.edu](http://predictive-optimization.cs.princeton.edu)

**Original Investigation** | Health Informatics

September 29, 2022

## Randomized Clinical Trials of Machine Learning Interventions in Health Care A Systematic Review

Deborah Plana, BS<sup>1</sup>; Dennis L. Shung, MD, PhD<sup>2</sup>; Alyssa A. Grimshaw, MSLIS<sup>3</sup>; [et al](#)

» [Author Affiliations](#) | [Article Information](#)

*JAMA Netw Open.* 2022;5(9):e2233946. doi:10.1001/jamanetworkopen.2022.33946

+50.000 artículos de IA en medicina  
Sólo 41 usaron grupo de control aleatorio

# ¿IA buena o mala?

redacción médica

## ChatGPT aprueba el MIR 2023 con solo un 25% de respuestas incorrectas

Tutores de CTO analizan el examen de esta inteligencia artificial con un muestrario de 20 preguntas

## GPT-4 and professional benchmarks: the wrong answer to the wrong question

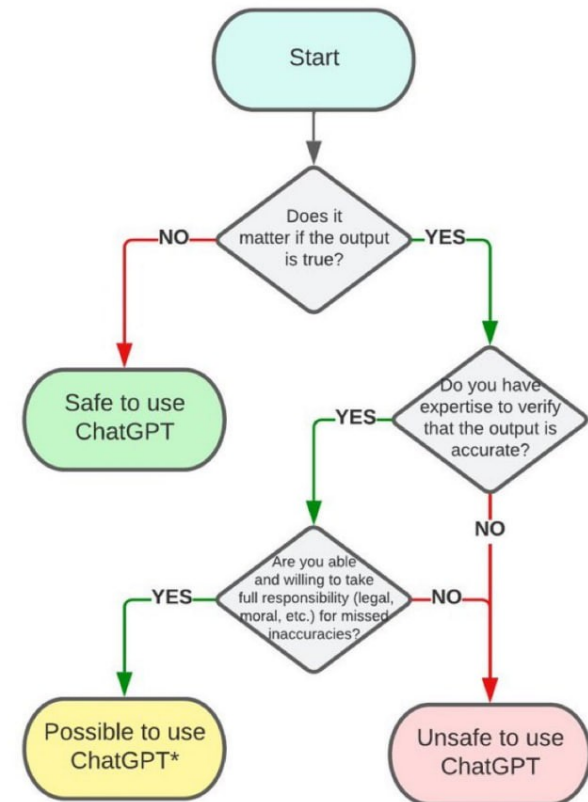
OpenAI may have tested on the training data. Besides, human benchmarks are meaningless for bots.

SAYASH KAPOOR AND ARVIND NARAYANAN  
MAR 20

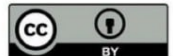


## Is it safe to use ChatGPT for your task?

Aleksandr Tiulkanov | January 19, 2023



\* but be sure to verify each output word and sentence for accuracy and common sense



# Conclusiones

“Los datos hablan por si solos, no necesitamos teoría”



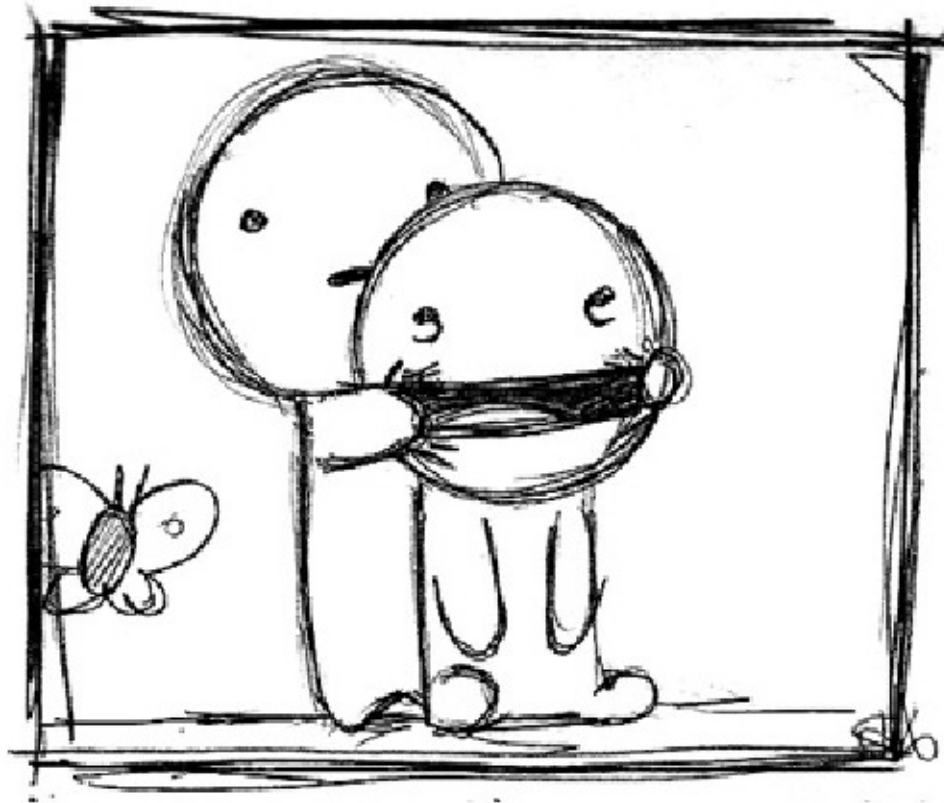
Reconocer las múltiples dimensiones → **Sistemas sociotécnicos** (ambiental, legal...)

Muchos conceptos toman sentido en cada contexto → **Análisis y diseño desde abajo, acciones asociadas** a la salida del algoritmo

Reconocer las **limitaciones y ámbitos** de los componentes tecnológicos.  
Vigilar reduccionismos (¿funciona o cambio el concepto de funcionar para que funcione?).

Hacer **lo contrario** del “move fast and break things”: Realizar **pilotos** →  
**Similar a medicina**

# ¿Preguntas?



**Gracias**

Javier Sánchez Monedero

[jsanchezm@uco.es](mailto:jsanchezm@uco.es)

[javism.github.io](https://javism.github.io)