

# SISTEMAS DE DECISIÓN AUTOMATIZADA JUSTOS

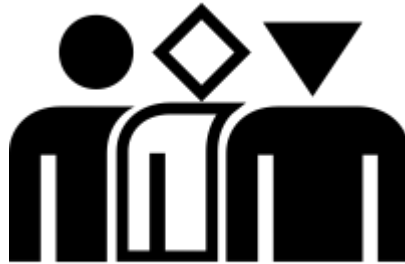
Patricia Díaz | @PatoDiazGNU

# 3 GRANDES DESAFÍOS

*PRIVACIDAD*



*DIVERSIDAD*



*TRANSPARENCIA*

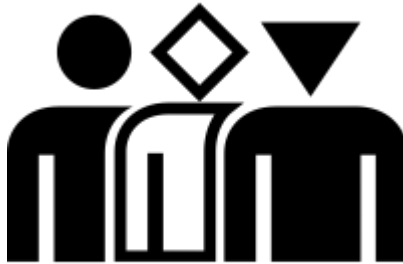


# 3 GRANDES DESAFÍOS

*PRIVACIDAD*



*DIVERSIDAD*



*TRANSPARENCIA*



# EL SESGO EN LOS ALGORITMOS AFECTA LA VIDA DE LAS PERSONAS

Resultados positivos	Resultados negativos
empleo ofrecido	empleo negado
préstamo aceptado	préstamo denegado
descuento ofrecido	descuento no ofrecido

# Discriminación en los pactos Internacionales

## **DISCRIMINACIÓN DIRECTA (trato desigual) por motivos de Raza (\*)**

Convención Internacional sobre la Eliminación de todas las Formas de Discriminación Racial la define como:

“toda **distinción, exclusión, restricción o preferencia basada en** motivos de raza, color, linaje u origen nacional o étnico **que tenga por objeto o por resultado** anular o menoscabar el reconocimiento, goce o ejercicio, en condiciones de igualdad, de los derechos humanos y libertades fundamentales.”

(\*) En la Convención sobre la eliminación de todas las formas de discriminación contra la mujer y la Convención sobre los derechos de las personas con discapacidad figuran definiciones similares.

## **DISCRIMINACIÓN INDIRECTA (impacto desigual):**

Comité de Derechos Económicos, Sociales y Culturales define la discriminación indirecta como “leyes, políticas o **prácticas en apariencia neutras pero que influyen de manera desproporcionada** en los derechos del Pacto afectados por los motivos prohibidos de discriminación”

# SESGO RACIAL

(Redlining)



Bloomberg



The northern half of Atlanta, home to 96% of the city's white residents, has same-day delivery. The southern half, where 90% of the residents are black, is excluded.

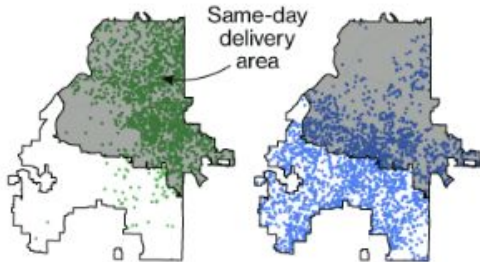
Dallas has the lowest overall coverage rate among cities with same-day delivery. White residents are more than twice as likely as black residents to have access to the service.

Out half of Chicago's black residents in the southern half of the city where they do not have access to Amazon's same-day delivery service.

Same-day service is unavailable in southeast Washington, D.C. including neighborhoods located blocks from the Capitol building and all areas across the Anacostia River.

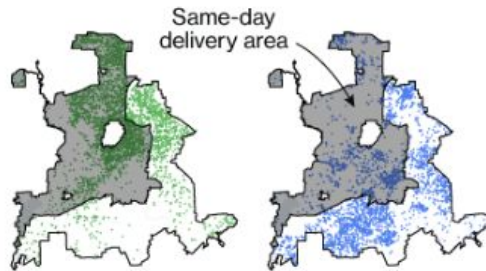
White residents

Black residents



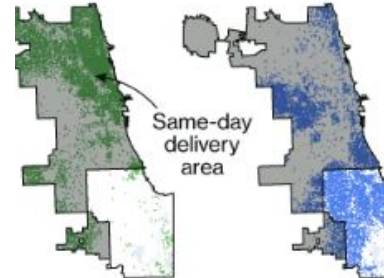
White residents

Black residents



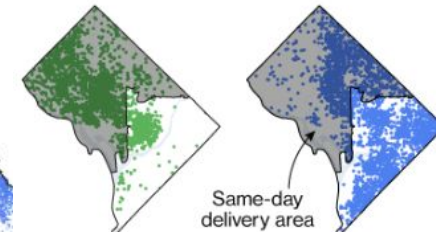
White residents

Black residents



White residents

Black residents



# MIT Researcher Exposing Bias in Facial Recognition Tech Triggers Amazon's Wrath

By Matt O'Brien | April 8, 2019



## SESGO RACIAL y de GÉNERO

(en sistemas de Reconocimiento Facial Automatizado)

[PROYECTO GENDER SHADES](#)

# Investigador del MIT que expone el sesgo en la tecnología de reconocimiento facial desencadena la ira de Amazon

Por Matt O'Brien | 8 de abril de 2019



## SESGO RACIAL y de GÉNERO

(en sistemas de Reconocimiento Facial Automatizado)

[PROYECTO GENDER SHADES](#)



## / Discriminación automatizada: Facebook utiliza estereotipos groseros para optimizar la entrega de anuncios

por *Nicolas Kayser-Bril*

Un experimento de AlgorithmWatch muestra que las plataformas en línea optimizan la entrega de anuncios de manera discriminatoria. Los anunciantes que los utilicen podrían estar infringiendo la ley.

**SESGO GÉNERO  
y EDAD**  
(en anuncios  
laborales)

## England exams row timeline: was Ofqual warned of algorithm bias?

Scrapping followed series of warnings over algorithm's volatility and fairness

**Ben Quinn and Richard Adams**

Thu 20 Aug 2020 19:53 BST



▲ Students opposite Downing Street protesting against the downgrading of A-level results on 16 August. Photograph: Matthew Chattle/Rex/Shutterstock

# CALIFICACIÓN A ESTUDIANTES DURANTE LA PANDEMIA EN UK

## Efecto Matthew

**“The rich get richer and the poor get poorer”**

El sistema discrimina a los estudiantes que históricamente tuvieron malas calificaciones

[\(Informe Automating Society, 2020\)](#)

# EL SESGO NO SIEMPRE ESTÁ EN LOS ALGORITMOS

Noticias hoy | Dólar blue hoy | Elecciones 2023 | Ganancias | Macri | Encuestas | Dónde vota Mendoza | Champions League | Sigue nuestro canal en Whatsapp | Horóscopo hoy | Virales | Martes, 19 de Septiembre de 2023

**Clarín** [Suscribite por \\$30](#) [Ingresar](#)

[En vivo](#) | [Sesión por Ganancias en Diputados](#) | [Elecciones 2023: 'Lo de Chaco fue apoteótico y el domingo ganamos en Mendoza', sentenció Bullrich](#)

**Sociedad**

## En Salta usan inteligencia artificial para prever embarazos adolescentes

El gobernador Juan Manuel Urtubey dijo que hay niñas "predestinadas a un embarazo adolescente". Y que las pueden identificar con nombre, apellido y domicilio para trabajar con ellas y prevenirlo.



**lm** Internacional Cultura Política Clima Sociedad Kiosco Especiales Opinión [Más](#)

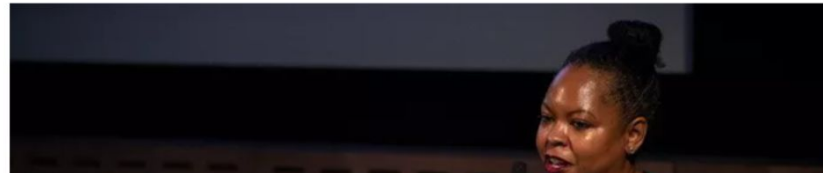
## Quin problema tenen amb les nostres llengües?

#LAMAREA96 [COMPRAR REVISTA](#)

### Sociedad

## Simone Browne: «La vigilancia biométrica es similar a la del tráfico de esclavos»

- Entrevista a la socióloga Simone Browne, autora de 'Dark Matters: On the Surveillance of Blackness'



# Los sistemas de **Procesamiento de Lenguaje Natural** también reproducen los sesgos sociales

DETECTAR IDIOMA   PERSA   **ESPAÑOL**   TURCO   ▾   ↔   **PERSA**   TURCO   ESPAÑOL   ▾

Él es enfermero   ×

Ella es investigadora

37/5000

او پرستار است ☆

او یک محقق است

🔊   📄   ✎   🔗

DETECTAR IDIOMA   **PERSA**   ESPAÑOL   TURCO   ▾   ↔   PERSA   TURCO   **ESPAÑOL**   ▾

او پرستار است ×

او یک محقق است

28/5000

Ella es enfermera ☆

El es investigador

🔊   📄   ✎   🔗

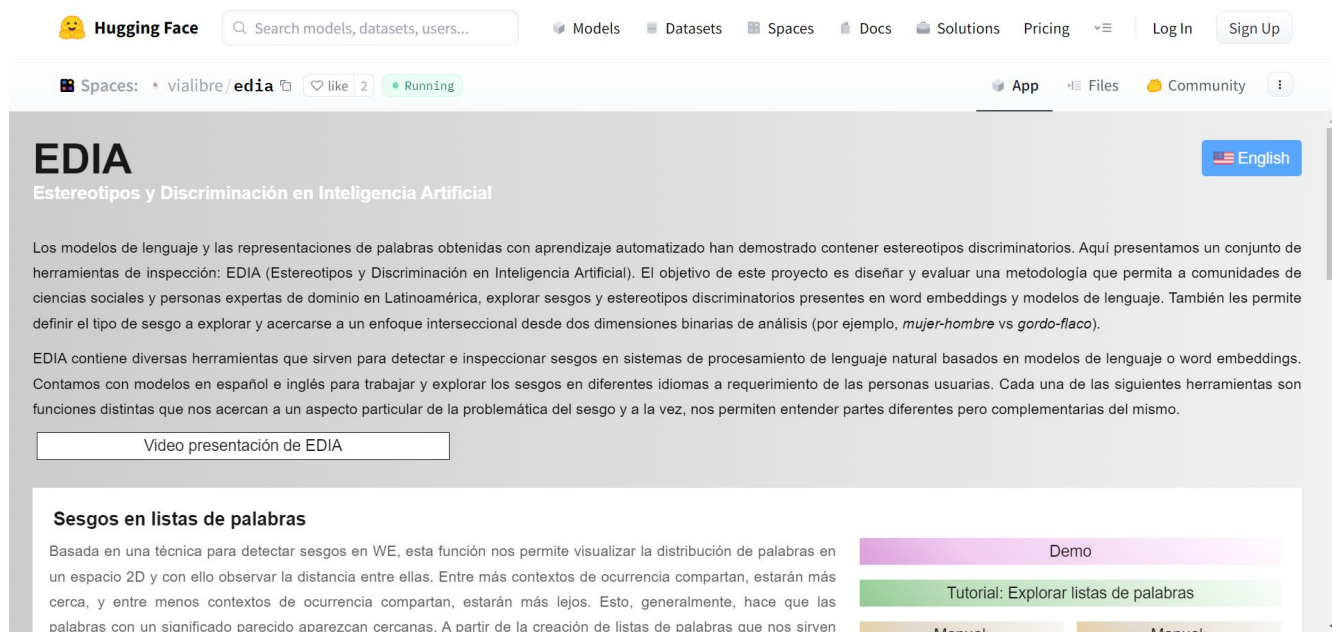
Enviar comentarios

# EDIA:

## Estereotipos y Discriminación en Inteligencia Artificial

### Metodologías para explorar el sesgo en MODELOS DE LENGUAJE NATURAL EN ESPAÑOL

[Herramientas desarrolladas por  
Fundación Vía Libre y por la  
Universidad Nacional de Córdoba](#)



The screenshot shows the Hugging Face interface for the EDIA space. At the top, there's the Hugging Face logo and a search bar. Below that, navigation links for Models, Datasets, Spaces, Docs, Solutions, Pricing, Log In, and Sign Up are visible. The main content area shows the space name 'vialibre / edia' with a 'like' button and a 'Running' indicator. The title 'EDIA' is prominently displayed, followed by the subtitle 'Estereotipos y Discriminación en Inteligencia Artificial'. A paragraph of text explains the project's goal: to design and evaluate a methodology for exploring bias in NLP models. Below this, there's a video player placeholder for 'Video presentación de EDIA'. At the bottom, there are three buttons: 'Demo', 'Tutorial: Explorar listas de palabras', and 'Manual'.

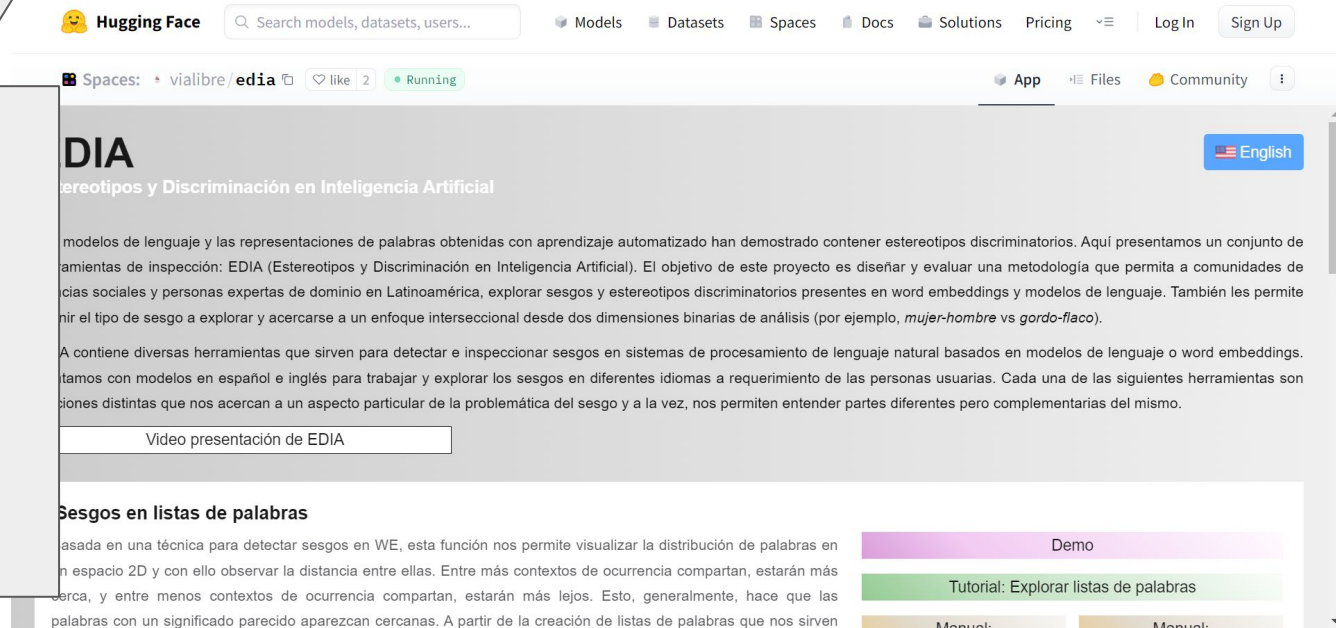
# EDIA:

## Estereotipos y Discriminación en Inteligencia Artificial

### Metodologías para explorar el sesgo en MODELOS DE LENGUAJE NATURAL EN ESPAÑOL

EJEMPLOS de herramientas que utilizan PLN:

- Google Traductor
- ChatGPT
- Asistente de voz Siri, Alexa
- Chatbots
- Motores de búsqueda (búsqueda predictiva)



**Hugging Face** Search models, datasets, users... Models Datasets Spaces Docs Solutions Pricing Log In Sign Up

Spaces: vialibre / edia like 2 Running App Files Community

## EDIA

### Estereotipos y Discriminación en Inteligencia Artificial

modelos de lenguaje y las representaciones de palabras obtenidas con aprendizaje automatizado han demostrado contener estereotipos discriminatorios. Aquí presentamos un conjunto de herramientas de inspección: EDIA (Estereotipos y Discriminación en Inteligencia Artificial). El objetivo de este proyecto es diseñar y evaluar una metodología que permita a comunidades de ciencias sociales y personas expertas de dominio en Latinoamérica, explorar sesgos y estereotipos discriminatorios presentes en word embeddings y modelos de lenguaje. También les permite definir el tipo de sesgo a explorar y acercarse a un enfoque interseccional desde dos dimensiones binarias de análisis (por ejemplo, *mujer-hombre* vs *gordo-flaco*).

A EDIA contiene diversas herramientas que sirven para detectar e inspeccionar sesgos en sistemas de procesamiento de lenguaje natural basados en modelos de lenguaje o word embeddings. Trabajamos con modelos en español e inglés para trabajar y explorar los sesgos en diferentes idiomas a requerimiento de las personas usuarias. Cada una de las siguientes herramientas son opciones distintas que nos acercan a un aspecto particular de la problemática del sesgo y a la vez, nos permiten entender partes diferentes pero complementarias del mismo.

Video presentación de EDIA

### Sesgos en listas de palabras

Basada en una técnica para detectar sesgos en WE, esta función nos permite visualizar la distribución de palabras en un espacio 2D y con ello observar la distancia entre ellas. Entre más contextos de ocurrencia compartan, estarán más cercanas, y entre menos contextos de ocurrencia compartan, estarán más lejos. Esto, generalmente, hace que las palabras con un significado parecido aparezcan cercanas. A partir de la creación de listas de palabras que nos sirven

Demo

Tutorial: Explorar listas de palabras

Manual: Manual:

# 3 GRANDES DESAFÍOS

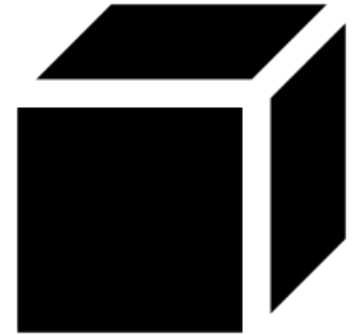
*PRIVACIDAD*



*DIVERSIDAD*

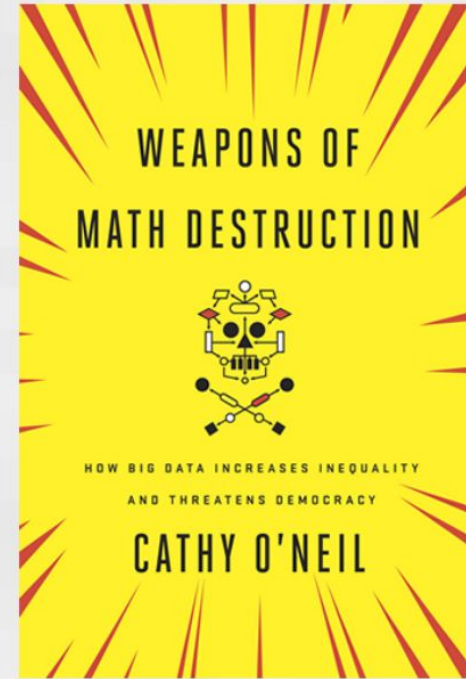
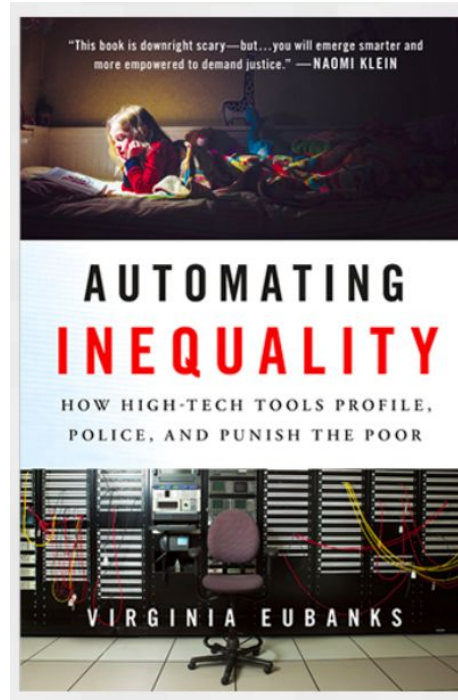


*TRANSPARENCIA*



# “LOS ALGORITMOS SON OPINIONES EMBEBIDAS EN CÓDIGO”

*(Cathy O’Neil, 2016)*





Education

## 'Creative ... motivating' and fired



Sarah Wysocki was out of work for only a few days after she was fired by DCPS last year. She is now teaching at Hybla Valley Elementary School in Fairfax County. (Jahl Chikwendu/The Washington Post)

By **Bill Turque**

March 6, 2012

## RANKINGS DOCENTES EN LAS ESCUELAS DE USA

Education

## 'Creative ... motivating' and fired



Sarah Wysoki was out of work for only a few days after she was fired by DCPS last year. She is now teaching at Hybla Valley Elementary School in Fairfax County. (Jahl Chikwendu/The Washington Post)

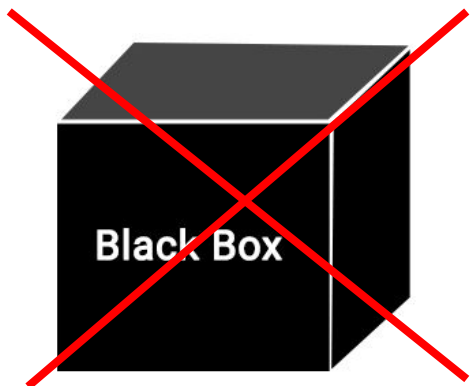
By **Bill Turque**  
March 6, 2012

## RANKINGS DOCENTES EN LAS ESCUELAS DE USA

¿Cómo detectar y corregir los  
sesgos si no comprendemos  
cómo se toman las decisiones  
automatizadas?

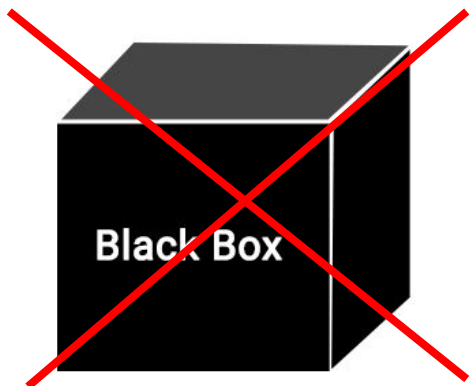
# ¿QUÉ ES LA TRANSPARENCIA ALGORÍTMICA?

- *transparencia algorítmica no es sinónimo de liberar el código fuente*



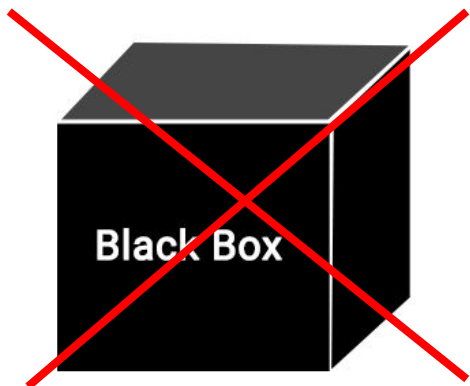
# ¿QUÉ ES LA TRANSPARENCIA ALGORÍTMICA?

- *transparencia algorítmica no es sinónimo de liberar el código fuente*
- *la transparencia algorítmica requiere transparencia de datos*

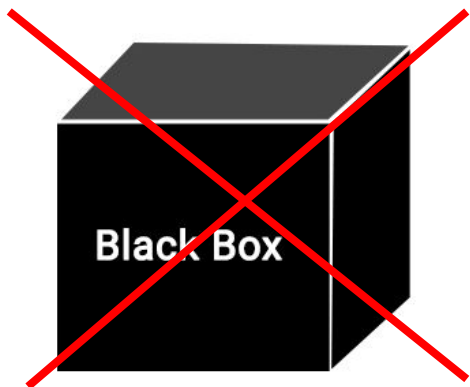


# ¿QUÉ ES LA TRANSPARENCIA ALGORÍTMICA?

- *transparencia algorítmica no es sinónimo de liberar el código fuente*
- *la transparencia algorítmica requiere transparencia de datos*
- *la transparencia de datos no es sinónimo de hacer públicos todos los datos*



# ¿QUÉ ES LA TRANSPARENCIA ALGORÍTMICA?



- *transparencia algorítmica **no es sinónimo de liberar el código fuente***
- *la transparencia algorítmica requiere **transparencia de datos***
- *la transparencia de datos **no es sinónimo de hacer públicos todos los datos***
- *la transparencia procesable requiere **interpretabilidad / explicabilidad***

*explicar los supuestos del sistema y sus efectos, involucrar a los actores que recibirán el impacto en la implementación*

# ¿EXISTE EL DERECHO A LA EXPLICABILIDAD?

## No exactamente...

URUGUAY Ley 18331

### Artículo 16 (Derecho a la impugnación de valoraciones personales)

“Las personas tienen derecho a no verse sometidas a una **decisión con efectos jurídicos** que les afecte de manera significativa, **que se base en un tratamiento automatizado de datos** destinado a evaluar determinados aspectos de su personalidad, como su rendimiento laboral, crédito, fiabilidad, conducta, entre otros.

El afectado podrá **impugnar los actos** administrativos o decisiones privadas que impliquen una valoración de su comportamiento, **cuyo único fundamento sea un tratamiento de datos personales que ofrezca una definición de sus características o personalidad.**

En este caso, el afectado tendrá **derecho a obtener información del responsable de la base de datos tanto sobre los criterios de valoración como sobre el programa utilizado** en el tratamiento que sirvió para adoptar la decisión manifestada en el acto.”

# ¿EXISTE EL DERECHO A LA EXPLICABILIDAD?

## No exactamente...

### URUGUAY Ley 18331

#### Artículo 16 (Derecho a la impugnación de valoraciones personales)

“Las personas tienen derecho a no verse sometidas a una **decisión con efectos jurídicos** que les afecte de manera significativa, **que se base en un tratamiento automatizado de datos** destinado a evaluar determinados aspectos de su personalidad, como su rendimiento laboral, crédito, fiabilidad, conducta, entre otros.

El afectado podrá **impugnar los actos** administrativos o decisiones privadas que impliquen una valoración de su comportamiento, **cuyo único fundamento sea un tratamiento de datos personales que ofrezca una definición de sus características o personalidad.**

En este caso, el afectado tendrá **derecho a obtener información del responsable de la base de datos tanto sobre los criterios de valoración como sobre el programa utilizado** en el tratamiento que sirvió para adoptar la decisión manifestada en el acto.”

#### PERO...

- **no se imponen obligaciones de transparencia proactiva.**
- **el mecanismo solo aplica sobre sistemas que tratan datos personales y cuando la decisión automatizada tenga “efectos jurídicos” que afecten a un individuo de forma significativa.**



# “Una evaluación y comunicación adecuadas de la funcionalidad debería ser un requisito mínimo para el despliegue masivo de sistemas algorítmicos”



≡ Navegación de artículos

## La falacia de la funcionalidad de la IA

**Inioluwa Deborah Raji** , Universidad de California, Berkeley, EE. UU., [deborahraji1@gmail.com](mailto:deborahraji1@gmail.com)

**I. Elizabeth Kumar** , Universidad de Brown, EE. UU., [iekumar@brown.edu](mailto:iekumar@brown.edu)

**Aaron Horowitz** , Unión Americana de Libertades Civiles, EE. UU., [ahorowitz@aclu.org](mailto:ahorowitz@aclu.org)

**Andrew Selbst** , Universidad de California, Los Ángeles, EE. UU., [aselbst@law.ucla.edu](mailto:aselbst@law.ucla.edu)

DOI: <https://doi.org/10.1145/3531146.3533158>

HECHO '22: [Conferencia ACM 2022 sobre equidad, responsabilidad y transparencia](#) , Seúl, República de Corea, junio de 2022

Los sistemas de IA desplegados a menudo no funcionan. Pueden construirse al azar, implementarse indiscriminadamente y promoverse de manera engañosa. Sin embargo, a pesar de esta realidad, los académicos, la prensa y los legisladores prestan muy poca atención a la funcionalidad. Esto conduce a soluciones técnicas y políticas centradas en implementaciones “éticas” o alineadas con el valor, a menudo omitiendo la pregunta anterior de si un sistema determinado funciona o proporciona algún beneficio. Para describir los daños de varios tipos de fallas de funcionalidad, analizamos un conjunto de estudios de casos para crear una taxonomía de problemas de funcionalidad de IA conocidos. Luego señalamos las políticas y las respuestas organizacionales que a menudo se pasan por alto y se vuelven más disponibles una

Esta presentación se encuentra licenciada con una licencia Creative Commons Atribución Compartir Igual



¡Puedes reutilizarla!

La iconografía utilizada es de uso libre y pertenece a [FLATICON](#).

ATENCIÓN:

Las capturas de prensa son usadas por la docente con fines educativos conociendo la falta de excepción al © y bajo su entera responsabilidad.