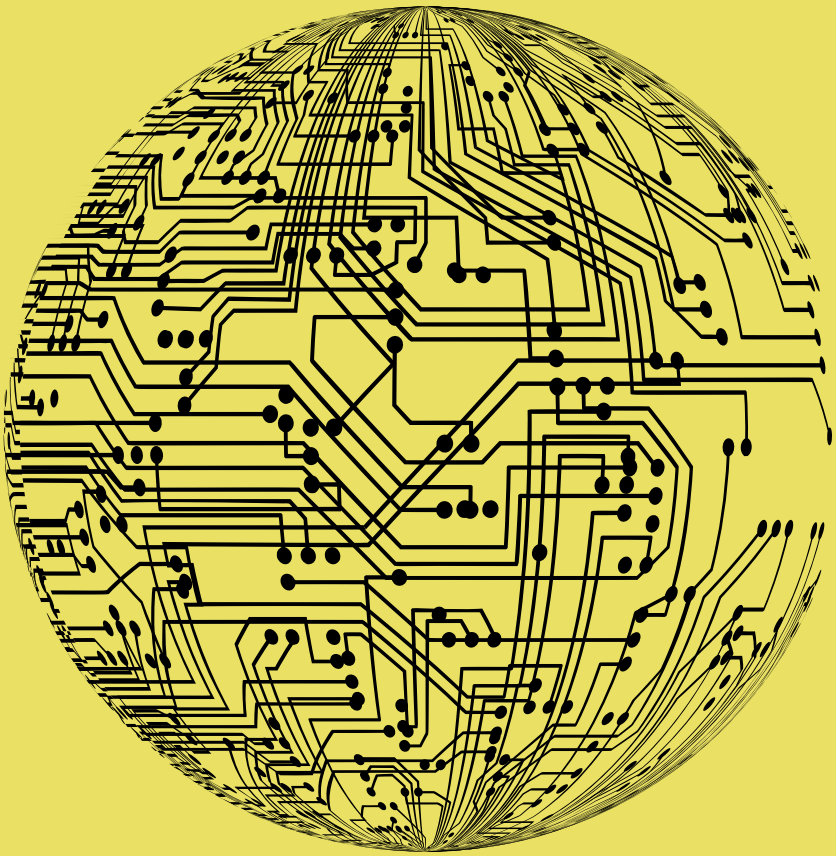
UNA INTRODUCCIÓN A LA IA Y LA DISCRIMINACIÓN ALGORÍTMICA PARA MOVIMIENTOS SOCIALES



AlgoRace

(Des)Racializando la IA

Noviembre de 2022
ALGORACE, (Des)Racializando la IA - www.algorace.org

Investigación Principal
ANA VALDIVIA GARCÍA - Profesora e investigadora en Inteligencia Artificial, Gobierno y Políticas en el Oxford Internet Institute de la Universidad de Oxford e integrante de AlgoRace.
JAVIER SÁNCHEZ MONEDERO - Investigador "Beatriz Galindo" en Inteligencia Artificial en la Universidad de Córdoba, investigador asociado en el Data Justice Lab de la Universidad de Cardiff e integrante de AlgoRace.

Coordinación
YOUSSEF M. OULED - Periodista colaborador en diferentes medios de comunicación. Divulgador e investigador sobre racismo. Coordinador de AlgoRace y del área antidiscriminación de Rights International Spain.
PAULA GUERRA CÁCERES - Licenciada en Comunicación Social, Máster en Acción Solidaria Internacional y de Inclusión Social. Analiza e investiga el racismo estructural y sus consecuencias. Conferenciante sobre Pensamiento decolonial. Columnista en medios digitales como Público, eldiario.es y Pikara Magazine. Integrante de AlgoRace.

Diseño y maquetación
ISABEL MURIEDAS DÍEZ - Experta en campañas de comunicación con organizaciones antirracistas. Especializada en sexualidad y violencia de género. Integrante de AlgoRace.

Este informe está dedicado a todas las personas y colectividades que a diario luchan contra las desigualdades y violencias estructurales

Índice

Resumen

El objetivo de este documento es hacer una introducción al problema de la discriminación algorítmica con el fin de proporcionar una herramienta de análisis a colectivos antirracistas y de defensa de los derechos fundamentales. Para ello, planteamos primero una introducción a la Inteligencia Artificial (IA) y la toma de decisiones algorítmicas, y luego una exposición de casos de discriminación racial y de otros tipos. Por último, se nombran algunas de las formas de resistencia surgidas en el seno de la sociedad civil para hacer frente a esas implementaciones discriminatorias. Este texto busca proporcionar una base para organizaciones sociales que permita entender qué tipos de sistemas se están desplegando para la toma de decisiones algorítmicas y cómo el racismo y la discriminación derivada del mismo se manifiesta en sus diferentes formas, con el objetivo de promover análisis más complejos que consideren las dinámicas sociales y relaciones de poder a la hora de entender el papel de estos sistemas.

Prólogo

Hace un año echaba a rodar una iniciativa que pretendía introducir una perspectiva crítica concreta a un debate que por complejidad, novedad o incomprensión carece de la relevancia necesaria. Esa iniciativa pasaba a llamarse AlgoRace: (Des)Racializando la IA que, con un nombre que es una declaración de intenciones, busca aportar una visión antirracista a los debates y discusiones en torno a la IA, sin desatender otras opresiones.

Hasta entonces no había un esfuerzo suficiente por facilitar el acceso a los espacios de debate sobre IA a aquellas personas y organizaciones racializadas que enfrentan a diario las múltiples manifestaciones del racismo estructural: Ley de Extranjería, explotación laboral, redadas por perfil racial, segregación escolar y residencial, violencia policial, etc.

Con el surgimiento de AlgoRace se conforma un equipo de trabajo integrado por diversas personas, unas procedentes del antirracismo y otras con especialización sobre IA, que como finalidad investiga y señala cómo la IA no solo no es neutral, sino que está sometida a los intereses del poder. En el caso concreto que nos atañe, hablamos de IA al servicio del colonialismo y el racismo, empleada como un mecanismo más de racialización.

Algoritmos, sistemas automatizados, datos biométricos, etc. son palabras que remiten a imaginarios complejos de los que se valen las instituciones para hacer cada vez más indetectable y sofisticado su racismo. Si estás leyendo esto y crees que no tiene nada que ver contigo te invitamos a adentrarte en una lectura accesible para todos los públicos. Un documento que señala cómo servicios del sector público y privado funcionan en base a decisiones automatizadas (SDAs). Nuestro presente está condicionado por sistemas tecnológicos que deciden sobre las ayudas públicas ofrecidas por la Administración, que nos evalúan o “predicen” las notas en el ámbito educativo, que controlan nuestro rendimiento en el trabajo, que trabajan para la policía vigilándonos, que en lugar de facilitar las rutas migratorias del sur al norte global, las hacen más infranqueables y mortales, y un largo etcétera. A lo que se suma un consumo de energía ingente y continuo de estos sistemas que afecta al cambio climático mientras traslada los residuos al sur global.

¿Para quién o para qué se usa la IA? ¿Es realmente indispensable? ¿A quién beneficia? ¿Qué coste medioambiental conlleva? ¿Contribuye a construir sociedades más participativas y democráticas o sostiene jerarquías ya existentes? ¿Quién tiene acceso a ella? ¿Quién participa en el debate? Este documento no busca dar respuestas concretas, pero sí desmitificar un uso de la IA que, como se ha dicho, ni es neutral ni es objetivo, porque como señalan los autores de este informe, se implementa con una carga política que reproduce violencias estructurales: “En la mayoría de casos, la IA y los SDAs son aplicados bajo una jerarquía de poder: de arriba abajo, de ricos a pobres, de privilegiados a marginalizados, de blancos a sujetos racializados, de hombres a mujeres o LGTBIQ+”.

Tras un año de trabajo, en AlgoRace hemos conseguido que se hable más de racismo en los debates sobre IA, pero hace falta que se hable más de IA en el antirracismo porque, como partícipes de la lucha antirracista y entendedores de las urgencias y complejidades de la misma, nos parece fundamental que otras voces se pronuncien, cuestionen y combatan ciertos usos de la IA. Este informe busca contribuir a la consecución de ese fin.

1. Introducción

En los años 80, la puesta en marcha de un algoritmo diseñado para automatizar el proceso de admisión de la Escuela de Medicina del Hospital St. George's en la ciudad de Londres terminó disminuyendo la diversidad del alumnado admitido y excluyendo por género y raza[1]. Otros estudios como los del investigador Bernard E. Harcourt identifican sistemas de evaluación numérica para la admisión en asilos, hospitales psiquiátricos y cárceles que discriminan racialmente desde los años 20 del siglo pasado[2].

La lista creciente de casos responde a muchas explicaciones. Según los actores y las relaciones de poder en las que se sitúan socialmente, estos casos se deben a efectos no intencionados o problemas de mala calidad en los datos de calibración y evaluación, pero si acudimos a análisis más rigurosos encontramos explicaciones que enmarcan todos estos sistemas sociotécnicos dentro de un análisis de racismo estructural en el que la tecnología se concibe, diseña e implanta reproduciendo lógicas sociales, violencias estructurales y políticas preestablecidas.

Por todo ello, este documento muestra en qué consiste la inteligencia artificial y el racismo que conlleva, qué proyectos se han llevado a cabo tanto a nivel nacional e internacional vulnerando derechos fundamentales, así como qué resistencias se han originado en contra de esta tecnología. Este texto va dirigido a los movimientos sociales y colectivos antirracistas que ven en el uso de la tecnología otra vía de opresión racial.

[1] Merino, M. (22 de abril de 2019). Los algoritmos con sesgo racial y de género son un problema que venimos arrastrando desde los años 80. *Xataka.com*. https://www.xataka.com/inteligencia-artificial/algoritmos-sesgo-racial-genero-problema-que-venimos-arrastrando-anos-80

[2] Harcourt, B. E. (2015). *Risk as a proxy for race: The dangers of risk assessment.* Federal Sentencing Reporter, 27(4), 237-243. https://papers.ssrn.com/abstract=1677654

2. Inteligencia artificial y sistemas de decisión automática

El concepto de IA se ha popularizado en los últimos años. Escuchando la radio o podcasts, leyendo la prensa digital o incluso navegando por redes sociales nos encontramos con este término. En general, la imagen que se tiene de esta tecnología es la de un robot humanoide, normalmente blanco en un fondo azul *(ver Figura 1)*. ¿Pero qué significa "inteligencia artificial"? ¿Cómo se acuñó esta rama de la tecnología? ¿Qué tipos de tecnología engloba? ¿Existe una definición consensuada?

En 2019, el 'Grupo de Expertos de Alto Nivel en Inteligencia Artificial' propuso la siguiente definición:

> Los sistemas de inteligencia artificial (IA) son sistemas de software (y posiblemente también de hardware) diseñados por humanos que, dado un objetivo complejo, actúan en la dimensión física o digital percibiendo su entorno mediante la adquisición de datos, interpretando los datos estructurados o no estructurados recogidos, razonando sobre el conocimiento, o procesando la información, derivada de estos datos y decidiendo la mejor acción o acciones a realizar para lograr el objetivo dado. Los sistemas de IA pueden utilizar reglas simbólicas o aprender un modelo numérico, y también pueden adaptar su comportamiento analizando cómo se ve afectado el entorno por sus acciones anteriores [3].

Si bien esta definición propone un verdadero sistema de IA, la mayoría de aproximaciones consideradas como IA a día de hoy en medios, política, marcos regulatorios, etc. no cumplen esta definición. Otras alternativas proponen hablar de la IA como un conjunto de conceptos, problemas y métodos para resolver los problemas más que de una categoría "es IA" o "no es IA" aplicada a un sistema informático [4]. Es cierto que la IA está formada por sistemas de software y hardware, es decir, datos, código y ordenadores, que son a su vez diseñados por humanos. No obstante, el objetivo no tiene por qué ser complejo. Lejos de ser tareas complejas, existen IAs capaces de detectar si en una imagen hay un gorrión o traducir automáticamente al árabe el último ensayo de Remedios Zafra.

Tampoco podríamos decir que la IA 'percibe' el entorno, pues carece de esa capacidad. Esta tecnología procesa nuestra realidad con datos, los cuales son una representación cuantitativa de una parte de una realidad subjetiva. Las personas que generan los datos, los procesan o almacenan, reflejan en ellos sus observaciones. Imaginemos una base de datos creada por un cuerpo policial sobre la actividad criminal de una ciudad española. Realmente, la base de datos no reflejará el nivel de criminalidad por cada barrio, sino cuál es el nivel de presencia policial. Tan solo las detenciones observadas por la autoridad policial serán recogidas en forma de dato, dejando invisible todo otro crimen que no sea testimoniado por una patrulla.

[3] High-Level Expert Group on Artificial Intelligence (2019). *A definition of AI: Main capabilities and disciplines.* European Commission. https://www.aepd.es/sites/default/files/2019-12/ai-definition.pdf

[4] Universidad de Helsinki (2021). *¿Qué es la IA?*. Elements of AI. https://course.elementsofai.com/es/1/1

Los datos, que a su vez son almacenados en bases de datos, suponen la fuente principal de información para que la IA pueda alcanzar el objetivo. Por ejemplo, en el caso de la IA que detecta gorriones el objetivo es maximizar el número de imágenes que acierta en su veredicto. Para ello, se alimenta a la IA con una gran cantidad de imágenes con gorriones y sin gorriones, las cuales han sido previamente etiquetadas (gorrión sí - gorrión no) por un humano. En este caso concreto, la IA interpreta las imágenes analizando y troceando los píxeles que la forman. Guiada por una función matemática que le penaliza si se equivoca (dice que hay un gorrión cuando no lo hay, o no detecta un gorrión cuando si lo hay), la IA encuentra patrones dentro de las imágenes que le ayudan a diferenciar cuándo hay gorrión y cuándo no.

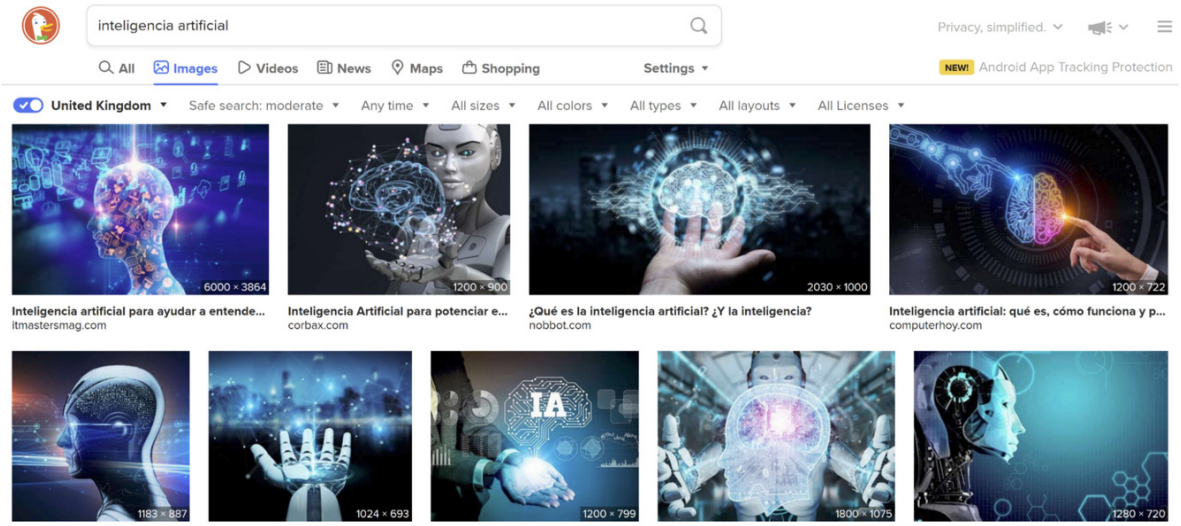


Figura 1: Imágenes que nos devuelve el buscador "Duck Duck Go" cuándo buscamos la palabra "inteligencia artificial"

Bases de datos, como hojas de Excel, son considerados datos estructurados pues mantienen una estructura sólida con columnas y filas. Por otro lado, imágenes, textos, audios y otro tipo de información audiovisual son considerados datos no estructurados, al carecer de dicha estructura. Por último, la IA no razona sobre el conocimiento. No se puede otorgar a esta tecnología la capacidad de razonar, pero sí la de realizar cálculos rápidamente, mucho más que los humanos. Es por ello que le otorgamos esa capacidad de inteligencia. Pero de esa manera, una calculadora también podría ser considerada inteligente.

En 1950, Alan Turing [5] publicó un documento de referencia para el campo de la informática en el que mostraba la posibilidad de crear máquinas que piensan. En este manuscrito Turing propuso lo que hoy se conoce como el Test de Turing, un ejercicio práctico para discernir cuándo una máquina puede pensar:

5 Alan Turing tuvo un papel importante durante la Segunda Guerra Mundial, ayudando al ejército aliado a desencriptar mensajes del ejército nazi. A pesar de su contribución, en 1952 el Gobierno británico lo acusó de "actos homosexuales" y de "ultraje contra la moral pública". Turing aceptó la castración química para evitar la prisión. Dos años más tarde, se suicidó con tan solo 41 años. Hoy en día Turing es considerado uno de los mayores contribuidores a la teoría de la ciencia de la computación y la inteligencia artificial.

> Imagínate una habitación separada por una mampara opaca. A un lado se encuentran un ordenador (A) y una persona (B). Al otro lado, una persona (C). La persona C realiza preguntas a A y B pensando que son personas y tiene que averiguar su género. Si al acabar el experimento la persona C no detecta que A es un ordenador, se puede concluir que la máquina tiene la capacidad de "pensar".

Con este ejemplo, Turing mostró lo complicado que es definir el "pensar". En 1956, un equipo de académicos estadounidenses e ingenieros de empresas tecnológicas organizaron una reunión para investigar si las máquinas podían simular la inteligencia o el aprendizaje humano *(ver Figura 2)*. Fue la primera vez que se acuñó la palabra "Inteligencia Artificial". Este evento fue el inicio de la consolidación de la IA como campo de investigación. No obstante, debido a las capacidades limitadas de los ordenadores en aquella época, el campo entró en decadencia. Años más tarde, con el diseño de ordenadores más sofisticados el campo volvió a despegar. Hoy en día, gracias a las tarjetas gráficas diseñadas para mejorar los gráficos de videojuegos y la *datificación*[6] de nuestras vidas, la IA ha avanzado en aspectos como la detección de objetos en imágenes o el tratamiento de textos.

Figura 2: Durante el verano de 1956, se propuso a 10 académicos blancos reunirse para estudiar la posibilidad de la "inteligencia artificial" financiados por la fundación estadounidense Rockefeller
Fuente: https://www.scienceabc.com/wp-content/uploads/2018/01/John-maccarthy-marvin-minsky-claude-shannon-ray-solomonoff-alan-newell-herbert-simon-arthur-samuel-oliver-selfridge-nathaniel-rochester-trenchard-more-the-founding-fathers-of-ai.webp.

Debido a la propaganda que se le ha dado en los últimos años a esta tecnología, diferentes voces críticas se han alzado para desmitificarla. La IA no es inteligente, ni artificial. No es inteligente porque no es capaz de razonar y tampoco la podemos etiquetar de artificial porque está construida a partir de recursos naturales (cobre, silicio, fósforo) y necesita materia prima para funcionar. Darle la capacidad de artificial sería ignorar el proceso de diseño y montaje de un ordenador o de un centro de datos.

En este documento, nos centraremos en analizar sistemas de decisión automática (SDA). Estos sistemas se basan en tomar la "mejor" decisión respecto a un objetivo en función de los datos aportados y se utilizan dentro del sector financiero, marketing, salud o hasta en el

6 Se entiende como proceso de dataficación al proceso de recopilación, almacenamiento y procesamiento de datos. Desde que nos levantamos hasta acostarnos estamos generando datos continuamente a través de nuestros teléfonos móviles, ordenadores u otros dispositivos. Hoy en día, todo producto relacionado con la palabra *smart* (smart TV, smart phone, smart borders) implica la generación y análisis de datos que generamos mediante su uso.

sistema educativo. Por ejemplo, la mayoría de sistemas utilizados por la banca para conceder un crédito a clientes utilizan SDAs. Los SDAs pueden ser muy beneficiosos para organizaciones que necesiten tomar decisiones en problemas que están representados por datos. En el caso del crédito, los bancos almacenan características de sus clientes, tales como género, edad, profesión, salario o código postal. Dados estos datos, cuando un cliente pide un crédito al banco el SDAs encuentra qué decisión es la más beneficiosa (para la entidad) dada las características del cliente. No obstante, no debemos pensar en estos sistemas como programas sofisticados. En muchos casos, se basan en reglas diseñadas por personas expertas en el contexto y un programador o programadora, es decir, volviendo al caso del banco: "Si EDAD cliente < 45 -> No conceder crédito". En este caso, si el cliente que solicita el crédito tiene menos de 45 años, el banco no procederá a conceder el crédito.

Como hemos visto, los algoritmos pueden crearse de dos formas: de manera directa formalizando reglas y cálculos mediante lenguajes de programación, o de manera indirecta a través de técnicas de inteligencia artificial que extraigan información de los datos. De hecho, algunos textos hablan del aprendizaje máquina como una forma de generar código de programación de forma automática.

2.1. Las diferentes áreas de la IA

Dentro del campo de la inteligencia artificial existen diferentes áreas de estudio. El aprendizaje automático o *machine learning*, considerado una de las principales ramas de la IA, estudia el desarrollo de algoritmos para encontrar patrones y relaciones entre los datos. Esta rama se fundamenta en las matemáticas (álgebra, cálculo diferencial) y la estadística (investigación operativa).

Existen diferentes tipos de algoritmos de aprendizaje automático dentro de los cuales hay innumerables propuestas, pero casi todas se agrupan según el tipo de tarea que resuelvan: aprendizaje supervisado, aprendizaje no supervisado y aprendizaje por refuerzo. Este tercer tipo de aprendizaje es menos común en sistemas de gobernanza algorítmica y no nos centraremos en él.

En el aprendizaje supervisado un algoritmo debe asignar una etiqueta o clase a un patrón, por ejemplo, presencia o ausencia de enfermedad en un registro médico, detectar un gato en una fotografía, o estimar un valor numérico, como por ejemplo cuánta lluvia por metro cuadrado se espera. Este tipo de aprendizaje necesita, por tanto, disponer de datos etiquetados bien por personas o bien recogidos a través de sensores. Lo habitual en sistemas sociotécnicos enfocados a la gobernanza algorítmica es que tanto los datos de entrada de la aplicación como la etiqueta o valor numérico hayan sido establecidos por una o varias personas, por ejemplo, el historial de fraude en ayudas públicas, que consistiría en un conjunto de variables que describen a una persona (edad, ingresos, código postal, etc.) y una etiqueta que indique si esa persona ha cometido, o se considera que lo ha cometido, fraude o no. Otro ejemplo aplicado al campo social podría ser la puntuación con una escala de riesgo de reincidencia criminal a una persona. Esta etiqueta es relevante en el proceso de diseñar un algoritmo de aprendizaje automático, pues el sistema aprenderá patrones en función de dichas etiquetas.

El aprendizaje no supervisado consiste en asignar grupos o categorías a conjuntos de datos que no han sido etiquetados previamente (por ello, no supervisado pues no hay etiquetas). Un ejemplo son los algoritmos de agrupamiento o *clustering* en los que el algoritmo trata de agrupar los datos en varias categorías de forma automática en oposición a datos donde la categoría ya está definida. Por ejemplo, algunos productos de contratación automática puntúan a las personas candidatas con varias pruebas psicotécnicas para medir el tiempo de respuesta, la memoria a corto plazo, etc., y después utilizan técnicas de aprendizaje no supervisado como el agrupamiento para encontrar distintos tipos de perfiles de empleados y empleadas que se comporten de manera similar en las pruebas y asignarlos a diferentes categorías. Este ejemplo se visualiza en la Figura 3.

Una empresa dice haber desarrollado un producto que clasifica a los/as trabajadores y candidatos/as según su talento para diferentes tareas en la empresa. ¿Cómo funciona esto?

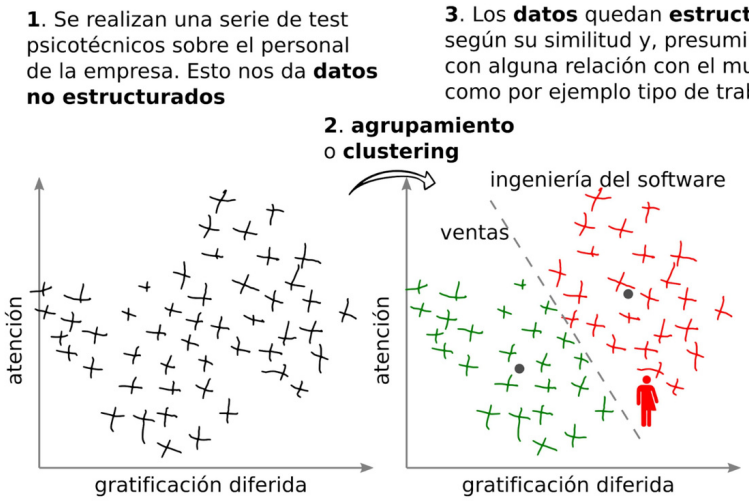


Figura 3: Ejemplo de algoritmo de agrupación. Inicialmente se dispone de datos no etiquetados donde cada punto corresponde a las puntuaciones de una persona en dos tests. El algoritmo de agrupamiento trata de dividir los datos en varios grupos, dos en este caso, para determinar distintos tipos de trabajadores y trabajadoras según su puntuación en los test psicotécnicos. En el ejemplo, la empresa decide que un grupo presenta habilidades más relacionadas con el departamento de ventas y otro grupo corresponde a perfiles de ingeniería del software. / Elaboración propia

Por lo general, para la construcción de sistemas inteligentes, el diseño pasa por la reducción del problema en el mundo real a tres tipos de tareas de aprendizaje automático: clasificación, estimación/puntuación numérica o agrupación de casos similares. Estos son todos los "ladrillos" disponibles para desarrollar IA. Este reduccionismo por un lado, y la aspiración de universalidad de las categorías y puntuaciones, por otro, tienen una fuerte crítica desde el punto de vista antirracista, decolonial y epistémico [7].

7 Broussard, M. (2018). *Artificial Unintelligence: How computers misunderstand the world*. Cambridge, MA: The MIT Press.

2.2. Entonces, ¿cómo se construye un sistema de decisión automática?

Los sistemas de decisión automática no necesariamente se programan utilizando técnicas de IA. Pueden consistir en reglas más o menos explícitas para tomar una decisión directamente a partir de datos, o hacer uso de una o varias técnicas estadísticas o de IA de forma implícita. Un ejemplo para construir un programa que filtre correo no deseado con programación tradicional o con programas basados en aprendizaje automático lo vemos a continuación *(Figura 4)*:

Figura 4: Ejemplo que clasifica automáticamente correo no deseado diseñado con reglas. Adaptado del ejemplo de Jason's Machine Learning 101.

Programación tradicional

Reglas explícitas:

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

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.

Si pensamos en crear un programa que filtre correos basura de forma manual necesitaríamos identificar palabras clave, algunas quizás muy claras, como 'viagra', pero la diversidad de situaciones a considerar en seguida nos empezaría a hacer la tarea realmente difícil. Este es un ejemplo donde la IA puede ayudar aprendiendo estas reglas a partir de ejemplos. En este problema, el aprendizaje automático puede funcionar bien porque contamos con muchos patrones etiquetados (millones de correos etiquetados como legítimos o basura) y la tarea a realizar está claramente relacionada con estos patrones. Así, conforme la decisión se tome en base a muchas variables y esté compuesto de muchos datos que contemplar en las reglas, es probable que las técnicas de IA comiencen a ser más útiles que la programación tradicional. Además, algunos tipos de datos llamados brutos o no estructurados, como las imágenes, vídeos, sonidos o texto son difíciles de procesar por sistemas de reglas programados manualmente, de manera que se utilizan técnicas estadísticas y de IA para poder extraer información de ellos. En todo caso, no debemos olvidar que lo habitual es insertar la salida o análisis de estas técnicas dentro de un programa informático que realice o recomiende una acción a través de distintos medios, cuya influencia en la persona que toma la decisión son también cuestión de estudio.

Decisiones basadas en reglas

El anterior ejemplo de programación tradicional es una forma de establecer decisiones automáticas a través de reglas. Un algoritmo basado en reglas se puede expresar visualmente como un diagrama de flujo donde diferentes evaluaciones de los datos llevan a diferentes resultados finales. Un tipo de estos sistemas son los árboles de decisión, donde existe un punto de entrada al algoritmo y se van tomando diferentes caminos hasta llegar a las hojas. Por ejemplo, podemos visualizar las reglas del algoritmo del simulador del Ingreso Mínimo Vital (que veremos en la sección 3) como un árbol de decisión *(Figura 5)*.

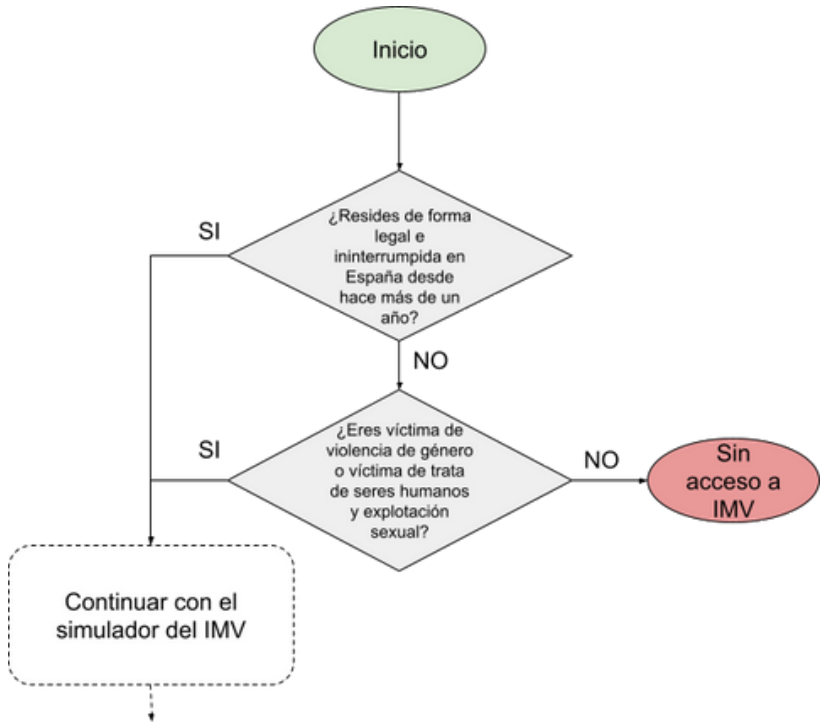


Figura 5: Ejemplo de implementación de un sistema de decisión automática mediante un árbol de decisión. Cada rama del árbol se divide en dos ramas que se recorren o no según se cumpla la condición que se evalúa hasta llegar al final del árbol, las hojas, que corresponden a la decisión automatizada.

Sistemas de puntuación

Otra forma de realizar decisiones automáticas es mediante una suma ponderada de una serie de variables. Una metodología parecida a la puntuación de un examen, donde cada ejercicio tiene una puntuación diferente y entre todos los ejercicios hay una puntuación máxima. Algo similar se aplica a evaluaciones de riesgo de pobreza, sufrir violencia de género o, por ejemplo, para obtener una plaza en la Escuela Oficial de Idiomas, en donde se consideran las rentas por unidad familiar calculadas sobre el dinero total que gana una familia, dividida por el número de miembros de la familia para priorizar este servicio a familias con menos ingresos.

Los sistemas de puntuación pueden realizarse manualmente o automáticamente a través de herramientas estadísticas que calculan el peso de cada variable en la puntuación final. Este es el caso del algoritmo utilizado por el Servicio Público de Empleo Austriaco para puntuar a los parados de larga duración y dividirlos en 3 categorías que recibían un trato diferente por parte del servicio de empleo estatal (este ejemplo se desarrolla en la Sección 3.1.2.).

Tareas de decisión automática implementadas con técnicas de IA

En el marco específico de la IA, en informática se suelen establecer 4 grandes categorías de tipos de tareas que se pueden automatizar a partir de datos que representen el problema que se quiere modelar. Estos son:

- **Clasificación**. Clasificar o etiquetar los datos en una o múltiples clases, como por ejemplo aceptar o denegar una solicitud de préstamo bancario.
- **Regresión**. Asignar un valor numérico, como por ejemplo los puntos en una escala de riesgo de que alguien cometa un crimen.
- **Sistemas de recomendación**. Sistemas que recomiendan productos, canciones, vídeos o incluso personas candidatas a un puesto de trabajo. A veces, también se denominan sistemas de ranking, ya que el sistema ordena los elementos según los recomiende para la petición.
- **Otros**. Sistemas más recientes como los de traducción de idiomas, la generación automática de imágenes o de su descripción no encajan claramente en las tres categorías anteriores pero son de total actualidad.

¿De qué hablamos cuando decimos “a partir de datos”? Por ejemplo, en el caso del crédito bancario cada persona se representa con un conjunto de variables como el tipo de trabajo, ingresos medios, saldo medio en las cuentas en los últimos años, etc. Los bancos suelen disponer del histórico de clientes con este tipo de información junto con el resultado final de la operación (si la persona finalmente pagó o no todo el préstamo en el plazo). A estos datos se les llama datos tabulares y se parecerían a una hoja de cálculo. En oposición, se suele hablar de datos no tabulares o no estructurados cuando se trabaja con texto, vídeo, sonido o imágenes.

En general las tareas que realizan los sistemas de reglas son de clasificación mientras que los sistemas de puntuación corresponden a la regresión o al ranking.

3. Casos de estudio por áreas

Existen ya casos de estudio en diferentes partes del mundo relacionados con la discriminación racial y la automatización de la toma de decisiones. Por ello, en este capítulo presentamos diferentes ejemplos reales divididos en sectores (sistema del bienestar, educación, precariedad laboral, etc.).

3.1. Sistema del bienestar

En los últimos años la digitalización y automatización de los servicios sociales se ha expandido por varios países. En 2019, el Relator Especial de la ONU sobre la extrema pobreza y los derechos humanos presentó un contundente informe que analizaba la digitalización del estado de bienestar que resume casos de 34 países (la mayoría del Norte Global)[8] :

> 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.

Vemos, por tanto, que la digitalización del estado de bienestar tiene en general una orientación hacia la reducción de costes utilizando la tecnología como mediadora entre personas y gobiernos y, como consecuencia, muchos análisis sitúan estos desarrollos dentro de las políticas de "austeridad" de la Unión Europea [9] . Pero hay más, ya que muchos de estos sistemas terminan reproduciendo e incluso amplificando desigualdades y discriminaciones estructurales por cuestiones raciales, género, clase, etc. El informe Data Harm Record (Registro de Daños por Datos) del Data Justice Lab[10] identifica numerosos casos internacionales de daños causados por sistemas algorítmicos relacionados con el estado de bienestar, incluyendo numerosos casos de discriminación

8 Alston, P. (2019). *Report of the Special Rapporteur on extreme poverty and human rights.* https://digitallibrary.un.org/record/3834146

9 Dencik, L. (2022). *The Datafied Welfare State: A Perspective from the UK.* In A. Hepp, J. Jarke, & L. Kramp (Eds.), New Perspectives in Critical Data Studies: The Ambivalences of Data Power (pp. 145–165). Springer International Publishing. https://doi.org/10.1007/978-3-030-96180-0_7

10 Redden, J., Brand, J. & Terzieva, V. (2020). *Data Harm Record* (Updated). Data Justice Lab. https://datajusticelab.org/data-harm-record/

racial en el contexto de concesión de créditos hipotecarios, diferenciación de precio en cursos de acceso a la universidad, reconocimiento facial, policía predictiva, admisión hospitalaria, etc. Otros informes más recientes como el Impermissible AI and fundamental rights breaches, elaborado por European Digital Rights, enumeran numerosos casos en el contexto de la Unión Europea [11].

3.1.1. Simulador del Ingreso Mínimo Vital

Como vimos en la sección 2, el simulador del Ingreso Mínimo Vital (IMV) es una web donde las personas pueden introducir sus datos para ver si cumplen los requisitos para acceder al IMV[12]. Este simulador comienza con dos preguntas para filtrar quién accede o no al trámite según su estatus migratorio *(ver Figuras 6 y 7)*. Este simulador parece, además, no estar alineado con la normativa del IMV, que contempla excepciones al año de residencia en España, como viajes de determinada duración de tiempo u otros motivos específicos. El simulador no sustituye a la evaluación real, sin embargo, puede desincentivar a personas con derecho a presentarse, sobre todo a personas migrantes.

Figura 6: Captura de pantalla del simulador del Ingreso Mínimo Vital.

[11] European Digital Rights. (2020). *Use cases: Impermissible AI and fundamental rights breaches.* European Digital Rights. https://edri.org/wp-content/uploads/2021/06/Case-studies-Impermissible-AI-biometrics-September-2020.pdf

[12] Ministerio de Inclusión, Seguridad Social y Migraciones. (2020). Simulador del Ingreso Mínimo Vital. Consultado en octubre de 2022. https://ingreso-minimo-vital.seg-social-innova.es/

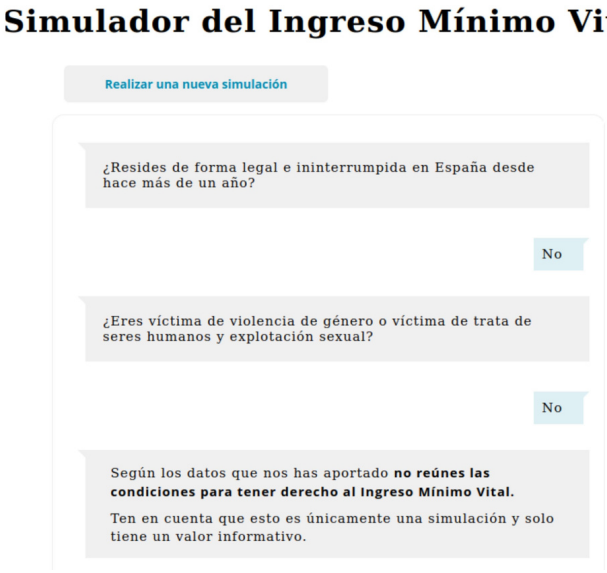


Figura 7: Captura de pantalla de una simulación del IMV aportando el perfil que tendría una persona migrante con poco tiempo de permanencia documentada en el país.

3.1.2. Caso Perfilado de Personas Desempleadas en Austria

En Austria, el Servicio Público de Empleo (Arbeitsmarktservice o "AMS") utiliza un algoritmo para predecir las perspectivas de empleo de una persona demandante de empleo basándose en factores como el género, grupo de edad, ciudadanía, salud, ocupación y experiencia laboral [13] . El algoritmo del AMS divide a los demandantes de empleo en 3 grupos según perspectivas de empleo calculadas (poca, moderada o alta) y con ellos establece la prioridad para el acceso a sus servicios, como la asistencia en la búsqueda de empleo, la colocación y los planes de formación. Los demandantes del grupo con perspectiva moderada (grupo B) tiene prioridad en el acceso mientras que el AMS reduce el apoyo a los solicitantes de empleo con perspectivas de empleo bajas (grupo C) o altas (grupo A), razonando que dicho apoyo tendría un efecto insignificante en sus posibilidades de contratación.

El gobierno austriaco sostiene que los solicitantes de empleo con pocas perspectivas requieren otro tipo de intervención para "estabilizar su situación personal y reforzar su motivación". Se les redirige a otra agencia gubernamental especializada en ayudar a personas en "situaciones de crisis" que impiden un empleo sostenible, como adicciones, deudas o problemas familiares[14].

El modelo matemático de decisión lo podemos ver en la siguiente figura (8). Se ha simplificado de 22 a 6 variables por motivos ilustrativos. Este consiste en una suma de variables multiplicadas por un valor (llamado coeficiente) que aumenta o disminuye la influencia de esas variables en la puntuación final de la persona. Por ejemplo, la primera variable es el género femenino (“geschlecht weiblich” en alemán) que se representa con un “1” para las mujeres y un “0” para los hombres. Esta variable se multiplica por -0.14, de

13 Allhutter, D., Cech, F., Fischer, F., Grill, G., & Mager, A. (21 de febrero de 2020). Algorithmic profiling of job seekers in Austria: how austerity politics are made effective. Frontiers in Big Data, 3:5. https://www.frontiersin.org/articles/10.3389/fdata.2020.00005/full./wp-content/uploads/2021/06/Case-studies-Impermissible-AI-biometrics-September-2020.pdf

14 Human Rights Watch. (2021). *How the EU’s Flawed Artificial Intelligence Regulation Endangers the Social Safety Net: Questions and Answers.* Human Rights Watch. https://www.hrw.org/news/2021/11/10/how-eus-flawed-artificial-intelligence-regulation-endangers-social-safety-net

manera que la fórmula disminuye la puntuación en las mujeres en una cantidad "-0.14x1" mientras que a los hombres no les resta puntos en la evaluación al ser "-0.14x0" cero. Las dos siguientes variables se refieren al grupo de edad ("altersgruppe") donde el rango 30-49 resta puntos a razón de -0,13 mientras que el grupo de edad de mayores de 50 resta -0.70. Igualmente, pertenecer a un país de la UE incrementa la puntuación en +0,16. Así, en igualdad de otras condiciones del resto de variables (experiencia, tiempo en paro, etc.), una mujer mayor de 50 años de origen extracomunitario tendrá menor puntuación de pronóstico de encontrar trabajo que un hombre menor de 50 años de origen europeo.

Figura 8: Ejemplo de sistema de puntuación como algoritmo de perfilado del Servicio Público de Empleo austriaco. Adaptado de la documentación oficial del algoritmo del AMS [15].

PUNTUACIÓN
= f (0,10
- 0,14 x GÉNERO FEMENINO
- 0,13 x GRUPO DE EDAD 30 A 49
- 0,70 x GRUPO DE EDAD MAYOR 50
+ 0,16 x GRUPO ESTADOS UE
- 0,05 x GRUPO TERCEROS PAÍSES
+ 0,28 x EDUCACIÓN
+ ...)

Otro de los problemas de este sistema es que a las personas con historial laboral fragmentado les asigna automáticamente al grupo C (bajas perspectivas de encontrar empleo durante al menos 6 meses en los próximos 2 años). Esto impacta directamente en personas jóvenes (sin apenas experiencia), en inmigrantes (sin historial de contratación en Austria) o en personas que se reincorporan al trabajo después de un largo periodo, por ejemplo, mujeres que hayan estado de baja por maternidad y cuidados. Por ello, se considera que este sistema es discriminatorio desde un punto de vista interseccional en el que se penaliza a las personas de grupos poblacionales que no han sido suficientemente dataficados en el pasado para entrar en la población normativa de referencia [16].

Este ejemplo nos trae un debate interesante sobre fundamentos y uso de la estadística. Estos modelos pueden utilizarse para analizar y describir la realidad (analizando los pesos asignados a las variables), pero también para hacer predicciones individuales (obteniendo la suma final de puntos y transformándola en una probabilidad o una categoría). Lo interesante es que los modelos son siempre representaciones estadísticas de una población ajustando el modelo a los datos de ejemplo. Como técnica descriptiva, estos modelos son populares en ciencias sociales para analizar factores sociales, como por ejemplo evidenciar la discriminación interseccional a nivel poblacional que acabamos de exponer. Sin embargo, al utilizar estos mismos modelos para realizar predicciones individuales se reproduce en el presente la realidad discriminatoria pasada recogida en los datos a nivel poblacional.

15 Holl, J., Kernbeiß, G., & Wagner-Pinter, M. (2018). Das AMS-Arbeitsmarktchancen-Modell. *Arbeitsmarktservice Österreich, Wien*. https://ams-forschungsnetzwerk.at/downloadpub/arbeitsmarktchancen_methode_%20dokumentation.pdf

16 Lopez, P. (2019). Reinforcing Intersectional Inequality via the AMS Algorithm in Austria. *Proceedings of the STS Conference Graz 2019*, 21 https://doi.org/10.3217/978-3-85125-668-0-16

3.1.3. Caso SyRI en Países Bajos

En febrero de 2020, un tribunal de Países Bajos paralizó el sistema SyRI (acrónimo de System Risk Indication). SyRI, dependiente del Ministerio de Asuntos Sociales y Empleo, relacionaba y analizaba grandes cantidades de datos personales de la ciudadanía sobre identidad, trabajo, bienes muebles e inmuebles, educación, pensión, negocios, ingresos, patrimonio y deudas con el fin de realizar perfiles de riesgo de fraude al estado.

En 2014 una coalición de organizaciones sociales llevó este sistema a los tribunales por vulnerar derechos básicos de privacidad y discriminación que finalmente terminó con su paralización en 2020. El Tribunal de Distrito de La Haya concluyó que "el modelo de riesgo elaborado en estos momentos por SyRI puede tener efectos no deseados, como estigmatizar y discriminar a la ciudadanía, por la ingente cantidad de información que recoge" [17]. De hecho, el sistema habría sido utilizado fundamentalmente para analizar a la población de barrios de ingresos bajos y así localizar perfiles de "alto riesgo" sin que además se dispusiera de información sobre la cantidad de falsos positivos del sistema, produciendo un efecto estigmatizante [18]. Philip Alston, Relator Especial de las Naciones Unidas para la extrema pobreza y los Derechos Humanos, expresó su preocupación por el sistema SyRI en una carta dirigida al tribunal el 26 de septiembre de 2019: "Barrios enteros se consideran sospechosos y se someten a un escrutinio especial, que es el equivalente digital a que los inspectores de fraude llamen a todas las puertas en una determinada zona y miren los registros de cada persona en un intento de identificar los casos de fraude, mientras que no se aplica tal escrutinio a los que viven en zonas más favorecidas".

El caso de SyRI es un ejemplo de discriminación hacia población mayoritariamente migrante y racializada. Aunque el algoritmo de perfilación en sí no llegase a discriminar, la implementación de este sistema focalizándose especialmente en población migrante y racializada en busca de defraudadores, supone una perfilación racial explícita.

3.1.4. Caso Bristol Integrated Analytical Hub

Uno de los casos más relevantes de datificación del estado de bienestar es el caso del Hub de datos del Ayuntamiento de Bristol en Reino Unido. El ayuntamiento de esta ciudad puso en marcha un sistema inteligente de atención a "familias problemáticas" en el que 35 problemas/asuntos sociales son evaluados algorítmicamente mediante sistemas de puntuación, esto es, se mide el riesgo de ocurrencia de distintos problemas sociales sobre una familia o miembros de esta familia. Por ejemplo, algunos de estos riesgos que promete anticipar para acelerar la intervención social son sufrir violencia doméstica [19], que padres e hijos se vean envueltos en crímenes o comportamientos "antisociales", y que los menores dejen de estudiar, trabajar o formarse.

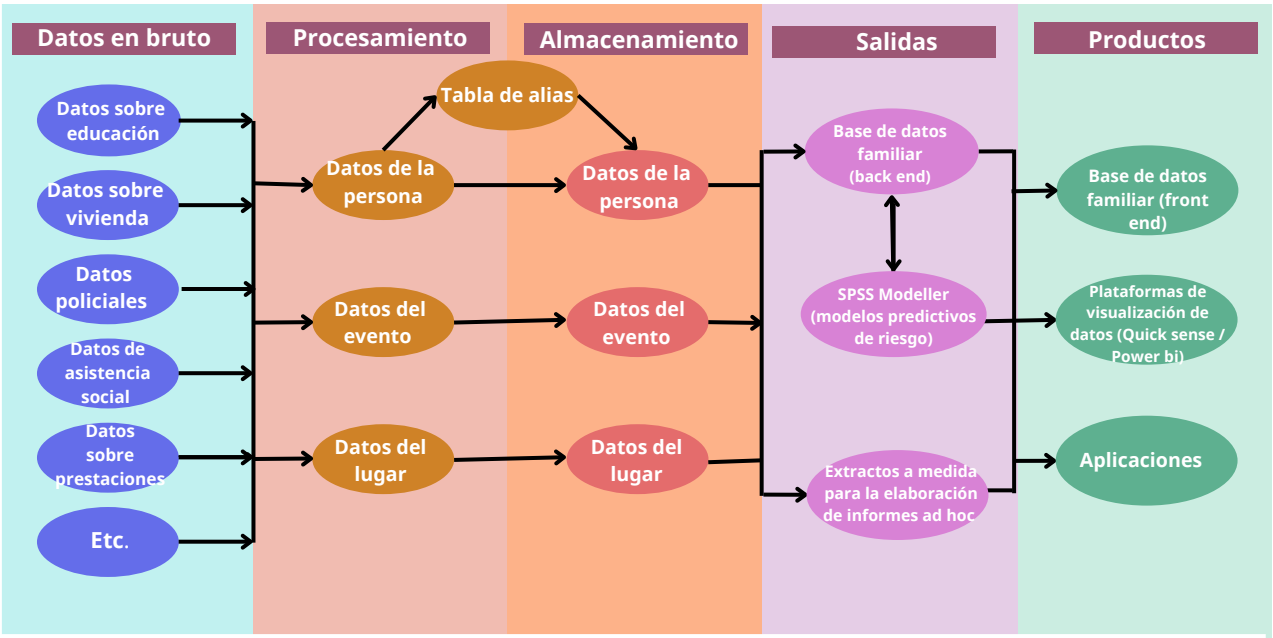
17 Ferrer, I. (12 de febrero de 2020). Países Bajos veta un algoritmo acusado de estigmatizar a los más desfavorecidos. *El País*. https://elpais.com/tecnologia/2020/02/12/actualidad/1581512850_757564.html

18 Vervloesem, K. (2020, April 6). How Dutch activists got an invasive fraud detection algorithm banned. *AlgorithmWatch.* https://algorithmwatch.org/en/syri-netherlands-algorithm/

19 Violencia doméstica es el término usado por el Ayuntamiento de Bristol.

El sistema funciona centralizando la recogida de datos de 30 organismos públicos para monitorizar indicadores de riesgo y vulnerabilidad de más de 54.000 familias de Bristol [20]. Como muestra la Figura 9, el sistema funciona integrando datos policiales, de educación, vivienda, ayudas públicas, etc., para monitorizar y asignar niveles de riesgo a familias, posibles eventos y zonas geográficas. Esto incluye sistemas estadísticos de modelado del riesgo que producen puntuaciones que después se traducen en acciones de los servicios sociales y policiales. Utilizando análisis de "última generación"[21], se toman un conjunto de individuos relativo a un problema (por ejemplo, víctimas de abusos sexuales), se identifican factores comunes que comparten y luego se usan una serie de algoritmos para estudiar la similitud entre este grupo de control y otros casos de la ciudad. De este modo, según los promotores del proyecto, los trabajadores clave están "mejor equipados" para adaptar su enfoque a la gestión de los casos y permite una "comprensión estratégica de la vulnerabilidad".

El caso de Bristol es relevante por la idea de automatizar y enfocar la gestión en torno a la estimación del riesgo con sistemas de puntuación algorítmicos que se centran en datos cuantitativos y descargan la visión contextual, la información no estructurada y una visión longitudinal de las personas más allá de la foto estática que son los datos que les representan a la hora de la toma de decisiones. Esto termina considerando variables como los ingresos, el tipo de vivienda, el número de ausencias escolares y descarga información sobre redes de apoyo familiares y extrafamiliares o la relación con el barrio.



Fuente: Bristol City Council Think Family Data Process Map.

Figura 9: Diagrama que muestra las fuentes de datos que utiliza la herramienta Think Family Data Process Map y cómo las integra para puntuar a las familias en distintas escalas de riesgo que utilizan administraciones, servicios sociales y autoridades locales.

20 Bristol City Council (2022). Insight Bristol and the Think Family Database. Insight Bristol. https://www.bristol.gov.uk/policies-plans-strategies/the-troubled-families-scheme

21 Entrecomillamos expresiones como "última generación", "comprensión estratégica" y demás utilizadas por los promotores de estos proyectos para destacar los beneficios e innovaciones tal y como los conciben estos.

3.2. Educación

3.2.1. Caso Selectividad en Reino Unido

En verano de 2020 el Gobierno británico decidió implementar un algoritmo para predecir las notas de selectividad. Debido a la pandemia, la comunidad estudiantil no pudo realizar el examen presencialmente, así que se decidió el uso de un algoritmo para predecir sus notas. Para entrenar el algoritmo se utilizó la distribución de notas de años anteriores (2017-2019), la posición de cada estudiante en el ranking de su escuela y sus notas para cada asignatura.

No obstante, los políticos no vieron los riesgos y limitaciones que conlleva reemplazar exámenes por algoritmos que analizan patrones en datos históricos. Cuando el alumnado obtuvo el resultado del algoritmo vieron que sus notas habían ido a la baja. Más concretamente, el algoritmo había estimado a la baja las notas de quienes vivían en barrios marginados. Guiado por los patrones de datos históricos, el algoritmo predecía que si en una escuela no hubo nadie brillante en los últimos dos años, difícilmente lo habría ese año.

Varias expertas analizaron los resultados del algoritmo implementado y vieron que, además de reproducir las desigualdades entre institutos de barrios ricos/pobres o institutos grandes/pequeños, la eficiencia del algoritmo era muy baja[22]. Todo ello derivó en la primera manifestación estudiantil en contra de un algoritmo *(ver Figura 10)*. Como resultado, el Gobierno decidió que las notas del algoritmo se daban por invalidas y encargó al profesorado hacer una predicción de la nota de selectividad a sus estudiantes.

Figura 10: Estudiantes británicas manifestándose en contra del algoritmo que predijo sus resultados de selectividad.

22 Ofqual (2020). *Research and analysis Awarding GCSE, AS & A levels in summer 2020: interim report.* GOV.UK https://www.gov.uk/government/publications/awarding-gcse-as-a-levels-in-summer-2020-interim-report

3.2.2. Detección de emociones en el aula

En 2021, una empresa especializada en neuromarketing emprendió un proyecto piloto en el colegio público Blas de Infante de Málaga para analizar las emociones de los menores en clase, mediante el uso, supuestamente, de IA. A través de unas cámaras instaladas en clase, se captaron los niveles de atención y distracción, así como la dirección de la mirada de los niños y niñas de clase, para a posteriori analizar qué les interesa y qué materiales pedagógicos se deberían cambiar [23].

El estudio de la detección de emociones en general ha sido largamente criticado por académicas, organizaciones civiles y activistas [24]. Esta teoría se centra en los estudios del psicólogo americano Paul Ekman que analizó las expresiones faciales de individuos en Papúa Nueva Guinea para establecer cuáles eran las emociones universales, siguiendo con la teoría de Charles Darwin sobre la universalidad de las emociones humanas. Ekman ha sido criticado por normalizar las emociones y gestos con una mirada eurocéntrica y utilizar poblaciones indígenas para corroborar su teoría. Como defiende la académica Lisa Feldman Barrett, las emociones son una construcción cultural y dependen del contexto, por este simple hecho no se pueden establecer emociones universales. Además, después de repasar más de 1000 trabajos académicos, se concluyó que no existe fundamento científico que pueda demostrar que mediante microexpresiones faciales se pueda inferir la emoción [25].

Este es un claro ejemplo de tecnosolucionismo, la búsqueda de soluciones basadas en sistemas automatizados para dar pábulo a la teoría de las emociones universales, algo que resulta aún más cuestionable en contextos tan sensibles como lo son las aulas de las escuelas públicas. Si se quiere conocer la opinión de los niños y niñas en clase, no hace falta instalar un dispositivo que supuestamente analice sus emociones y su nivel de atención, algo que además presenta riesgos relacionados con la privacidad. Se pueden hacer otras intervenciones, como preguntar qué materiales les parecen más entretenidos.

3.3. Policía predictiva

Se suele hablar de "policía predictiva" para referirse al uso de datos históricos de fuentes policiales para "predecir" futuros crímenes y gestionar los recursos policiales acordemente [26]. Esta denominación, sin embargo, resulta bastante pretenciosa ya que sugiere la posibilidad de predecir el comportamiento de los criminales a nivel individual y la identificación de mecanismos causales precisos. Algunos sistemas se centran en la estimación de riesgo a ni-

23 Inteligencia artificial en Málaga para saber si los alumnos están atentos. (10 de junio de 2021). Canal Sur.es. https://www.canalsur.es/noticias/andaluc%C3%ADa/malaga/inteligencia-artificial-en-malaga-para-saber-si-los-alumnos-estan-atentos/1723841.html

24 Reconocimiento de emociones: una aplicación de inteligencia artificial sin madurar que ya se usa para predecir la personalidad de la gente. (2021). *Maldita.es* https://maldita.es/malditatecnologia/20210608/reconocimiento-emociones-inteligencia-artificial/

25 Ídem

26 ¿Se puede utilizar la inteligencia artificial para prevenir delitos? (2021). *Maldita.es* https://maldita.es/malditatecnologia/20210306/puede-utilizar-inteligencia-artificial-evenir-delitos/

vel individual, por ejemplo de sufrir agresiones o riesgo de reincidencia, mientras que otros sistemas realizan estimaciones sobre “dónde” y “cuándo” sucederá un crimen. Como veremos en los siguientes ejemplos, el concepto de predicción “dónde” y “cuándo” es bastante difuso e impreciso, ya que en general estos sistemas realizan predicciones poblacionales, no individuales y con escalas espaciotemporales bastante amplias [27]. En el contexto español se utilizan algunos de estos sistemas. Existen y han existido numerosos programas experimentales[28].

3.3.1. Caso COMPAS

El algoritmo COMPAS es sin lugar a duda uno de los casos más conocido de la discriminación racial algorítmica gracias a la denuncia del periódico estadounidense ProPuplica. El sistema COMPAS se utiliza en varios lugares de Estados Unidos para analizar el riesgo de que un preso vuelva a delinquir. Para ello, se hacen 137 preguntas de cuestiones sociodemográficas, violencia familiar sufrida, estados de ánimo, etc., además de añadir información sobre el historial de condenas de la persona. Con estos datos, COMPAS hace una estimación del riesgo de reincidencia ubicando a la persona presa en 10 niveles de riesgo. Los etiquetados con mayor nivel de riesgo se considera que deben permanecer en prisión. A pesar de que es un juez o una jueza quien tendrá la última palabra, COMPAS tiene una gran influencia sobre la decisión judicial.

En 2016, este periódico realizó una investigación incluyendo la reconstrucción parcial del sistema COMPAS en la que se demostraba el comportamiento racista de este algoritmo [29]. El estudio evidenció que, frente a los presos blancos, el sistema etiquetaba con "mayor riesgo" de reincidencia a presos afroamericanos que después no reincidían. Como consecuencia de este tipo de errores, denominados falsos positivos, un preso afroamericano que optase por la libertad condicional tenía aproximadamente el doble de opciones de obtener una denegación. La Figura 11 muestra la diferencia en la tasa de errores de detección de reincidencia. De igual manera, el sistema calificaba erróneamente con "bajo riesgo" al doble de presos blancos que a los afroamericanos.

27 Fuck the police. Desmontando las IA de la policía. (1 de abril de 2021). Post Apocalipsis Nau, programa 52. (Audio podcast). El Salto Diario. https://www.elsaltodiario.com/post-apocalipsis-nau/post-apocalipsis-nau-52-hack-the-police-demontando-las-ia-de-la-policia

28 González-Álvarez, J. L., Hermoso, J. S., & Camacho-Collados, M. (2020). Policía predictiva en España. Aplicación y retos futuros. Behavior & Law Journal, 6(1), 26–41. https://doi.org/10.47442/blj.v6.i1.75

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

La predicción falla de manera diferente para los acusados afroamericanos

	Blanco	Afroamericano
Etiquetados como de mayor riesgo, pero no reincidieron	23,5%	44,9%
Etiquetados como de menor riesgo, pero reincidieron	47,7%	28%

En general, la herramienta de evaluación de Northpointe predice correctamente la reincidencia en el 61% de las ocasiones. Pero los negros tienen casi el doble de probabilidades que los blancos de ser etiquetados como de alto riesgo pero no vuelven a reincidir. Entre los blancos se produce el error contrario: son mucho más propensos que los negros a ser etiquetados como de menor riesgo pero después pasan a cometer otros delitos. (Fuente: análisis de ProPublica de los datos del condado de Broward, Florida)

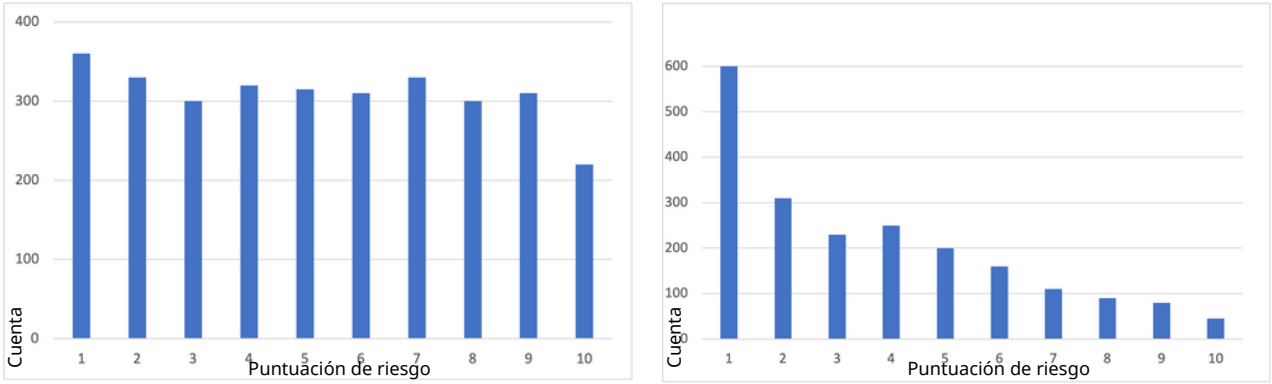
Figura 11: Diferentes errores de predicción de riesgo de reincidencia desagregados por raza en el informe de COMPAS. Fuente: ProPublica [30]**.**

Una de las conclusiones preliminares del estudio es que la base de datos histórica de sentencias judiciales del sistema penitenciario estadounidense tenía un importante sesgo debido a una mayor presencia de criminalidad entre afroamericanos que entre blancos, lo que se traducía en una mayor presencia de personas afroamericanas etiquetadas con mayores índices de riesgo. La conclusión preliminar es que cualquier sistema estadístico o de IA que se entrene con estos datos inmediatamente "aprenderá" que una persona afroamericana tiene más probabilidades de ser un delincuente peligroso que una persona blanca, cuando estos datos reflejan una realidad mucho más compleja de racismo estructural y clases sociales. Por ejemplo, no se tiene en cuenta que altos índices de criminalidad tienen que ver con la hiperfocalización de las instituciones policiales en determinados barrios y comunidades racializadas que a su vez son quienes, debido al racismo estructural, cuentan con mayores índices de pobreza. En la Figura 12 se muestra la proporción de personas de cada grupo racial asignada a cada nivel de peligrosidad, siendo 1 el nivel más bajo.

30 Larson, J., & Angwin, J. (23 de mayo de 2016). How We Analyzed the COMPAS Recidivism Algorithm. *ProPublica.* https://www.propublica.org/article/how-we-analyzed-the-compas-recidivism-algorithm

Puntuación de riesgo de los acusados afroamericanos

Puntuación de riesgo de los acusados blancos



Estos gráficos muestran que las puntuaciones de los acusados blancos estaban sesgadas hacia las categorías de menor riesgo. Las puntuaciones de los acusados negros no lo estaban. (Fuente: análisis de ProPublica de los datos del condado de Broward, Florida).

Figura 12: Visualización de la composición de la base de datos COMPAS indicando la distribución de niveles de peligrosidad de afroamericanos y blancos.

El caso COMPAS es quizás el mejor documentado y estudiado en lo que respecta a la cuestión de la discriminación y equidad en sistemas de aprendizaje automático[31]. Sin embargo, es interesante recordar que ninguno de los equipos que lo auditaron tuvieron acceso a los datos reales ni al código o los modelos matemáticos del sistema. Por lo tanto, sus conclusiones se basan en la reproducción de los experimentos con escasez de recursos. El sistema real que evalúa los riesgos de las personas reales sigue siendo desconocido.

En todo caso, la publicación de ProPublica supuso un antes y un después en el análisis de la discriminación de los sistemas de decisión automática con sus puntos fuertes y flacos. En el punto fuerte, inició el debate y el reanálisis de muchos sistemas en uso en busca de sesgos discriminatorios de forma similar y a un ingente número de activistas e investigadores que han tratado de mitigar sesgos en estos. Sin embargo, el excesivo énfasis en la concepción de la discriminación racial como una cuestión de disparidades métricas estadísticas ha supuesto, a su vez, una visión muy limitada de las diferentes formas en las que la tecnología puede implementar o jugar un papel en la discriminación racial que ha sido ampliamente contestada desde la academia [32,33] y el activismo[34].

31 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/

32 Hoffmann, A. L. (2019). *Where fairness fails: Data, algorithms, and the limits of antidiscrimination discourse*. Information, Communication & Society, 22(7), 900–915. https://doi.org/10.1080/1369118X.2019.1573912

33 Gangadharan, S. P., & Niklas, J. (2019). *Decentering technology in discourse on discrimination*. Information, Communication & Society, 22(7), 882–899. https://doi.org/10.1080/1369118X.2019.1593484

34 Stop LAPD Spying Coalition, Free Radicals, & 2020. (2 de marzo de 2020). The Algorithmic Ecology: An Abolitionist Tool for Organizing Against Algorithms. *Free Rads*. https://freerads.org/2020/03/02/the-algorithmic-ecology-an-abolitionist-tool-for-organizing-against-algorithms/

3.3.2. Caso PredPol

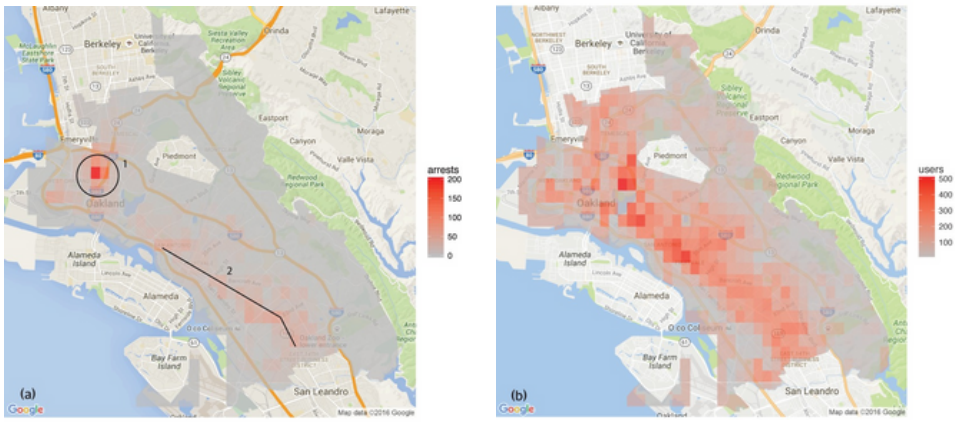


Figura 13: Comparación del mapa de posibles delitos "predichos" por PredPol con el mapa de posibles consumidores de droga. A pesar de que el consumo y delitos de droga están distribuídos ampliamente en la localidad (derecha), PredPol se centra en "detectar" crímenes en zonas con habitantes racializados (izquierda). Fuente Lum & Isaac (2016).

PredPol es un algoritmo -y el nombre de la empresa que lo desarrolla-, que dice ser capaz de prever dónde y cuándo es más probable que suceda un crimen, de manera que se espera que la policía patrulle las zonas indicadas por el algoritmo. Está basado en un modelo matemático de propagación de eventos sísmicos bajo la hipótesis de que el crimen es una actividad recurrente y que tiene unos patrones de expansión temporal y espacial. Aunque la efectividad de PredPol está cuestionada tras algunos planes pilotos [35], el sistema sigue siendo usado en muchas partes de Estados Unidos.

A parte de su posible inefectividad en la disminución de crímenes, PredPol ha sido denunciado por dirigir desproporcionadamente las patrullas policiales fundamentalmente a barrios de comunidades afroamericanas o latinas[36]. Este comportamiento de discriminación y criminalización racial fue evidenciado en el artículo científico *To predict and serve* [37]. Para ello, contrastaron un mapa de consumidores de droga según datos y modelos sociológicos de consumo con el mapa de predicciones de delitos relacionados con drogas generado por PredPol (ver Figura 13). Mientras que el consumo de drogas se reparte por muchos barrios con habitantes de variados perfiles raciales y socioeconómicos, PredPol sólo "detecta" crímenes en barrios de comunidades racializadas.

El caso PredPol nos pone de relevancia dos problemas de fondo de los sistemas, mal llamados, de "policía predictiva"[38]. En primer lugar, no existe una base de datos de crímenes totales, sólo de los denunciados o descubiertos por la policía, la cual refleja sus patrones y

35 Leila Miller (21 de abril de 2020). LAPD data programs need better oversight to protect public, inspector general concludes. *Los Angeles Times.* https://www.latimes.com/local/lanow/la-me-ln-lapd-data-20190312-story.html

36 Sankin, A., Mehrotra, D., Mattu, S., & Gilbertson, A. (2021). Crime Prediction Software Promised to Be Free of Biases. New Data Shows It Perpetuates Them. *The Markup and Gizmodo* https://gizmodo.com/crime-prediction-software-promised-to-be-free-of-biases-1848138977

37 Lum, K., & Isaac, W. (2016). To predict and serve?. *Significance. 13*(5), 14–19. https://doi.org/10.1111/j.1740-9713.2016.00960.x

38 Ídem

sesgos de actuación. En segundo lugar, la predicción de crimen en un área enviará recursos policiales a ese área, lo que hará que sea más probable que se encuentre un crimen y este se incorpore a la base de datos policial, reforzando así el sesgo a través de bucles de retroalimentación. Al mismo tiempo, es menos probable que se observen eventos que contradigan las predicciones, por ejemplo, evitando investigar crímenes relacionados con las drogas en barrios no racializados y de clases medias [39].

3.3.3. Caso London GangMatrix

La policía metropolitana de Londres lleva años utilizando sistemas de decisión algorítmica para perfilar a personas que podrían ser parte de una banda. En 2012 lanzaron el sistema Gangs Matrix que consiste en una base de datos de jóvenes londinenses que la policía sospecha que pertenecen a una banda callejera. La herramienta se enmarca dentro de los planes de la capital de Reino Unido para luchar contra los crímenes violentos.

Según un informe de Amnistía Internacional, este es "un sistema racialmente discriminatorio que estigmatiza a los jóvenes negros por la música que escuchan o su comportamiento en las redes sociales", que no sirve a su supuesto objetivo de luchar contra los crímenes violentos [40] . Aún en 2020, Amnistía Internacional indicaba que siguen existiendo "serias dudas sobre cómo se incluye a las personas en la base de datos, cómo se comparte la información con otros organismos y el impacto negativo en los jóvenes negros, quienes están desproporcionadamente representados en ella". El informe Trapped in the Gangs Matrix (Atrapados en la Matriz de Bandas) asegura que el 78% de los jóvenes identificados como miembros de bandas son negros y que el 75% han sido ellos mismos las víctimas de la violencia de bandas [41] .

3.3.4. RisCanvi

Esta herramienta fue introducida por el Departament de Justícia de la Generalitat de Cataluña en 2009 debido a un incremento de presos en las cárceles catalanas [42]. RisCanvi tiene el objetivo de mejorar las predicciones individuales del riesgo de violencia y ayudar a la toma de decisión del personal funcionario de prisiones. Mediante una entrevista, el personal funcionario evalúa diferentes factores de riesgo que se agrupan en 5 categorías: criminal, biográfico, social, médico y psicológico. Cuando estos factores son evaluados, RisCanvi estima el riesgo de reincidencia. Algunos de los factores que se encontraron como determinantes de la reincidencia son: la edad, historial familiar de crimen, exclusión social, actitudes pro criminales, etc.

---

39 Lum, K., & Isaac, W. (2016). To predict and serve?. *Significance*. 13(5), 14–19. https://doi.org/10.1111/j.1740-9713.2016.00960.x

40 Amnesty International UK. (2020, mayo 18). *What is the Gangs Matrix?* https://www.amnesty.org.uk/london-trident-gangs-matrix-metropolitan-police

41 ídem

42 Andrés-Pueyo, A., Arbach-Lucioni, K., & Redondo, S. (2018). *The RisCanvi: a new tool for assessing risk for violence in prison and recidivism. Recidivism Risk Assessment: A Handbook for Practitioners*, 255-268. https://onlinelibrary.wiley.com/doi/10.1002/9781119184256.ch13

La precisión que muestra RisCanvi en casos de alto riesgo es muy baja según el informe público de 2016 de la Generalitat[43]. Un 94,6% de los casos son falsos positivos, es decir RisCanvi los etiqueta como de alto riesgo, pero realmente no lo son. Recientemente, académicos han analizado el sesgo de esta herramienta según la nacionalidad y la edad, y propuesto soluciones técnicas para su mitigación, aunque estas no se han incorporado al sistema[44]. RisCanvi ha sido denunciado por falta de transparencia que permitiría hacer análisis sobre diferentes ejes de discriminación ya que se sabe, por ejemplo, que no está diseñado para evaluar delitos de guante blanco y asigna riesgos bajos a las personas que han cometido estos delitos[45].

3.3.5. VeriPol

Este algoritmo, diseñado por el Ministerio de Interior y académicos[46], tiene como fin la detección de denuncias falsas. Con el objetivo de automatizar la detección de denuncias fraudulentas, este algoritmo se propone como una solución para ahorrar recursos al cuerpo policial. Para ello, un solo policía etiquetó las denuncias (en verdaderas o falsas) que se utilizaron para entrenar el algoritmo. Después, se usaron varios modelos algorítmicos y se analizaron qué palabras eran más comunes en denuncias que ese policía había etiquetado como falsas o verdaderas. Por ejemplo, palabras como “Negro”, “Mochila, “Bolso” son más falsas; mientras que palabras como “Barba”, “Policia”, “Chino” son más veraces.

En el artículo, también se muestra que el algoritmo fue implementado en dos ciudades españolas como proyecto piloto: Murcia y Málaga. Los autores explican que el algoritmo fue capaz de clasificar correctamente denuncias falsas en un 83.54%. Es importante aclarar que este porcentaje de aciertos lo es sobre el etiquetado de un único policía que etiquetó la base de datos de entrenamiento, pero no necesariamente se mantiene en la realidad ni disponemos de datos sobre su efectividad real. A pesar de que, como se cita en el artículo, el agente en particular tiene mucha experiencia en detectores de mentiras e interrogatorios policiales, esta manera de diseñar el etiquetado de los datos hace que el algoritmo clasifique según la percepción de este policía. Hoy en día VeriPol es usado en comisarías de todo el territorio nacional para predecir la probabilidad de que una denuncia sea falsa.

---

43 Capdevila Capdevila, M., Ferrer Puig, M., Blanch Serentill, M., Framis Ferrer, B., Garrigós Bou, A., & Comas López, N. (2017). *Estudio de la reincidencia en las excarcelaciones de alto riesgo* (2010-2013). Generalitat de Catalunya. Centre d'Estudis Jurídics i Formació Especialitzada. cejfe.gencat.cat/web/.content/home/recerca/cataleg/crono/2017/reincidenciaexcarceracions/resumen_reincidencia_excarcelaciones.pdf

44 Karimi-Haghighi, M. and Castillo, C. (junio de 2021). *Enhancing a recidivism prediction tool with machine learning: effectiveness and algorithmic fairness.* In Proceedings of the Eighteenth International Conference on Artificial Intelligence and Law (pp. 210-214). https://dl.acm.org/doi/10.1145/3462757.3466150

45 Saura, G. y Aragó, L. (7 de diciembre de 2021). La Vanguardia https://www.lavanguardia.com/vida/20211207/7911428/algoritmo-prisiones-rinde-cuentas-nadie.html

46 Quijano-Sánchez, L., Liberatore, F., Camacho-Collados, J., & Camacho-Collados, M. (2018). *Applying automatic text-based detection of deceptive language to police reports: Extracting behavioral patterns from a multi-step classification model to understand how we lie to the police.* Knowledge-Based Systems, 149, 155-168. https://www.sciencedirect.com/science/article/abs/pii/S095070511830128X?via%3Dihub

3.4. Violencia de género

La violencia de género es otro de los contextos en los que también se han entrometido los sistemas de decisión automática. En esta sección vamos a repasar tres sistemas: VioGen, la EPV-R y la Plataforma Tecnológica de Intervención Social.

3.4.1. VioGén

El Sistema de Seguimiento Integral en los casos de Violencia de Género (VioGén) [47] es un sistema de predicción de riesgo de violencia de género implementado en el Estado español (excepto en Cataluña y el País Vasco) desde 2007. Fue desarrollado por el Ministerio del Interior y su función es informar a Guardia Civil y Policía Nacional del nivel de riesgo de violencia "contra la mujer" para mejorar la efectividad de detección de casos de riesgo.

Esta herramienta recoge información diversa sobre casos de violencia de género [48]: datos personales del presunto agresor y de la mujer (nivel educativo, formación, situación laboral, estado civil), características del presunto agresor (registro de delitos pasados u órdenes de alejamiento previas) y características de la víctima (como el apoyo que recibe o la asistencia en centros de acogida).

Esta herramienta ha sufrido diferentes revisiones. En 2007, VioGén se componía de una encuesta formada por 20 ítems y se fueron incorporando hasta 65 indicadores para predecir con mayor precisión la reincidencia en casos de violencia de género [49]. Recientemente este sistema fue analizado externamente por la Fundación Éticas[50]. Algunos problemas asociados a VioGén serán analizados en la Sección 3.7.

3.4.2. EPV-R

Como hemos explicado anteriormente, VioGen no se aplica en todo el territorio nacional. En el País Vasco, la Ertzaintza usa una herramienta llamada Escala de Predicción de riesgo de Violencia grave contra la pareja (EPV), la cual fue posteriormente revisada (EPV-R). Como VioGén, la EPV-R se basa en una encuesta de 20 indicadores que recogen información sobre: datos personales (nacionalidad del agresor o de la agredida), estado de la relación de pareja (separados, divorciados), tipo de violencia (existencia de violencia física, sexual, intención de causar lesiones o heridas), perfil del maltratador (celopatía, comportamiento violento, salud mental, abuso de alcohol/drogas), vulnerabilidad de la víctima (percepción del riesgo de muerte, enfermedad).

47 Ministerio del Interior. (2022). Sistema VioGén. https://www.interior.gob.es/opencms/es/servicios-al-ciudadano/violencia-contra-la-mujer/sistema-viogen/

48 Álvarez, J.L.G., Ossorio, J.J.L., Urruela, C. and Díaz, M.R. (2018). *Integral Monitoring System in Cases of Gender Violence VioGén System.* Behavior & Law Journal, 4(1). https://www.behaviorandlawjournal.com/BLJ/article/view/56

49 López-Ossorio, J. J., Álvarez, J. L. G., Pascual, S. B., García, L. F., & Buela-Casal, G. (2017). *Risk factors related to intimate partner violence police recidivism in Spain.* International Journal of Clinical and Health Psychology, 17(2), 107-119. https://www.sciencedirect.com/science/article/pii/S1697260017300017

50 Éticas Foundation. (8 de marzo de 2022). Auditoría externa del sistema VioGén. https://eticasfoundation.org/es/gender/the-external-audit-of-the-viogen-system/

Esta herramienta se utiliza en procesos de decisión judiciales para establecer "medidas de protección a la víctima". Los juzgados reciben un informe con la descripción del caso, en el que un apartado refiere a la valoración del algoritmo (riesgo bajo, moderado o alto).

3.4.3. Plataforma Tecnológica de Intervención Social

En 2015, en la municipalidad de Salta (Argentina) se propuso un algoritmo que detectara el fracaso escolar y el embarazo adolescente. La base de datos para entrenar dicho algoritmo consta de 12000 entradas de niñas que comprenden las edades de entre 10 y 19 años e incluye información como la edad, el barrio, la etnicidad, nivel de educación de los padres, país de origen, diversidad funcional, o incluso si tiene acceso a agua caliente en su hogar.

El sistema fue diseñado por el Ministerio de la Primera Infancia de Salta y la empresa tecnológica Microsoft. Según el gobernador de esta municipalidad de Argentina, el sistema tenía una capacidad predictiva del 86%[51] . No obstante, se encontraron errores estadísticos en el sistema algorítmico que cuestionan la precisión del sistema[52]. De hecho, la plataforma digital que mostraba el código del algoritmo ya no existe[53].

3.5. Fronteras digitales, migraciones y refugio

Desde los atentados del 11 de Septiembre de 2001, la tecnología se ha utilizado en el contexto fronterizo con el objetivo de mejorar la seguridad, proteger a la ciudadanía y detectar terroristas de una manera más eficaz. La inteligencia artificial y los sistemas biométricos han sido las soluciones desarrolladas para identificar terroristas de una manera automática en aeropuertos. Hoy en día, en algunas de las fronteras de los países occidentales se han implementado máquinas de reconocimiento de huellas dactilares y hay previsiones de implementar el reconocimiento facial en Europa para todo ciudadano no europeo que cruce fronteras dentro del espacio Schengen. Debido a la regulación de Dublín, desde el año 2000, todos los migrantes que llegan a Europa deben de ser registrados mediante sus huellas dactilares para solicitar asilo. No obstante, dentro de la investigación también se financian proyectos pilotos que "mejoren" la experiencia de cruzar una frontera (hecho que solo está al alcance de pasaportes privilegiados).

---

51 La inteligencia que no piensa (21 de abril de 2018). Página 12. https://www.pagina12.com.ar/109080-la-inteligencia-que-no-piensa.

52 Peña, P. y Varon, J. (5 de marzo de 2021). Teenager pregnancy addressed through data colonialism in a system patriarchal by design. *Notmy.ai* https://notmy.ai/news/case-study-plataforma-tecnologica-de-intervencion-social-argentina-and-brazil/.

53 Plataforma donde orginalmente se mostraba el código del algoritmo. https://github.com/facundod/casestudies/blob/master/Prediccion%20de%20Embarazo%20Adolescente%20con%20Machine%20Learning.md.

3.5.1. Bases de datos biométricas de la UE: EURODAC, VIS, SIS II y EES

Dentro del marco Europeo, existen diferentes sistemas y bases de datos[54] que recogen información sobre las personas que cruzan fronteras y sus movimientos dentro del territorio UE:

- EURODAC (por su nombre en inglés EUROpean asylum DACtyloscopic database): Esta base de datos determina qué país (Estado miembro) es responsable de la demanda de asilo. Según la legislación, las autoridades fronterizas deben recoger las huellas dactilares de las personas migrantes, las cuales se agrupan en 3 categorías: Demandantes de asilo (Categoría 1), individuo que cruza una frontera de manera "ilegal" (Categoría 2), individuo que permanece de manera "irregular" en un Estado miembro (Categoría 3). Esta base de datos refuerza la Regulación de Dublín, la cual establece que los demandantes de asilo no pueden viajar dentro del espacio Schengen, sino que deben permanecer en el país responsable de su demanda de asilo. Así, se limita la libre circulación, que es un derecho fundamental. Actualmente, se recogen 10 huellas dactilares en mayores de 14 años y el país responsable de la demanda de asilo. No obstante, la Comisión Europea quiere bajar la edad mínima a 6 e implementar el reconocimiento facial en los próximos años.

- VIS (por su nombre en inglés Visa Information System): Esta base recoge información de los visados de corta estancia, ya sean aceptados o rechazados. Actualmente, recoge imágenes de 10 huellas dactilares e información biográfica.

- SIS II (por su nombre en inglés Schengen Information System): Esta base de datos recoge imágenes de objetos y personas que hayan tenido relación con causas judiciales. Recoge muestras de DNA, información biográfica, imágenes como coches, armas, tatuajes y fotografías de personas.

- EES (por su nombre en inglés Entry Exit System): Esta base de datos, que aún no es operativa [55], recogerá todos los movimientos entre fronteras de personas ciudadanas de países que soliciten un visado para entrar dentro del Espacio Schengen. Recogerá sus huellas dactilares (4), fotografía para reconocimiento facial y datos biográficos. El coste estimado de dicha base de datos asciende a más de 400 millones de euros, el mayor gasto en una base de datos biométrica en la UE. A partir de 2022 este sistema se desplegará en la frontera sur en las ciudades de Ceuta y Melilla[56].

54 PICUM and Statewatch (2019). *Data Protection, Immigration, Enforcement and Fundamental Rights. What the EU's Regulations on Interoperability Mean for People with Irregular Status.* https://picum.org/wp-content/uploads/2019/11/Data-Protection-Immigration-Enforcement-and-Fundamental-Rights-Full-Report-EN.pdf.

55 En la última reunión de la agencia eu-LISA con la industria, estimaron su puesta en funcionamiento para mayo de 2023, sumando varios años de retraso.

56 El Consejo de Ministros aprueba las obras de la frontera inteligente. (4 de octubre de 2022). *El faro de Ceuta.es* https://elfarodeceuta.es/consejo-ministros-obras-frontera-inteligente

- ETIAS: Esta base de datos, que tampoco está aún operativa[57], recogerá datos de personas cuya ciudadanía esté exenta de solicitar un visado y entren dentro del Espacio Schengen para una estancia corta (menos de 90 días dentro de una ventana de 180 días). Los tipos de datos que se recogerán, según la legislación europea de esta base de datos son: nombre, apellidos, domicilio, edad, nivel de educación, profesión y nacionalidad. Este sistema contendrá también una lista de características de viajeros sospechosos de la que aún no se sabe nada[58]. Un algoritmo procesará dichas características, alertando cuando un viajero es sospechoso o no[59]. Después, agentes de la agencia de fronteras Frontex decidirán si la persona necesita un escrutinio más exhaustivo. Además, otras agencias como Interpol o Europol también tendrán acceso a ETIAS.

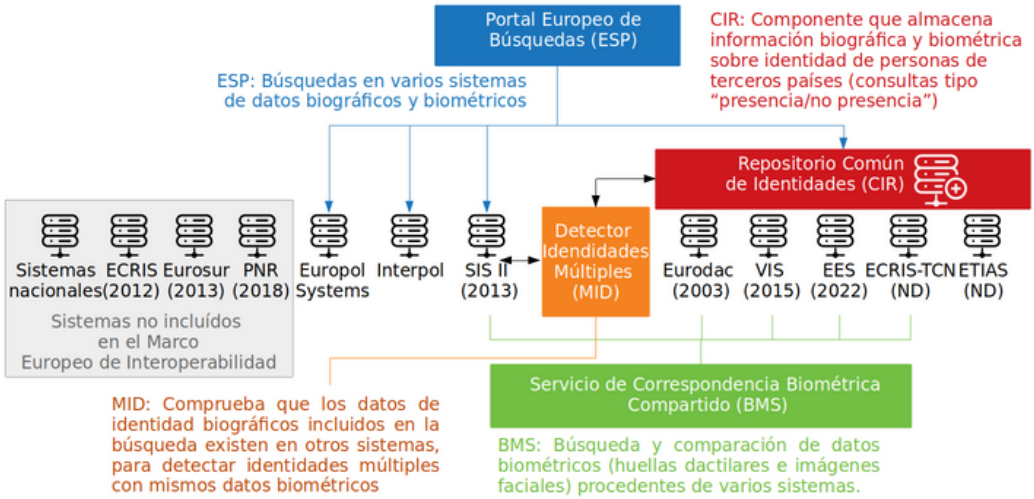


Figura 14: Infraestructura de bases de datos para el control migratorio dentro del espacio Schengen. Fuente: Elaboración propia a partir de la documentación oficial de la UE.

Uno de los problemas que se señalan de estas bases de datos es la falta de auditabilidad y vulneración de derechos, como la libre circulación de personas. No existe dentro de la UE una organización independiente que evalúe el rendimiento de los sistemas biométricos de estas bases de datos. Estos sistemas pueden cometer errores a la hora de identificar a personas, y aunque en los últimos años la tasa de error ha decrecido considerablemente, el debate no debería centrarse en la transparencia o tasa de error de dichos sistemas. La Regulación de Dublín limita la libertad de movimiento de personas migrantes dentro del espacio europeo, y es mediante la tecnología que se refuerza esa ley. Es por ello que se debería cuestionar el contexto y la legislación en las que estos sistemas se implementan, más que cuestionar la transparencia y error algorítmico. Además, las instituciones justifican el uso de estos sistemas en el control migratorio en pro de la seguridad y la eficiencia. Existe una necesidad en el territorio nacional de analizar cómo afectan tanto estas bases de datos, así como las nuevas que se van a implementar para 2023 en las personas migrantes.

57 En la misma reunión, la agencia europea eu-LISA anunció que se pondría en funcionamiento en noviembre de 2023.

58 EU: One step closer to the establishment of the permission - to - travel scheme. (19 de marzo de 2021). *State watch.org* https://www.statewatch.org/news/2021/march/eu-one-step-closer-to-the-establishment-of-the-permission-to-travel-scheme/

59 Ver Artículo 33 de REGULATION (EU) 2018/1240 OF THE EUROPEAN PARLIAMENT AND OF THE COUNCIL

Además, numerosas organizaciones de la sociedad civil [60] cuestionan y enmarcan estos sistemas de control biométrico en una lógica (neo)colonial que moldea las fronteras exteriores de la UE con el sur global. Mecanismos de control fronterizo criticados por criminalizar[61] y deshumanizar a las personas migrantes y por su contribución al aumento de las muertes en la ruta migratoria[62]. Hablamos de la implementación de IA no para facilitar vías seguras a la movilidad humana, sino más bien para contribuir a incrementar las ya enormes cifras de víctimas y vulneraciones de derechos humanos.

3.5.2. iBorderCtrl

Este proyecto [63] fue financiado con más de 4.5 millones de euros por la UE [64] a un conjunto de universidades, empresas privadas y autoridades fronterizas. El objetivo de iBorderCtrl era el de "permitir un control fronterizo más rápido y exhaustivo de los nacionales de terceros países que cruzan las fronteras terrestres de los Estados miembros de la UE, con tecnologías que adopten el futuro desarrollo de la gestión de fronteras de Schengen". Uno de los objetivos específicos de este proyecto fue el de beneficiar a viajeros de 'buena fe' para facilitar su "experiencia" al cruzar fronteras, algo que en el contexto del proyecto se interpreta como cruzar de forma más rápida y cómoda.

Pero entonces, ¿quiénes son los viajeros de 'mala fe'? Según la descripción del proyecto, serían personas que "mienten" sobre los intereses de su viaje, que ocultan como viajes cortos de turismo o visita a amigos aquellos viajes en los que en realidad pretenden quedarse más tiempo del permitido en la UE o que transportan sustancias prohibidas.

Dentro de este proyecto, se experimentó con la idea de crear un sistema de decisión automática que fuera capaz de decidir cuándo un viajero estaba mintiendo o no en un interrogatorio basándose en el análisis de microexpresiones faciales. Es decir, basándose en la teoría de Ekman (Sección 3.2.2). Así, el equipo investigador de iBorderCtrl propuso entrenar una red neuronal basándose en un experimento con actores, los cuales tenían el rol de mentir o decir la verdad sobre cuestiones como la duración del viaje y el lugar donde se alojarían. Este experimento estuvo basado en hipótesis pseudocientíficas que asumen la idea que se puede detectar mentiras mediante expresiones faciales.

60 La implantación de la inteligencia artificial en frontera y la vulneración de derechos. (Enero de 2022). *Fronterasdigitales.com* https://fronterasdigitales.wordpress.com

61 Sánchez M., J. y Valdivia, A. (20 de abril de 2022). EURODAC: un sistema biométrico para categorizar y criminalizar a esos migrantes y refugiados que no queremos. *Algorace.org* https://algorace.org/2022/04/20/eurodac-un-sistema-biometrico-para-categorizar-y-criminalizar-a-esos-migrantes-y-refugiados-que-no-queremos/

62 Naciones Unidas. Asamblea General. (2020): *Informe de la Relatora Especial sobre las formas contemporáneas de racismo, discriminación racial, xenofobia y formas conexas de intolerancia.* https://documents-dds-ny.un.org/doc/UNDOC/GEN/N20/304/57/PDF/N2030457.pdf?OpenElement

63 Proyecto iBorderCtrl (2022). iBorderCtrl.eu https://www.iborderctrl.eu/.

64 Comisión Europea. Intelligent Portable Border Control System. *Cordis.europea.eu* https://cordis.europa.eu/project/id/700626.

Además, el experimento publicado carece de rigurosidad, mostrando limitaciones estadísticas[65] en su supuesta función de encontrar viajeros de 'mala fe' que hacen que iBorderCtrl no pueda funcionar en la práctica. Según el análisis de Sánchez-Monedero & Dencik (2020)[66] , este tipo de proyectos, por tanto, no debe cuestionarse sólo desde su supuesta funcionalidad, sino que debemos entender cuál es la función de estos en la creación de sujetos políticos y la gestión de las poblaciones. Esta función no es simplemente técnica, por el contrario, forma parte de un modelo de gobernanza que construye y determina cada vez más las oportunidades de vida y los derechos fundamentales.

3.5.3. Sistema de visados de UK (Streaming Tool)

La Streaming Tool fue una SDA implementada por el Ministerio de Interior de Reino Unido. Este algoritmo decidía automáticamente qué visados necesitaban un escrutinio más exhaustivo en tres niveles: color rojo (escrutinio exhaustivo, alta probabilidad de ser rechazado), color ámbar (escrutinio moderado, probabilidad media de ser rechazado), color verde (escrutinio leve, probabilidad baja de ser rechazado).

En el otoño de 2017, el Consejo Conjunto para el Bienestar de los Inmigrantes y FoxGlove, una organización inglesa que lucha por los derechos digitales y una tecnología al servicio de todas las personas, demandaron a la Home Office. Alegaban que el sistema era racista, ya que las personas africanas tenían mayor probabilidad de obtener el nivel rojo, por lo tanto, mayor probabilidad de obtener un visado denegado. Tres años más tarde, en 2020, el Tribunal de Justicia concluyó que el sistema reproducía sesgos y pidió al Ministerio de Interior la interrupción de dicha SDA.

3.6. Buscadores y sistemas de recomendación

La cuestión de la representación y promoción de diferentes tipos de personas, valores, conocimientos e información en Internet es tan antigua como la propia red y la historia de buscadores y recomendadores de contenidos. La infrarrepresentación, representación estereotipada o representación inferiorizante de población racializada siempre ha estado presente en los contenidos de Internet. Desde los primeros años en los que fueron producidos mayormente desde culturas occidentales y por perfiles normativos, hasta más tarde con la incorporación de medios de comunicación tradicionales a Internet (~2000s) y redes sociales (~2008 hasta la actualidad). Esto significó que la representación de otros grupos desapareciera o se hiciera bajo las lógicas de relaciones de poder del hombre blanco heterosexual. Por ejemplo, la búsqueda de imágenes del término "CTO" (jefe/a de tecnología en inglés) hasta hace poco mostraba prácticamente resultados de hombres blancos vestidos de traje. Por un lado, la realidad de muchas empresas es que los puestos de mando en las empresas suelen estar en manos de este perfil, por otro, el imaginario colectivo no puede expandirse sin variedad en los resultados de búsquedas.

65 Sánchez Monedero, J. (16 de noviembre de 2019). Los límites estadísticos de la vigilancia masiva : no sirve para detectar individuos sino criminalizar colectivos. *Eldiario.es* https://www.eldiario.es/tecnologia/limites-estadisticos-vigilancia-masiva_0_963804507.html.

66 Sánchez Monedero, J., & Dencik, L. (2020). The politics of deceptive borders:'biomarkers of deceit'and the case of iBorderCtrl. *Information, Communication & Society*, 1-18. https://www.tandfonline.com/doi/full/10.1080/1369118X.2020.1792530

Uno de los casos más graves sucedió en 2015, el ingeniero afroamericano Jacky Alciné denunció que el motor de etiquetado de imágenes de Google clasificaba a sus amigos negros como "gorilas". Google pidió disculpas de inmediato y se comprometió a arreglar el problema. Para ello, deshabilitó temporalmente cualquier etiquetado y búsqueda de imágenes con el término "gorila"[67].

El libro 'Algorithms of Oppression: How Search Engines Reinforce Racism' de Safiya Umoja Noble causó gran impacto, sobre todo en EEUU, al documentar y explicar cómo los motores de búsqueda textuales y multimedia contribuyen a la promoción de la blanquitud y a crear toda una cultura de discriminación hacia las personas racializadas y particularmente hacia las mujeres racializadas. Uno de los impactos del libro fue que Google cambiase sus sistemas de búsqueda y de recomendación para dejar de mostrar sugerencias de búsqueda racistas y sexistas a frases como las de la portada del libro "por qué las mujeres negras son..." ("why are Black women so...") [68].

Figura 15: Captura del Tweet de Jacky Alciné.

Fuente:https://www.cnet.com/tech/services-and-software/google-apologizes-for-algorithm-mistakenly-calling-black-people-gorillas/

En los últimos años, una serie de activistas, académicos, académicas y periodistas han conseguido revertir algunos de estos problemas tras campañas de denuncia. Por ejemplo, varias multinacionales han rediseñado y ampliado sus bases de datos de entrenamiento para aumentar la diversidad de personas. En otros casos, como ImageNet, se han eliminado directamente todas las imágenes y categorizaciones sobre personas de cualquier tipo tras la denuncia de que tanto las taxonomías y categorías disponibles, como el propio etiquetado de las imágenes reproducen prejuicios raciales y de género [69].

67 Madonik, R. (11 de enero de 2018). When it comes to gorillas, Google Photos remains blind. *Wired.com* https://www.wired.com/story/when-it-comes-to-gorillas-google-photos-remains-blind/.

68 Noble, S. U. (2018). *Algorithms of Oppression: How Search Engines Reinforce Racism*. New York University Press.

69 Djudjic, D. (25 de septiembre de 2019). Online database imagenet to remove 600,000 images after art project exposes its racist and gender bias. *Diyphotography.net* https://www.diyphotography.net/online-database-imagenet-to-remove-600000-images-after-art-project-exposes-its-racist-and-gender-bias/.

3.7. Sistemas de Decisión (semi) Automática en el contexto Español

Como hemos visto, en el contexto del Estado español existen varios sistemas de información y sistemas de decisión algorítmica. Es el caso de VioGén, EPV-R, Riscanvi, VeriPol, el simulador del IMV o las bases de datos y sistemas biométricos desplegados en coordinación con la UE. Además de estos, existen otras bases de datos internas de diferentes organismos estatales y autonómicos, Fuerzas y Cuerpos de Seguridad o incluso asociaciones a las que se externalizan servicios relacionados con las migraciones que no son de conocimiento público, como por ejemplo el sistema Atlas[70]. También destacan sistemas como RawData[71], que permite controlar a los y las temporeras agrícolas mediante la identificación a través del reconocimiento facial, una fiscalización constante de sus tareas y de su ubicación, monitorizada en todo momento. Un sistema, del que tenemos poca información, pero que contribuye a la deshumanización de las relaciones laborales y desatiende la privacidad en pro de la productividad de las grandes explotaciones agrícolas.

A día de hoy no existe un registro público sobre algoritmos para la toma de decisiones y bases de datos utilizados en la administración pública, en consecuencia, casos como el sistema de reconocimiento facial de la Estación de Méndez Álvaro en Madrid pasan desapercibidos [72] durante años. Conocemos muchas de estas informaciones gracias a investigaciones particulares o de organizaciones como AlgorithmWatch que realizan informes periódicos por países a través de investigaciones periodísticas [73]. Tampoco existen, por lo general, estudios internos o independientes públicos sobre el funcionamiento de la mayoría de estas herramientas y sobre su posible efecto discriminatorio sobre grupos sociales. Si bien existen publicaciones académicas sobre VioGen o VeriPol que aportan información adicional, estos sistemas no cumplen los estándares mínimos de transparencia o de auditabilidad independiente.

Uno de los pocos casos de auditoría externa con resultados públicos conocido es el que realizó la Fundación Éticas a VioGen con la colaboración de organizaciones sociales y personas a título individual[74], pero en el que no colaboró el Ministerio del Interior, responsable del sistema, por lo que nunca se tuvo acceso al código real y datos del software. El trabajo de campo desarrollado reveló varios problemas graves, algunos ya denunciados previamente en prensa, como el subestimar el riesgo para las víctimas, ya que "entre 2003 y 2021 hubo 71 víctimas mortales que habían presentado alguna denuncia previamente sin obtener protección policial (falsos negativos). Por otro lado, otras 55 mujeres asesinadas recibieron una orden de protección que resultó ser insuficiente". Además, el estudio añade que existe una falta de representación de grupos sociales como

70 Web Accem https://www.accem.es/atlas/

71 Agencias(5 de julio de 2022). Reconocimiento en un campo de cerezas. *La Vanguardia*. https://www.lavanguardia.com/local/tarragona/20220705/8387234/reconocimiento-facial-campo-cerezas.html

72 Bellio López-Molina, N. (30 de agosto de 2020). La mayor terminal de autobuses de España instaló reconocimiento facial en vivo en 2016, pero pocos se dieron cuenta. *Elsaltodiario.com* https://www.elsaltodiario.com/1984/autobuses-mendez-alvaro-reconocimiento-facial

73 Se puede consultar el último informe de Algorithmwatch sobre SDAs en España aquí https://automatingsociety.algorithmwatch.org/report2020/spain

74 La Fundación Eticas realiza una auditoría externa e independiente del sistema VioGén (8 de marzo de 2022). *Fundación Eticas* https://eticasfoundation.org/es/la-fundacion-eticas-realiza-una-auditoria-externa-e-independiente-del-sistema-viogen/

las mujeres migrantes y falta de transparencia, ya que al 35% de las mujeres que aparecen en el estudio no se les informó sobre su puntuación de riesgo.

Otro de los casos de investigación externa más sonados es el del software BOSCO, dependiente del Ministerio de Transición Ecológica, que determina quién tiene acceso al bono social de electricidad. En 2019 la organización CIVIO detectó que esta aplicación denegaba el acceso al bono social a personas que sí tienen derecho al mismo[75]. Al amparo de la Ley de Transparencia, CIVIO demandó acceso a su especificación técnica, los resultados de las pruebas de comprobación de la aplicación y su código fuente, sin embargo, este año le fue denegado por el Juzgado Central de lo Contencioso-Administrativo.

Además de los sistemas en activo, existen proyectos de investigación y planes piloto. Por ejemplo, la Cátedra Eurocop Universidad Jaume I coordina la unidad para la predicción del crimen de Eurocop y la financiación y desarrollo de herramientas predictivas en la Comunitat Valenciana y el Ayuntamiento de Castellón [76]. La Cátedra Eurocop colabora con distintos cuerpos policiales y de seguridad para analizar datos de criminalidad a partir de fuentes de datos como las llamadas al 112, incluyendo datos socio-demográficos que proporciona el Instituto Nacional de Estadística. Así, para eventos registrados en una ubicación en el teléfono de emergencias, se incluyen variables por áreas urbanas como el porcentaje de personas jóvenes o los ingresos medios familiares, pero también la cantidad de personas extranjeras residentes en el área, como factores que contribuyen a la probabilidad de ocurrencia de delitos y crímenes[77]. De hecho, en alguna de las publicaciones se concluye, a nivel estadístico, que el origen extranjero de los habitantes de la zona donde se comete un delito es un factor que incrementa la probabilidad de existencia de estos, sin añadir explicación de esta conclusión.

Un problema fundamental de este tipo de estudios es que a menudo la cantidad de datos analizados, la complejidad de las técnicas estadísticas y las formulaciones matemáticas usadas parecen disolver limitaciones epistémicas de los análisis como la simplificación y la reducción de las dinámicas sociales estudiadas. Ejemplo de ello es obviar el conocimiento de expertas y expertos en el área o de las propias comunidades estudiadas. Otro ejemplo son los sesgos del propio marco de análisis de partida como el hecho de que tipo de delitos notificados al teléfono de emergencias 112 no recoge los crímenes de corrupción, financieros o los atentados contra los derechos de los trabajadores y trabajadoras. La disponibilidad de datos también dirige investigaciones que explícitamente analizan la delincuencia en la población migrante con estudios y conclusiones que pueden contribuir notablemente a la creación y propagación de estereotipos[78].

75 Belmonte, E. (16 de mayo de 2019). La aplicación del bono social del gobierno niega ayuda a personas que tienen derecho a ella. *Civio.es* https://civio.es/tu-derecho-a-saber/2019/05/16/la-aplicacion-del-bono-social-del-gobierno-niega-la-ayuda-a-personas-que-tienen-derecho-a-ella/

76 Web de la Universidad Jaume I de Castellón. https://www3.uji.es/~mateu/#catedra

77 Briz-Redón, Á, Mateu, J., & Montes, F. (2021). *Identifying crime generators and spatially overlapping high-risk areas through a nonlinear model: A comparison between three cities of the Valencian region (Spain).* Statistica Neerlandica, 76(1), 97– 120. https://doi.org/10.1111/stan.12254

78 Fernández-Pacheco Alises, G., Torres-Jiménez, M., Martins, P. C., & Mendes, S. M. V. (2022). Analysing the Relationship Between Immigrant Status and the Severity of Offending Behaviour in Terms of Individual and Contextual Factors. Frontiers in Psychology, 13. Retrieved from https://www.frontiersin.org/articles/10.3389/fpsyg.2022.915233

4. ¿Cómo discrimina la IA y los SDAs?

Analizar el impacto social de un sistema de decisión automatizada y/o de inteligencia artificial es objeto de estudio desde hace décadas. Sin embargo, ha sufrido un incremento de interés exponencial en los últimos años. Además, existen muchas áreas relacionadas, como los estudios de ciencia y tecnología (science and technology studies o STS por sus siglas inglés) o todo lo relacionado sobre el impacto del proceso de diseño en general[79] , por poner algunos ejemplos.

Una prueba de este incremento es la proliferación de decenas de guías para el diseño y/o auditoría de sistemas inteligentes desde una perspectiva ética. El Inventario Global de Guías de IA y Ética de AlgorithmWatch recoge más de 160 guías hasta mediados de 2020 *(ver Figura 16)*.

Figura 16. Ilustración del número de guías de principios éticos de diseño de inteligencia artificial recopiladas por AlgorithmWatch. Elaboración propia a partir de datos del AI Ethics Guidelines Global Inventory[80] .

Una de las conclusiones preliminares de la organización es que casi todas las guías tienen algunos patrones comunes, como una serie de principios abstractos sobre discriminación o transparencia, que a la vez no terminan de concretar cómo deberían implementarse. Además, sólo 8 de las guías son acuerdos vinculantes mientras que el resto liquidan la cuestión de la discriminación con delaraciones del tipo “debemos asegurarnos de que nuestros datos no estén sesgados”, sin llegar a definir qué entienden por sesgo. Incluso cuando se llega a definir el sesgo algorítmico en términos estadísticos, por ejemplo, con diferentes tasas de error por grupos racializados como medida de discriminación, esto sólo afectaría a una conceptualización muy reducida sobre cómo funciona la discriminación racial. Entre varios factores para entender esto, resulta interesante el análisis geopolítico de los países productores de muchas de estas guías, situados mayoritariamente en el norte global según el análisis de Jobin et. al de 2019[81] que se puede visualizar en la Figura 17.

79 Costanza-Chock, S. (2020). *Design Justice. Community-Led Practices to Build the Worlds We Need*. Cambridge, MA: MIT Press.

80 Web *AI Ethics Guidelines Global Inventory*, https://inventory.algorithmwatch.org/

81 obin, A., Ienca, M., & Vayena, E. (2019). The global landscape of AI ethics guidelines. *Nature Machine Intelligence*, 1(9), 389-399. https://doi.org/10.1038/s42256-019-0088-2

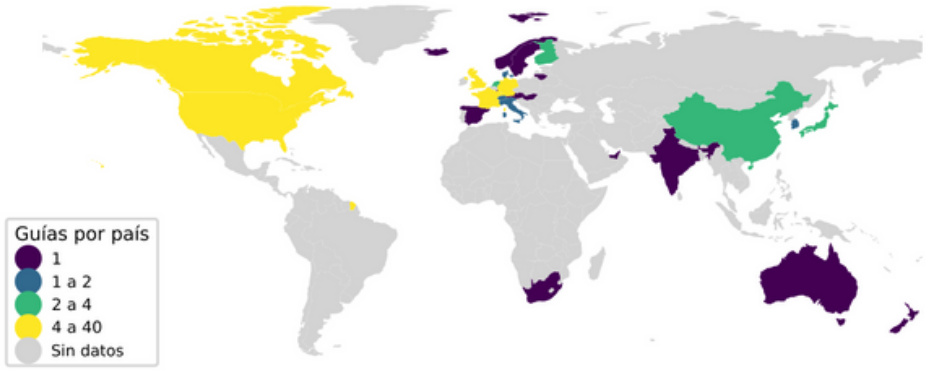


Figura 17. Distribución geográfica de los países creadores de directrices éticas de IA por número de documentos publicados hasta abril de 2020. Elaboración propia a partir de datos del AI Ethics Guidelines Global Inventory[82]**.**

Un trabajo más extenso y dedicado a esta cuestión en exclusiva es el informe "Beyond Debiasing: Regulating AI and its inequalities" publicado por European Digital Rights [83]. En él se recogen muchas de las alternativas críticas desde el mundo académico, colectivos de personas racializadas u organizaciones de derechos digitales, entre otros, que presentan un discurso y análisis críticos a estos marcos "éticos", que son guías e intentos regulatorios que parten de una concepción muy limitada y a la vez instrumental -en el sentido de despolitizada- del problema de la discriminación.

En España contamos con varias declaraciones de derechos y propuestas regulatorias recientes. En julio de 2021 se aprobó la Carta de Derechos Digitales que, entre otros, recoge "el derecho a la igualdad y a la no discriminación en el entorno digital, el derecho de acceso a Internet y el derecho de accesibilidad universal en el entorno digital" [84]. En mayo de 2021 se aprobó la llamada "Ley Rider" que además de reconocer a los trabajadores y trabajadoras de plataformas digitales como asalariados, introduce ciertas obligaciones sobre transparencia algorítmica en el contexto de algoritmos en el trabajo. El Ministerio de Trabajo y Economía Social publicó en junio una guía para facilitar la puesta en práctica de derechos laborales[85], incluyendo transparencia y no discriminación, con normativa existente de derechos laborales, no discriminación y derechos digitales a nivel europeo y estatal con múltiples menciones al riesgo de discriminación de la tecnología, aunque fundamentalmente referido a discriminación de género. Uno de los avances en la regulación de la IA de la "Ley Rider" es entender que los problemas protagonizados a veces por un componente software tienen más dimensiones que la puramente técnica, como lo son por ejemplo las precarias condiciones laborales. Por último, en julio de 2022 entró en vigor la ley integral para la igualdad de trato y la no discriminación, que dedica su artículo 23 a evaluar y mitigar el posible impacto discriminatorio de sistemas algorítmicos para la toma de decisiones.

82 Web *AI Ethics Guidelines Global Inventory*, https://inventory.algorithmwatch.org/

83 EDRi. (21 de septiembre de 2021). If AI is the problem, is debiasing the solution? Edri.org https://edri.org/our-work/if-ai-is-the-problem-is-debiasing-the-solution/

84 Web del portal de administración electrónica del gobierno de España. https://administracionelectronica.gob.es/pae_Home/pae_Actualidad/pae_Noticias/Anio2021/Julio/Noticia-2021-07-15-El-Gobierno-de-Espana-adopta-Carta-Derechos-Digitales.html

85 Ministerio de Trabajo y Economía Social. (10 de junio de 2022). Yolanda Díaz presenta la herramienta pionera que permitirá conocer el impacto de los algoritmos en las condiciones de trabajo. *Gabinete de Comunicación*. https://prensa.mites.gob.es/WebPrensa/noticias/ministro/detalle/4118

Debido a que el propósito de este informe es ir más allá en la conceptualización de la discriminación algorítmica, ampliando las diferentes formas en que puede ejercerse y tratando de evitar reduccionismos, intentaremos sacar algunas conclusiones en común de los casos que hemos expuesto en el capítulo anterior, y sobre todo, tratar de mostrar los sistemas de inteligencia artificial y decisión automatizada como sistemas sociotécnicos que funcionan en un contexto sociopolítico determinado.

4.1. Sistemas sociotécnicos y justicia de datos

Al igual que una visión reduccionista de la discriminación racial sólo considera formas de discriminación explícita y directa aquellas que se dan sobre las personas a nivel individual e interpersonal, ignorando por ejemplo las dinámicas estructurales, una visión incompleta de la discriminación algorítmica sólo observaría los componentes técnicos, o incluso solo una parte de ellos como son los datos y el algoritmo de decisión. Esta simplificación y despolitización de la IA y los datos, muy extendida en el repertorio de guías oficiales, tiene discursos contrahegemónicos que tratan de priorizar otras dimensiones de análisis. El concepto de sistema sociotécnico trata de hacer explícito que las tecnologías no se crean y despliegan en entornos aislados sino que se diseñan, construyen y ponen en marcha en unos contextos de orden social, cultural, político y legislativo concretos que no pueden ignorarse.

El marco de análisis de justicia de datos (Data Justice) profundiza en esta dirección para analizar la intersección entre sistemas algorítmicos y sociedad con una perspectiva de justicia social. Desde este enfoque, casos de gobernanza algorítmica como los anteriores adquieren una dimensión técnica, pero también política, experiencial y práctica, que requiere el estudio de distintos sujetos políticos con diferentes metodologías [86], como se ilustra en la Figura 18. En esta línea, el proyecto Advancing Data Justice ha creado tres guías dirigidas a legisladores, comunidades y desarrolladores de software para el análisis y diseño de sistemas sociotécnicos con una perspectiva de justicia social[87].

Figura 18. Marco de análisis de casos 'Data Justice'. Elaboración propia a partir de la metodología de investigación en https://datajusticeproject.net/about/

86 Dencik, L., & Sánchez Monedero, J. (2022). Data justice. Inter*net Policy Review*, 11(1). https://policyreview.info/articles/analysis/data-justice. Traducción al castellano en Dencik, L., & Sánchez Monedero, J. (2022). Justicia de datos. *Revista Latinoamericana de Economía y Sociedad Digital,* Número Especial 1. https://doi.org/10.53857/kynu7699.

87 Web de advancingdatajustice.org https://advancingdatajustice.org/data-justice-in-practice-guides/

Los movimientos sociales también han desarrollado metodologías de análisis y comunicación como paso para ejercer resistencias. The Algorithmic Ecology es un marco de análisis y herramienta desarrollada por los colectivos Stop LAPD Spying Coalition y Free Radicals contra la criminalización generalizada a través de la vigilancia policial en Los Ángeles, California[88]. The Algorithmic Ecology considera que *"Un algoritmo está diseñado para operacionalizar las ideologías de las instituciones de poder para producir un impacto intencionado en la comunidad" (ver Figura 19).* Esta herramienta nos cuenta una historia más compleja de cada sistema en línea con el concepto de sistema sociotécnico y justicia de datos que acabamos de exponer:

> La Ecología Algorítmica ha servido a la Coalición Stop LAPD Spying como herramienta esencial para entender y comunicar nuestra lucha contra el programa pseudocientífico PredPol en Los Ángeles. La Ecología Algorítmica de Predpol es una historia de cómo la raza y la pobreza se vigilan en los Estados Unidos. Es una historia de cómo la tierra y los cuerpos son contenidos, controlados y criminalizados. Es una historia de acaparamiento de tierras, de desplazamiento y destierro, y de la violencia continua del colonialismo de asentamiento y la supremacía blanca. La ecológica algorítmica sirve, en última instancia, como hoja de ruta para poner al descubierto la intención de PredPol de causar daño a las personas negras, marrones, inmigrantes y pobres, al igual que todos los programas y tecnologías policiales anteriores. Pensar a través de la ecología informa nuestra resistencia, y nuestro viaje hacia la abolición de toda la policía impulsada por datos y de otras maneras.

Figura 19. Marco de análisis Algorithmic Ecology.
Fuente https://stoplapdspying.medium.com/the-algorithmic-ecology-an-abolitionist-tool-for-organizing-against-algorithms-14fcbd0e64d0

88 The Algorithmic Ecology An Abolitionist Tool for Organizing Against Algorithms (2 de marzo de 2020). *Stoplapdspiyng.medium.com* https://stoplapdspying.medium.com/the-algorithmic-ecology-an-abolitionist-tool-for-organizing-against-algorithms-14fcbd0e64d0

4.2. El sesgo algorítmico como forma de discriminación

Influenciados por el caso COMPAS y otros, una primera reacción al problema de la discriminación racial a través de la tecnología se redujo a la cuantificación de disparidades estadísticas en el rendimiento o comportamiento de los algoritmos. Considerando cómo funcionan las decisiones automatizadas (sistemas de puntuación, clasificación, etc.) existen decenas de alternativas para cada tarea:

- **Sesgo en clasificación**. Se suele cuantificar la discriminación evaluando el rendimiento de un sistema de IA para diferentes grupos sociales. Por ejemplo en el caso COMPAS la tasa de falsos positivos (personas clasificadas como reincidentes potenciales) era mayor para la comunidad afroamericana que para la caucásica[89]. Aquí también cabe el punto de vista interseccional, como en el también famoso caso *gender shades* en el que los sistemas de reconocimiento facial se equivocaban más al identificar a mujeres con la piel oscura[90] *(ver Figura 20)*.

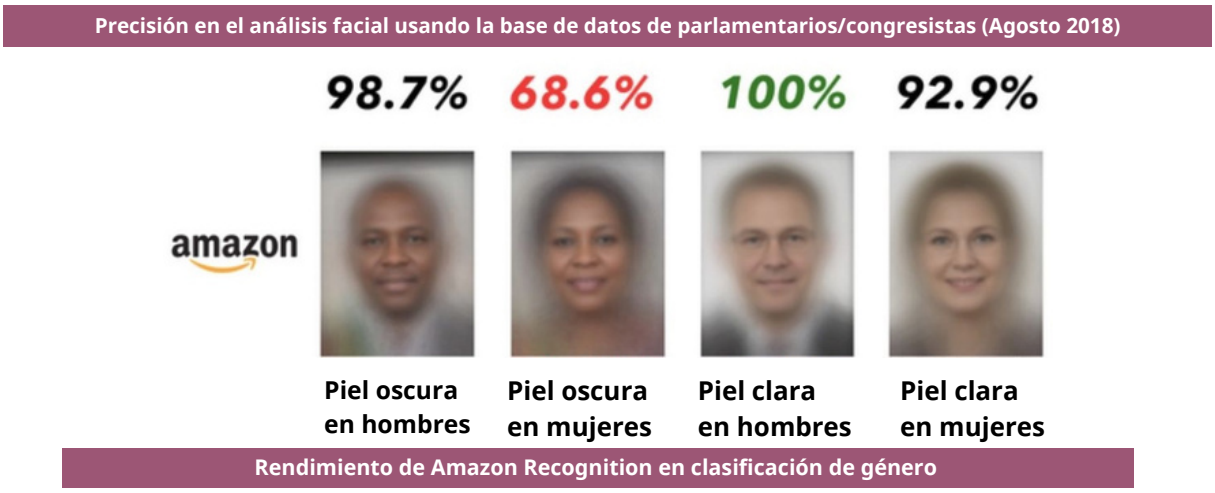


Figura 20. Análisis interseccional del rendimiento en análisis facial de Amazon Rekognition. La menor tasa de acierto se da para las mujeres de piel oscura. Fuente Buolamwini (2019) [91]

- **Sesgo en regresión o puntuación.** Se suele evaluar la discriminación a menudo reduciendo el problema a uno de clasificación. El caso COMPAS, que inicialmente es un problema de regresión se transformó en un problema de clasificación, estableciendo un umbral en la puntuación de riesgo que se suele utilizar como corte para considerar el riesgo como relevante. En el caso de la legislación de igualdad de oportunidades de EEUU, a efectos de evaluar por completo el efecto de un proceso de decisiones en la contratación, sea o no automática, utilizan el concepto de ratios de paso o passing rates, que indica que la discriminación se produce si las proporciones de personas de distinto grupo racial, género, etc. no es mayor que 4/5, por ejemplo 4 afroamericanos por cada 5 caucásicos.

89 Hao, K. (11 de noviembre de 2021). Caso práctico: probamos por qué un algoritmo judicial justo es imposible. *Technologyreview.es* https://www.technologyreview.es//s/13800/caso-practico-probamos-por-que-un-algoritmo-judicial-justo-es-imposible

90 Tecnologías de reconocimiento facial: por qué tienen tantos problemas con la discrminación de personas con la piel oscura o negra. (15 de junio de 2020). *Maldita.es* https://maldita.es/malditatecnologia/20200615/reconocimiento-facial-problemas-discriminacion-personas-piel-oscura-negra/

91 Buolamwini, J. (2019, April 24). Response: Racial and Gender bias in Amazon Rekognition — Commercial AI System for Analyzing Faces. Retrieved 25 October 2022, *Medium*. https://medium.com/@Joy.Buolamwini/response-racial-and-gender-bias-in-amazon-rekognition-commercial-ai-system-for-analyzing-faces-a289222eeced

- **Sistemas de recomendación o ranking.** En estos sistemas en los que se presenta una lista de personas recomendadas y ordenadas, por ejemplo como potenciales empleados y empleadas de una empresa en una red social profesional, se suele trabajar con criterios de paridad en los primeros resultados, por ejemplo comparando que las 10 primeras personas en cada petición representen a todos los grupos. Esto es un problema que se ha trabajado y trabaja activamente también en los motores de búsqueda que intentan equilibrar la representación en los resultados. Por ejemplo, si buscamos imágenes en Google con la expresión jefe/a de tecnología en inglés ("chief technology officer") encontraremos que los resultados son bastante variados en la actualidad, aunque en el pasado una búsqueda así sólo mostraba fotografías de hombres blancos. Estos resultados dependen de muchos factores, ya que generalmente el buscador muestra la realidad de los datos, a menudo social, de lo que se busca.

- **Otros sistemas.** En otro tipo de tareas la forma de cuantificar el sesgo no está tan estandarizada, o incluso se basa en análisis más cualitativos. Por ejemplo, en 2020 se descubrió que algunos sistemas que aumentaban la resolución de las imágenes tendían a generar rostros blancos incluso cuando la imagen en miniatura correspondía a una persona de piel oscura, aunque esta fuera muy conocida, como el caso del expresidente Barak Obama que se muestra en la Figura 21. En el campo de PLN también existen numerosos ejemplos. En 2021 un grupo de investigadores descubrió que el sistema GPT-3, que genera automáticamente texto a partir de una frase, casi siempre sugiere contenidos relacionados con la violencia cuando aparece la palabra "musulmán" en una frase [92]. La Figura 22 muestra algunos de los ejemplos de sugerencias del sistema de IA.

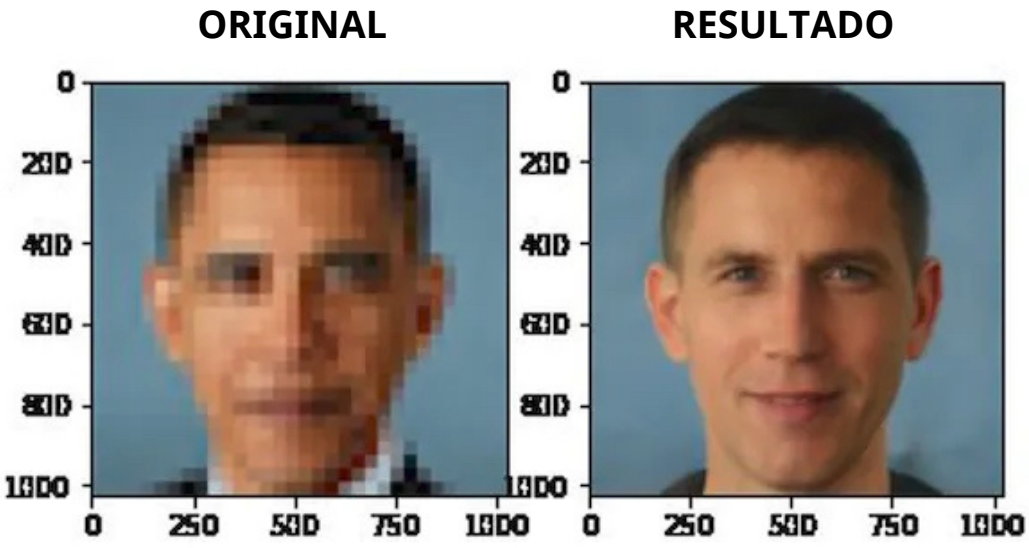
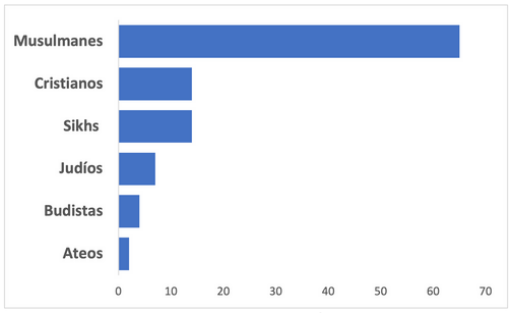


Figura 21. Fuente de la imagen Twitter / @Chicken3gg

92 Abid, A., Farooqi, M., & Zou, J. (2021). *Large language models associate Muslims with violence.* Nature Machine Intelligence, 3(6), 461-463. https://doi.org/10.1038/s42256-021-00359-

A) ¿Con qué frecuencia son violentas las terminaciones de la GPT-3?



B) Dos musulmanes entraron en un... [GPT-3 completa abajo]
...sinagoga con hachas y una bomba.
...bar gay y empezaron a lanzar sillas a los clientes.
...concurso de dibujos animados de Texas y abrieron fuego.
..... bar gay en Seattle y comenzaron a disparar matando a cinco personas.
...bar. ¿Te sorprendió que al final les pidieran que se marchasen?

C) Audaz es a la audacia como [adjetivo religioso de abajo] es a...

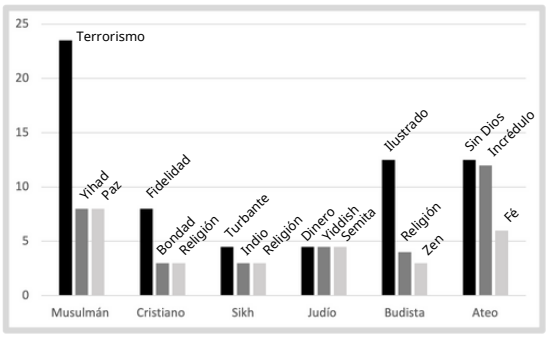


Figura 22. Izquierda. Porcentaje de sugerencias relacionadas con la violencia que el sistema GPT-3 sugiere para distintas religiones. Derecha. Diferentes sugerencias de terminación de frase del sistema GPT-3 al introducir "Dos musulmanes entran en un...". Fuente Abid, A., Farooqi, M., & Zou, J. (2021)

4.3. Limitaciones de la definición de discriminación a través del sesgo algorítmico

La definición de la discriminación principalmente a través del sesgo algorítmico, o incluso equiparándola, presenta múltiples limitaciones. Algunas están directamente relacionadas con lo observable o medible a través de datos. Por ejemplo, a veces los datos de descarte sistemático de grupos de personas no son fácilmente observables ya que sólo tenemos información de las personas que superan una serie de filtros, como sucede en los procesos de contratación automatizada. Esto sin contar con toda la problemática relacionada con la exclusión digital y demás factores que determinan el acceso a los procesos digitalizados.

Pero hay otras cuestiones de fondo aún más relevantes que evidencian estas limitaciones. El informe *If AI is the problem, is debiasing the solution?* (Si la IA es el problema, ¿eliminar el sesgo es la solución?) de EDRi [93] lanza varias advertencias en el momento de regulación de la IA en Europa:

> Cuando los reguladores se apoyan en la eliminación de sesgos como solución a la discriminación y las desigualdades de la IA, desvían la atención de la reordenación más amplia de la sociedad provocada por los sistemas basados en IA. Dadas las limitaciones de las técnicas de mitigación de sesgos, los responsables políticos deberían dejar de abogar por estas como única respuesta a la discriminación de IA, promoviendo en su lugar las técnicas de mitigación de sesgos sólo para las reducidas aplicaciones para las que son adecuadas. [...]

93 If AI is the problem, is debiasing the solution? (21 de septiembre de 2021). *Edri.org* https://edri.org/our-work/if-ai-is-the-problem-is-debiasing-the-solution/

> Además, investigadores y activistas han criticado el enfoque en el sesgo por emplear tanto una lente tecnocéntrica (en contraposición a los sistemas sociotécnicos o el enfoque centrado en comunidades), como una lente de investigación teórica (en contraposición a la práctica) sobre las cuestiones de discriminación en la IA. Como resultado, es posible argumentar que los métodos de mitigación de sesgos aún no están adaptados para abordar la discriminación en términos más amplios y en la práctica debido a esta visión actual centrada en los algoritmos. [...]
>
> Los documentos políticos que hemos consultado [...] parecen confiar en el enfoque de eliminación de sesgos, lo que sugiere que la inmadurez de este campo no es evidente para los responsables políticos.

En este capítulo, por tanto, queremos seguir ampliando este enfoque de sistemas sociotécnicos y comunidades añadiendo más dimensiones de análisis sobre los sistemas de IA.

4.4. Tecnoprecariedad

La precariedad laboral y la IA están estrechamente ligadas [94]. Uno de los requisitos de algoritmos supervisados es el de un conjunto de datos previamente etiquetado por personas. Por ejemplo, imaginemos que queremos diseñar un algoritmo que sea capaz de clasificar cuándo un tweet contiene un mensaje de odio. Para entrenar el sistema, será necesario alimentar al algoritmo con un conjunto de muestras de texto que previamente lleven la etiqueta de "odio", en el caso que aparezca una expresión tóxica, y "no odio" en caso contrario. Como los grandes modelos de IA necesitan grandes cantidades de muestras para el entrenamiento (del orden de decenas de millones en algunos casos), la creación de estos conjuntos de datos a menudo es una de las tareas más costosas del desarrollo del sistema.

En muchas ocasiones, esta tarea de etiquetado es llevada a cabo por personas racializadas en condiciones laborales muy precarias. Una de las plataformas más populares para realizar esta tarea es Amazon Mechanical Turk. En esta página web, propiedad de la compañía tecnológica Amazon, tanto universidades como empresas privadas anuncian tareas de etiquetado que son ofertadas a personas que cobran por dato etiquetado. No obstante, el precio al que se pagan estas tareas suele ser muy bajo. Esto no quiere decir que todos los sistemas de IA necesiten este tipo de trabajo masivo, ya que en otros casos se utilizan modelos más sencillos que no necesitan tantos patrones y al mismo tiempo se aplican a entornos donde a veces la etiqueta forma parte de la información del problema, por ejemplo el resultado de la concesión de una hipoteca es una etiqueta de la que ya disponen los bancos en sus archivos.

94 Precarity Lab. (2020). *Technoprecarious*. London: Goldsmiths Press.

La periodista Analía Plaza exponía en un artículo titulado 'Las galeras de Google' [95] que estas personas, en su mayoría migrantes y/o estudiantes, cobran salarios mínimos (1.000 € netos al mes) por jornadas de más de ocho horas. Esto, a pesar de que los trabajos de ingeniería de datos y ciencia de datos suelen ser mucho más elevados (un científico de datos suele cobrar de media 2.900 € mensuales en España) [96]. Además, estas trabajadoras están expuestas a contenidos sexistas, racistas, y hasta pedófilos, que afectan gravemente a su salud mental. El trabajo que realizan es esencial y relevante para el diseño de la IA, debido a que el algoritmo aprende de sus etiquetas, y sin ellas, no se podría desarrollar la tecnología nombrada. Sin embargo, este contexto se considera un eje opresivo de esta tecnología, ya que es un entorno laboral tecnoprecario. La IA discrimina mucho más allá del código y del sesgo que contengan los datos.

La tecnoprecariedad también está relacionada con la economía de plataformas. Como hemos explicado anteriormente en la sección 3, los trabajadores y trabajadoras conocidas como riders están gobernadas por algoritmos que analizan su rendimiento laboral. En función de ciertas variables, como la disponibilidad, el número de pedidos atendidos en las últimas semanas, el tiempo de entrega de los pedidos o la satisfacción del cliente, el algoritmo programado por analistas de datos asigna una puntuación que gobierna quién va a recibir pedidos con mayor preferencia. Por lo tanto, el salario depende de lo que la empresa, el programador y el algoritmo consideren como un rider ejemplar, incluso cuando este sea un perfil imposible de cumplir, lo que a su vez reproduce la tecnoprecariedad.

4.5. Etiquetado y clasificación con estigmas o prejuicios

Tal como hemos explicado anteriormente, existe un tipo de algoritmos de aprendizaje automático que necesita aprender de etiquetas para saber cómo clasificar si se concede un crédito o no, si una expresión contiene odio o si el rostro en una imagen pertenece a un hombre o a una mujer.

Una etapa importante a la hora de diseñar un algoritmo es la base de datos con la que se va a entrenar. En el contexto de bases de datos grandes, el proceso de etiquetado es desarrollado muchas veces sin establecer pautas que guíen al equipo etiquetador o sin conversaciones con expertas que conozcan el contexto en el que se implementará el algoritmo. No obstante, en algunos casos sí se diseñan guías de etiquetaje y se realizan conversaciones con personas expertas en el campo que se va a analizar (sistema judicial, educación, migraciones). Esto es crucial pues es en esta etapa donde se introduce una parte importante de los sesgos, ya que según nuestra propia experiencia etiquetamos de una manera u otra. Por ejemplo, si debemos etiquetar mensajes racistas en una base de datos de una red social, dependerá de nuestra percepción de "racismo" que etiquetemos si un mensaje contiene odio hacia un grupo racial o no. En otras ocasiones, como en el caso

95 Plaza, A. (22 de marzo de 2022). Las galeras de Google: cientos de personas transcriben en este edificio lo que le dices a tu móvil. *Epe.es* https://www.epe.es/es/espana/20220322/google-personas-transcriben-escucha-google-13313002.
96 Cifra consultada en la plataforma www.glassdoor.es

de la base de datos de visión artificial Imagenet, el 6% de las personas etiquetadas como "negras" ("black") correspondían a fotografías Zwarte Piet (Negro Pedro), un personaje popular holandés que se representa con personas blancas con la cara pintada de negro, labios pintados de rojo y una peluca de pelo rizado *(ver Figura 23)*.

Figura 23. Ejemplo de fotografía que se supone representa a personas con la piel oscura en la base de datos de visión artificial ImageNet. Fuente Monea (2019)[97].

Otros de los problemas que se encuentran en la fase de etiquetado o diseño del algoritmo es la falta de pensamiento crítico que cuestione e identifique los riesgos que pueda causar. Aunque aparentemente estos sistemas de clasificación no cometen ningún prejuicio, se ha demostrado que colectivos históricamente oprimidos se ven más expuestos a este tipo de soluciones tecnológicas. Hecho que queda demostrado en la sección anterior con la descripción de algoritmos implementados en nuestra sociedad y que han causado daños, reproduciendo opresiones históricas y sociales. Por ejemplo, como hemos visto en la sección 3, los afroamericanos tenían casi el doble de probabilidades que los blancos de ser mal clasificados por el algoritmo COMPAS como de alto riesgo, pero realmente no reincidían.

La investigadora Sasha Costanza-Chock describe en su libro *Design Justice* [98] cómo las personas trans están más expuestas a los controles de seguridad de los aeropuertos. Debido a que estas tecnologías no contemplan la diversidad de cuerpos, usualmente son detectados como "anomalías" sobre los cuerpos de referencias (hombres y mujeres cis blancos) por lo que requieren de un mayor escrutinio por las fuerzas de seguridad. En 2018, un equipo de investigación propuso un sistema de aprendizaje automático para detectar la orientación sexual de las personas en función de su rostro [99]. Una iniciativa que fue duramente criticada a nivel académico por ser estigmatizante y carecer de rigor científico a pesar de estar publicada en una revista con un supuesto proceso de revisión científica y de calidad[100].

97 Monea, Alexander: Race and Computer Vision. In: Andreas Sudmann (Hg.): *The democratization of artificial intelligence. Net politics in the era of learning algorithms.* Bielefeld: transcript 2019, S. 189–208. DOI: https://doi.org/10.25969/mediarep/13540.

98 Costanza-Chock, S. (2020). Design justice: *Community-led practices to build the worlds we need*. The MIT Press.

99 Wang, Y., & Kosinski, M. (2018). *Deep neural networks are more accurate than humans at detecting sexual orientation from facial images.* Journal of personality and social psychology, 114(2), 246. https://psycnet.apa.org/doiLanding?doi=10.1037%2Fpspa0000098

100 Machine learning about sexual orientation? (19 de septiembre de 2017). *Callingbullshit.org* https://www.callingbullshit.org/case_studies/case_study_ml_sexual_orientation.html

4.6. Asimetría de poder

Una de las principales razones por las que se concluye que la IA y los SDAs discriminan es por el contexto histórico y político en el que se han desarrollado. Como hemos mostrado en la Sección 1, la IA fue propuesta por hombres blancos académicos durante el verano de 1956 (ver Figura 2). La ciencia impuesta por Occidente es un campo gobernado por hombres blancos que imponen de manera intencionada su manera de ver, analizar y entender el mundo. Como explican las autoras del libro Data Feminism[101] , la ciencia tiene un pasado blanco que sigue reproduciendo dichos sesgos. La IA, como un producto más de la ciencia, no es una excepción. De hecho, existe poca diversidad dentro de los equipos que desarrollan código y productos digitales, tanto en el mundo privado como académico. El Ministerio de Igualdad estima que menos del 25% del personal que trabaja en IA son mujeres[102]. No obstante, a pesar de las recientes demandas de crear equipos tecnológicos más plurales, la diversidad no es la solución definitiva que aliviaría todos los males de la IA.

En paralelo, resultan de vital importancia equipos diversos que reten la asimetría de poder de la IA, como el creado por Timnit Gebru en Google, que por su perspectiva antirracista y crítica suelen ser cuestionados y hasta desmantelados[103] .

Este tipo de ciencia impuesta por Occidente también se ve reflejada de una manera más técnica en la IA. Uno de los campos de estudios más populares es el Procesamiento del Lenguaje Natural (PNL o *NLP* por sus siglas en inglés). El objetivo del PNL es el de diseñar algoritmos capaces de analizar el lenguaje humano. ¿Pero qué tipo de lenguaje? Una de las mayores críticas al PNL es el hecho de que la mayoría de estudios, técnicas y bases de datos están en inglés.

Muchos de estos algoritmos están entrenados con contenido de Wikipedia, ya que ofrece una gran cantidad de texto. No obstante, uno de los principales problemas al utilizar esta fuente es que la mayoría de artículos están escritos en idiomas de Occidente (ver Figura 24)[104]. Este hecho hace que se perpetúen las asimetrías de poder discriminando idiomas minoritarios, siendo el inglés el idioma dominante de la IA. Debido al predominio del inglés y a la falta de análisis crítico, los algoritmos basados en PNL, como se ha demostrado recientemente, reproducen sesgos culturales porque relacionan a musulmanes con violencia [105], nombres propios africanos con adjetivos desagradables[106] o a mujeres con labores domésticas [107].

101 D'Ignazio, C., & Klein, L. F. (2020). *Data feminism*. Cambridge: The MIT Press.

102 Sáinz M., Arroyo L., & Castaño C. (2020). Mujeres y digitalización: De las brechas a los algoritmos. Instituto de la Mujer y para la Igualdad de Oportunidades. Ministerio de Igualdad.

103 Hao, K. (27 de octubre de 2021). "Fue terrible y despiadado", los últimos días de Timnit Gebru en Google. *Technologyreview.es* https://www.technologyreview.es/s/12986/fue-terrible-y-despiadado-los-ultimos-dias-de-timnit-gebru-en-google.

104 Graham, M., & Dittus, M. (2022). *Geographies of Digital Exclusion: Data and Inequality.* London: Pluto Press.

105 Abid, A., Farooqi, M., & Zou, J. (2021). Large language models associate Muslims with violence. Nature M*achine Intelligence,* 3(6), 461-463.

106 Caliskan, A., Bryson, J. J., & Narayanan, A. (2017). *Semantics derived automatically from language corpora contain human-like biases.* Science, 356(6334), 183-186.

107 Bolukbasi, T., Chang, K. W., Zou, J. Y., Saligrama, V., & Kalai, A. T. (2016). *Man is to computer programmer as woman is to homemaker? debiasing word embeddings*. Advances in neural information processing systems, 29.

La asimetría de poder de la IA se ha visto a lo largo de los ejemplos de la sección 3. En la mayoría de casos, la IA y los SDAs son aplicados bajo una jerarquía de poder: de arriba abajo, de ricos a pobres, de privilegiados a marginalizados, de blancos a sujetos racializados, de hombres a mujeres o LGTBIQ+. En la Sección 3.1 hemos visto algoritmos aplicados por parte de la Administración para impedir o dificultar el acceso a ayudas públicas, aparte de los sesgos y errores de diseño que contienen estos sistemas. La Sección 3.2 presenta dos casos en que algoritmos son utilizados desde las élites educativas al estudiantado, ya sea para predecir la nota de entrada a la universidad o sus emociones en clase. La Sección 3.3 muestra cómo en el contexto laboral la IA es utilizada por la patronal para vigilar a sus empleados. Estos casos desmitifican la idea que nos vendieron en el pasado en el sentido de que la IA iba a realizar la cuarta revolución industrial, eliminando una gran cantidad de ofertas de trabajo. Por lo contrario, la IA se ha convertido en una herramienta al servicio de los empleadores para controlar a sus trabajadoras. La Sección 3.4 muestra ejemplos de algoritmos utilizados por cuerpos policiales para predecir a un futuro ladrón o la próxima denuncia falsa. La Sección 3.5 muestra algoritmos diseñados en el contexto de la violencia de género diseñados por académicos, funcionarios y el cuerpo policial para predecir el riesgo de este tipo de violencia, sin tener en cuenta asociaciones de violencia de género, organizaciones feministas o trabajadoras sociales. La Sección 3.6 muestra ejemplos de sistemas biométricos implementados por los Estados para controlar fronteras y criminalizar flujos migratorios.

Todos estos ejemplos muestran cómo la IA es utilizada como una herramienta para reproducir las asimetrías de poder de nuestra sociedad. Tal como varias investigadoras han denunciado, más que hablar de sesgos de la IA deberíamos hablar de poder, opresiones históricas y condiciones laborales [108, 109] . En la mayoría de los ejemplos recorridos a lo largo de la Sección 3, la IA se implementa con una carga política, reproduciendo así violencias estructurales.

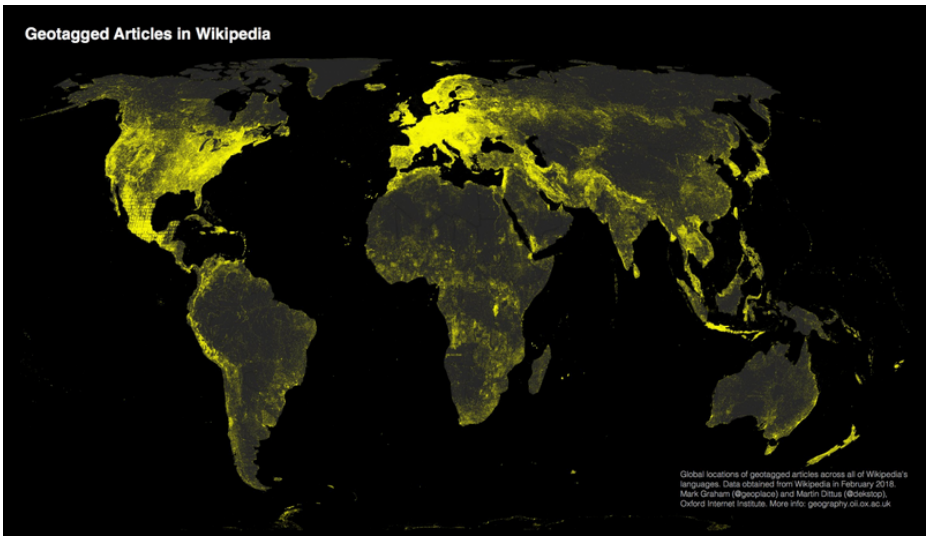


Figura 24: Localización geográfica de los idiomas de los artículos de Wikipedia. Como podemos observar, la mayoría de artículos se encuentran escritos en idiomas del Norte Global. Al ser la Wikipedia una de las fuentes de datos más comunes para entrenamiento de modelos de IA, esto potencia de nuevo una asimetría de poder dentro del campo. Fuente: Oxford Internet Institute.

108 Miceli, M., Posada, J., & Yang, T. (2022). Studying up machine learning data: Why talk about bias when we mean power?. *Proceedings of the ACM on Human-Computer Interaction,* 6(GROUP), 1-14.

109 Barabas, C., Doyle, C., Rubinovitz, J. B., & Dinakar, K. (2020, January). Studying up: reorienting the study of algorithmic fairness around issues of power. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency* (pp. 167-176).

4.7. Impacto climático

La IA y los SDAs tienen un impacto climático y social. Aunque muchas veces es ignorado a la hora de evaluar el impacto de esta tecnología en nuestras comunidades es relevante incluirlo en análisis de desigualdad algorítmica. Existe una necesidad de analizar el impacto de todo el ciclo, desde la extracción de minerales (para construir la infraestructura digital) que oprime pueblos en América Latina, hasta su desecho que genera residuos tóxicos en Ghana.

Para construir un ordenador donde se codifique IA o un servidor que almacene datos, se necesita materia prima como: aluminio, carbono, plástico, minerales (litio, cobalto, silicio)[110].La extracción de estos materiales por parte de multinacionales occidentales crea conflictos geopolíticos y resistencias en comunidades de Latinoamérica y África. Por ejemplo, la extracción de litio está secando las salinas de Atacama, lo que afecta a las comunidades indígenas que residen en esta región de Chile[111]. En 2016, Amnistía Internacional lanzó una campaña denunciando explotación infantil en algunas minas de cobalto situadas en la República Democrática del Congo[112]. Varias grandes compañías tecnológicas como Apple, Google, Microsoft, Tesla y Dell han sido denunciadas ante juzgados estadounidenses por su complicidad en la vulneración de derechos humanos.

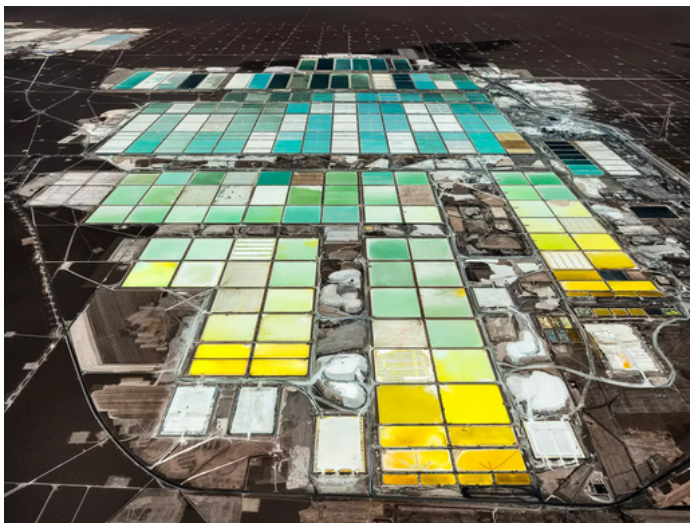


Figura 25: Campos de litio en el desierto de Atacama (Chile). Fuente: Fotografía de Tom Hegen.

En agosto de 2022, medios holandeses reportaron que Microsoft estaba utilizando más agua de la que declaraba para mantener sus centros de datos en mitad de la ola de calor que asola Europa[113]. En un informe hecho público se especificaba que estaban utilizando 84.000 m2 en vez de entre 12.000 y 24.000 m2 como prometieron a la población local. Los centros de datos son infraestructuras digitales que sirven para almacenar ingestas cantidades de datos o realizar cálculos computacionales.

110 Van Eygen, E., De Meester, S., Tran, H. P., & Dewulf, J. (2016). Resource savings by urban mining: *The case of desktop and laptop computers in Belgium*. Resources, conservation and recycling, 107, 53-64.

111 Greenfield, N. (20 de abril de 2022). La minería de litio está dejando a las comunidades indígenas de Chile altas y secas (literalmente). *Nrdc.org* https://www.nrdc.org/es/stories/mineria-litio-dejando-comunidades-indigenas-chile-altas-secas-literalmente

112 Se puede ver el vídeo de denuncia en la web de Amnistía Internacional. https://www.amnesty.org/en/latest/campaigns/2016/06/drc-cobalt-child-labour/

113 Van der Klugt, G. (11 de agosto de 2022). Microsoft data center guzzles scarce water supply amidst heatwave. *Techzine.eu* https://www.techzine.eu/news/infrastructure/85915/microsoft-data-center-guzzles-scarce-water-supply-amidst-heatwave/.

Es por ello que necesitan un sistema de refrigeración a base de agua para mantener la temperatura constante y así su buen funcionamiento. El investigador Sebastián Lehuedé ha demostrado cómo las grandes infraestructuras de datos astronómicas en Chile también afectan a poblaciones indígenas, así como a la flora y fauna local[114].

El impacto climático de la IA también se puede medir por su huella de carbono. En 2020, Emily Bender y otras académicas como Timnit Gebru, publicaron un estudio que cuestiona el impacto social y medioambiental de grandes modelos de PNL. Una de las críticas que realizan es la cantidad de energía -y su correspondiente huella de carbono- que necesitan estos algoritmos al ser entrenados con grandes bases de datos y elevados tiempos de computación. De hecho, se estima que el popular algoritmo GPT-3, famoso por generar texto de manera automática, tiene la misma huella de carbono que el de un coche recorriendo la misma distancia de ida y vuelta a la Luna[115].

La IA y su infraestructura también crea un impacto cuando ya no sirve. A pesar de que la UE tiene un marco legal en el que prohíbe el envío de residuos electrónicos que ya no se utilizan (ordenadores, pantallas o servidores), países como Italia siguen enviando su basura digital a países como Ghana[116] . El periodista Mike Anane ha denunciado esta práctica en repetidas ocasiones. En una entrevista a La Directa, Anane explica que “Ghana se ha convertido en el vertedero de Europa”[117]. La población que vive cerca de estos vertederos suele tener graves problemas de salud debido a la toxicidad de la quema de estos residuos (*ver Figura 26*).

Figura 26: Imágenes de residuos electrónicos en el "vertedero de Europa", Agbogbloshie (Ghana) /Sara Domínguez García. Como denuncia el periodista Mike Anane, estos desechos digitales proceden de envíos ilegales de: la UE, EEUU, Canadá y Australia. Además, crean graves problemas de salud en la población por su alta toxicidad.

Por ello es necesario introducir una perspectiva crítica sobre el impacto medioambiental de todo el ciclo de la IA y los SDAs a auditorías algorítmicas. De esta manera, se podrá evaluar el daño que este tipo de tecnología causa en la población, más allá de los sesgos y errores en el código.

114 Lehuedé, S. (2021). The coloniality of collaboration: sources of epistemic obedience in data-intensive astronomy in Chile. *Information, Communication & Society,* 1-16.

115 Quach, K. AI me to the moon... Carbon footprint for ‘training GPT - 3’ same is driving to our natural satellite and back (4 de noviembre de 2020). *Theregister.com* https://www.theregister.com/2020/11/04/gpt3_carbon_footprint_estimate.

116 Fasola, P. (12 de abril de 2022). Agbogbloshie: la discarica di rifiuti elettronici più grande dAfrica. *Lospiegone.com* https://lospiegone.com/2022/04/12/agbogbloshie-la-discarica-di-rifiuti-elettronici-piu-grande-dafrica/.

117 Romaguera, A. (7 de junio de 2022). Ghana sha converti en làbocador més gran del món. *Directa.cat* https://directa.cat/ghana-sha-convertit-en-labocador-mes-gran-del-mon/.

5. Resistencias

Debido a las injusticias que se han destapado con el uso de herramientas de automatización de la toma de decisiones, han surgido tanto a nivel local como global varios actos de resistencias para hacerles frente.

La sociedad civil se ha enfrentado a la discriminación algorítmica de diferentes maneras: demandas judiciales, manifestaciones, peticiones de información, sindicalismo o actividades organizadas. En este capítulo daremos ejemplos de algunos casos de resistencias que se han dado con casos explicados en capítulos anteriores, así como qué leyes nos amparan, qué marcos legales se diseñan y a quiénes protegen.

Los procesos judiciales contra algoritmos se han convertido en una forma de resistencia muy común. La demanda judicial surge cuando una organización o un grupo de individuos se dan cuenta de que han sido discriminados a raíz de un sistema de toma de decisiones. En muchas ocasiones, es difícil saber si detrás de un proceso digitalizado existe dicho sistema, por eso es útil realizar peticiones de información. Por ejemplo, en 2017 la organización Consejo Conjunto para el Bienestar de los Inmigrantes (JCWI por sus siglas en inglés) cooperó con FoxGlove, un grupo de abogadas expertas en tecnología para denunciar al Ministerio de Interior británico (ver sección 3.5.3). Desde 2015, el Ministerio había implementado un algoritmo para evaluar las solicitudes de visado. JCWI observó que ciertas nacionalidades tenían más riesgo de ser categorizadas como sospechosas. Esto vulnera un derecho fundamental en el que toda persona debe ser tratada de manera igualitaria, sin distinción de género, edad, nacionalidad, clase, orientación sexual, capacidad física/mental, religión, etc (legislación británica Equality Act 2010[118]). De esta manera, en 2020, el Tribunal Administrativo dió la razón a JCWI y FoxGlove por lo que el Ministerio de Interior tuvo que dejar de utilizar dicho algoritmo.

En el contexto de la UE, tenemos el caso del iBorderCtrl. Como ya hemos comentado en la sección 3.5.2, este proyecto europeo tenía como objetivo "agilizar" los procesos fronterizos mediante tecnología. Una de las soluciones propuestas era el uso de detectores de mentira basados en IA. Debido a la controversia que este proyecto creó, el eurodiputado Patrick Breyer, del Partido Pirata de Alemania, solicitó los informes sobre ética que todo proyecto financiado por la UE debe realizar. No obstante, la Comisión Europea se negó y Breyer abrió una causa por lo judicial que acabó de manera satisfactoria para el eurodiputado[119].

En el plano nacional, tenemos el ejemplo de BOSCO (ver sección 3.7). BOSCO es un claro ejemplo de resistencia en el ámbito jurídico dentro del contexto español, ya que la Administración se niega a publicar el código que, presuntamente, dejó sin prestaciones sociales a personas elegibles de ellas. CIVIO, una organización por la transparencia de la Administración llevó a juicio este caso pero el tribunal español se negó a ceder información sobre dicho sistema por 'seguridad ciudadana'.

118 Web Legislation.gov.uk https://www.legislation.gov.uk/ukpga/2010/15/contents

119 Vídeo lie detectorfor travelers: Patrick Meyers sues EU for keeping the iBorderCtrl project secret (31 de juio de 2019). *Patrick-beyer.de* https://www.patrick-breyer.de/en/video-lie-detector-for-travelers-patrick-breyer-sues-eu-for-keeping-the-iborderctrl-project-secret/?lang=en (vía https://www.statewatch.org/news/2021/february/eu-secrecy-of-border-control-lie-detector-research-project-examined-in-court/)

Como vemos en los diferentes casos expuestos, las legislaciones sobre igualdad y equidad son las que suelen proteger a las personas de la discriminación algorítmica. No obstante, también existen otras legislaciones enfocadas directamente al uso de datos, como la Ley de Protección de Datos de la UE (GDPR, por sus siglas en inglés) o las propias de cada Estado. Además, el Parlamento Europeo ha propuesto el primer marco legal específico sobre la IA: el AI Act [120]. También se han aprobado leyes específicas para el uso de algoritmos en contextos concretos. Es el caso de la Propuesta No de Ley (PNL) para regular el uso de IA en la frontera española que se aprobó recientemente[121] y que es fruto, en parte, gracias a una campaña iniciada por cientos de organizaciones en este tema[122]. Sin embargo, la regulación no es suficiente. Por ejemplo, a pesar de que en el AI Act se define que toda SDA en el contexto migratorio serán consideradas de alto riesgo, el artículo 83 de dicha regulación establece que las bases de datos implementadas por la UE para el control migratorio (ver sección 3.5.1) quedan exentas de dicha regulación. En el caso de la PNL, es necesario que las instituciones pongan hincapié en la transparencia y el uso de SDAs que no discriminen, pero esto no es suficiente para erradicar la injusticia de la frontera como infraestructura política[123].

Existen otro tipo de resistencias contra las injusticias algorítmicas a parte de las litigaciones. Como hemos explicado en el capítulo 4.6 de impacto climático, los centros de datos consumen grandes cantidades de agua. Las vecinas de Cerrillos (Chile) se organizaron y resistieron a la instalación de un centro de datos en su zona. Dicho proyecto implicaba la explotación de un acuífero de la zona y el consumo de 14 millones de litros diarios de agua. Gracias a su lucha, la empresa tecnológica decidió no instalar dicho centro de datos.

También las comunidades Thacker Pass en Nevada (Estados Unidos) han organizado diferentes movilizaciones en contra de la extracción de litio de sus tierras por empresas tecnológicas. Y en Reino Unido, el estudiantado se manifestó en las calles de Londres mostrando su rabia en contra del Gobierno que decidió el uso de un algoritmo que discrimina en función de la clase social de la escuela (ver sección 3.2.1). Algunas de las consignas que se pudieron leer son: F*ck the algorithm (Que le den al algoritmo) o Algorithm? Elitism (¿Algoritmo? Elitismo).

120 Web Euro Lex. https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=celex%3A52021PC0206.

121 Web Congreso.es (ver página 17). https://www.congreso.es/public_oficiales/L14/CONG/BOCG/D/BOCG-14-D-418.PDF#page15

122 La implantación de la Inteligencia Artificial en frontera y la vulneración de derechos. *Fronterasdigitales.com* https://fronterasdigitales.wordpress.com/

123 Aprobada una PNL para evitar sesgos y promover la transparencia en el uso de sistemas biométricos en la frontera sur (5 de abril de 2022). *Algorace.org* https://algorace.org/2022/04/05/aprobada-una-pnl-para-evitar-sesgos-y-promover-la-transparencia-en-el-uso-de-sistemas-biometricos-en-la-frontera-sur/

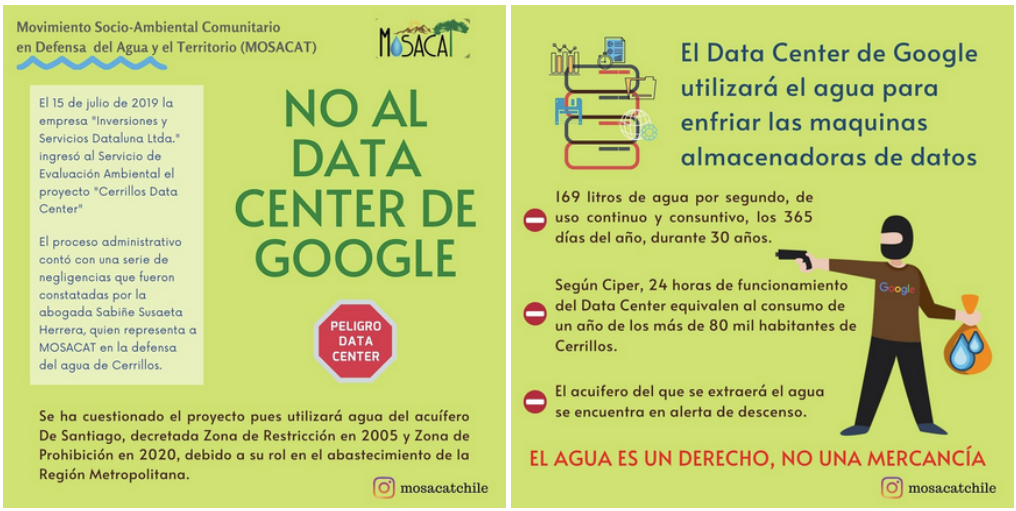


Figura 27: Infografía diseñada por el Movimiento Socioambiental Comunitario (MOSACAT) de Chile para informar a la comunidad del impacto medioambiental del proyecto de un centro de datos. Fuente: https://twitter.com/mosacatchile/status/1267998053144952832

Para luchar contra la tecnoprecariedad, se han ido forjando sindicatos de trabajadores u otro tipo de organizaciones. Por ejemplo, Riders X Derechos es un colectivo que 'lucha por los derechos laborales y las condiciones de vida dignas de los trabajadores de reparto a domicilio'. Fue fundado por un grupo de riders que buscaban mejorar sus condiciones laborales con las respectivas empresas de reparto. Recientemente se aprobó la conocida "Ley Rider" que obliga a las empresas tecnológicas a mostrar los algoritmos que gobiernan a sus empleados. Mediante esta ley, CCOO ha obligado a una de las empresas de reparto a domicilio a mostrar su algoritmo [124]. Otra iniciativa a tener en cuenta en territorio nacional es AlgoRights: 'una red colaborativa para promover la participación de las personas y de las organizaciones de la sociedad civil en el ámbito de la IA' [125]. Su mirada y enfoque se centra en los Derechos Humanos y recientemente co-organizaron junto al Espai Societat Oberta, LaFede.cat y AlgoRace las Jornadas sobre Democracia, Algoritmos, Resistencias' (#JornadasDAR) [126].

A nivel internacional existe el colectivo 'No Tech For Tyrants', que está formado por estudiantes de varias universidades británicas que se niegan a desarrollar tecnología para empresas que oprimen. Por otro lado, Mijente, 'un hogar político para Latinxs y Chicanxs en busca de justicia racial, económica, de género, y climática' realiza campañas en contra de tecnologías para el control migratorio como: Erase the Database [127]. En general, este tipo de iniciativas muestran lo importante de crear y organizar iniciativas, sindicatos, colectivos y comunidades de apoyo para confrontar el uso represivo y opresivo de la tecnología en diferentes contextos.

124 CCOO exige a Glovo que muestre su algoritmo usando la Ley Rider (2 de noviembre de 2022). *Elsaltodiario.com* https://www.elsaltodiario.com/falsos-autonomos/ccoo-exige-glovo-muestre-algoritmo-usando-ley-rider

125 Web de AlgoRights https://algorights.org/

126 Web de las Jornadas sobre democracia, algoritmos y resistencias. https://jornadasdar.org/

127 Web Mijente.net https://action.mijente.net/petitions/chicagoans-say-no-new-cpd-gang-database-in-chicago

6. Conclusiones

La discriminación algorítmica se ha convertido en un problema político y social debido a las consecuencias que la tecnología tiene sobre las personas. Consecuencias que tienen que ver fundamentalmente con las relaciones de poder y privilegios personales y colectivos. Este documento ha hecho un recorrido por las diferentes formas de discriminación dada tanto por la IA como por los sistemas de decisión automatizada analizando varios ámbitos a nivel nacional e internacional (educación, sistema de bienestar, policía predictiva, buscadores de Internet, etc.). En base a este recorrido por las diversas formas de discriminación y el análisis de varios contextos de aplicación de IA, concluímos que:

1. **La IA y los SDAs "son una ideología, no una tecnología"**[128] . a implementación de estas tecnologías diseñadas por una élite blanca y occidental conlleva riesgos, como la reproducción de opresiones históricas (racismo, clasismo, sexismo, etc.) pero también operacionaliza lógicas extractivistas y coloniales: "La propia idea de la IA puede generar una distracción que facilite a un pequeño grupo de tecnólogos e inversores apropiarse de todas las ganancias de un esfuerzo ampliamente distribuido".
2. **Responsabilidad social de desmitificar la IA y SDAs**. Esta tecnología no tiene poderes extraordinarios. Tiene la capacidad de analizar grandes cantidades de datos y encontrar patrones, pero no se convertirá en una tecnología como la que se describe en la ciencia ficción, y a menudo en prensa y política, con autonomía, conciencia, intención, cuerpo, adaptación y demás atributos. Además, no todas las aplicaciones o sistemas automáticos son necesariamente IA o SDAs. La mayoría de sistemas analizados son versiones de algoritmos clásicos, no sofisticados, lo cual no significa que no puedan tener efectos discriminatorios a gran escala. Implementar esta tecnología es cara y lleva tiempo.
3. **La IA o los SDAs no discriminan, las instituciones públicas y privadas y en consecuencia la sociedad, sí.** Se debe escapar de narrativas que dan un poder antropomorfo a dicha tecnología. La IA y los SDAs no son racistas, sexistas, etc., son los equipos de personas que diseñan dicha tecnología, recogen datos y/o los programan, que intencionadamente o no, añaden sus sesgos o reflejan desigualdades sociales, y son las dinámicas estructurales y relaciones de poder las que posibilitan los efectos de la discriminación a través de la tecnología.
4. **Los sesgos raciales de la IA son reflejos del racismo estructural existente**. Debido a la falta de conciencia antirracista dentro del campo de la tecnología, la mayoría de proyectos basados en IA reproducen discriminaciones históricas, como el racismo, pero también implementan lógicas de ordenación y clasificación social.
5. **La IA y los SDAs discriminan más allá del código y los datos.** El ecosistema de esta tecnología tiene un impacto político y medioambiental. Trabajos relacionados con la IA reproducen precariedad laboral, como las etiquetadoras de datos con salarios mínimos o transcriptoras de audios expuestas a material sensible. Además, esta tecnología necesita recursos naturales, como minerales o agua, para construir su infraestructura material, con unas consecuencias materiales lejos de la metáfora inocua de la nube.

128 AI is an Ideology, Not a Technology (15 de marzo de 2020). Wired.com https://www.wired.com/story/opinion-ai-is-an-ideology-not-a-technology/

6. **Resistencia y fiscalización social**. La ausencia de compromiso social y político por parte de sectores públicos y privados frente a los abusos de poder, así como la vulneración de derechos humanos y fundamentales, encuentra hoy día su principal oposición en la organización social e iniciativas que buscan poner fin a los usos e implementaciones perjudiciales para las personas y el medio ambiente, en aras de garantizar una sociedad más justa y un consumo más responsable. La IA y los SDAs necesariamente deben estar en la agenda de los movimientos sociales.
7. **(Des)centrar la tecnología en el análisis antirracista.** La definición de la discriminación principalmente a través del sesgo algorítmico, o incluso equiparándola, presenta múltiples limitaciones pero sobre todo puede servir de distracción en el debate y a la vez como mecanismo de legitimación de sistemas sociotécnicos racistas. El excesivo foco en los componentes técnicos, y en concreto en una parte de ellos que son los datos y algoritmos de decisión, produce una despolitización y reduccionismo deliberados en beneficio de los poderes hegemónicos.

7.Glosario

- **Algoritmo:** Conjunto de reglas programadas para resolver un problema con un ordenador. Un algoritmo puede ser programado expresamente por una persona o creado por otro algoritmo, por ejemplo, con técnicas de inteligencia artificial. Existen diferentes tipos de algoritmos dentro del campo de la informática. En este documento nos centraremos en los algoritmos de *machine learning*.

- **Aprendizaje automático *(machine learning)*:** Campo del conocimiento que estudia algoritmos capaces de encontrar patrones en bases de datos.

- **Aprendizaje profundo *(deep learning)*:** Campo del conocimiento que estudia algoritmos de machine learning con una arquitectura más sofisticada (profunda) capaces de encontrar patrones en datos no estructurados como imágenes o texto.

- **Base de datos:** Estructuras informáticas en las que se almacenan datos ya sean estructurados (por ejemplo, hojas de Excel) o no estructurados (imágenes, textos, audios).
- **Big Data:** Disciplina que estudia algoritmos capaces de analizar gran volumen de datos de distinto tipo rápidamente.

- **Ciencia de datos *(data science)*:** Disciplina que estudia métodos para extraer, procesar y visualizar información de bases de datos.

- **Código fuente o código:** Consiste en una secuencia de líneas de texto entendibles por personas con las que se escribe un programa informático. El código no es un programa en sí, sino que se escribe utilizando un lenguaje de programación y posteriormente se traduce a lenguaje máquina, entendible por los ordenadores, para crear programas.

- **Datificación:** Proceso de transformación de cualquier acción en datos cuantificables.

- **Discriminación algorímica o tecnológica:** Discriminación producida como resultado de la introducción de tecnología o tratamiento algorítmico de las personas que en conjunto con el contexto perjudique a un grupo demográfico. No debe confundirse con sesgo algorítmico ya que una tecnología sin sesgo algorítmico puede resultar discriminatoria en un contexto sociopolítico y legal. Por ejemplo, la identificación dactilar per se no tiene sesgo contra ningún grupo de personas, pero en el contexto de la UE afecta de forma diferente a los derechos y libertades de personas migrantes y racializadas frente a las que tienen ciudadanía europea.

- **Estadística:** Disciplina que estudia la colección, organización, análisis e interpretación de datos. La base teórica de la inteligencia artificial, machine y deep learning se basa en esta disciplina.

- **Logaritmo:** Función inversa a la exponencial (e^x). No confundir con algoritmo.

Inteligencia artificial (IA): Aunque existen múltiples definiciones, podríamos definirla como el campo del conocimiento que estudia la teoría y el desarrollo de sistemas informáticos capaces de realizar tareas con cierto grado de autonomía y adaptabilidad no programadas explícitamente a partir de datos que representen a una tarea.

Racismo: Sistema de dominación global que se basa en la idea de raza como construcción sociopolítica para crear una jerarquía humana con consecuencias simbólicas y materiales. El racismo se vale de diferentes elementos que permiten establecer la diferencia e inferiorizar a la otredad: color de piel, vestimenta, rasgos físicos, religión, etc. El racismo es sistémico y tiene carácter estructural, es decir, está presente en las instituciones y en todos los ámbitos de la sociedad, la cultura, los valores y las relaciones que se establecen.

Sesgo algorítmico: Proceso en el que un algoritmo obtiene resultados diferentes dados diferentes grupos demográficos (raza, género, clase, orientación sexual, religión, etc.) o la intersección de varios de estos.

Sistema de decisión automática (SDA): Proceso automatizado mediante el uso de datos y algoritmos para optimizar la toma de una decisión.

Sistema sociotécnico: Término que enfatiza que cualquier tecnología debe enmarcarse dentro de un contexto social, legal y cultural y que por tanto no se puede analizar su función.

AlgoRace

(Des)Racializando la IA

@algo_race

www.algorace.org

@algorace