Revista Novos Rumos Sociológicos | vol. 10 nº 18 | Ago/Dez/2022
Artigo recebido em 07-11-22 | Aprovado em 04-12-22
ISSN: 2318-1966

DOSSIÊ

# UN CONCEPTO VIAJERO: COMPRENSIONES ACERCA DE LA INTERSECCIONALIDAD EN ESTUDIOS DE SESGO EN LA CIENCIA DE DATOS

## A traveling concept: understandings about intersectionality in studies on bias in data science

Noelia Beltramelli Gula[1]

Camila Ferro[2]

María Goñi Mazzitelli[3]

Lorena Etcheverry[4]

Martín Rocamora[5]

**Resumen**

El uso de grandes volúmenes de datos, el diseño de algoritmos y la utilización de técnicas de aprendizaje automático parecen ser cada vez más frecuentes para la toma de decisiones en diferentes ámbitos como la salud, educación y seguridad pública, entre otros. El campo de estudio de la ciencia de datos, en la línea de investigación de justicia o *fairness*, investiga cómo estos sistemas pueden reproducir y acentuar sesgos y desigualdades presentes en nuestras sociedades. Se han detectado sesgos mayoritariamente de género y raciales en sistemas predictivos de reincidencia criminal, algoritmos que asisten decisiones médicas, entre muchos otros. Recientemente se han manifestado carencias en la detección de sesgos cuando se toman las variables de forma independiente, debido a que las desigualdades pueden presentarse en la intersección de las mismas, siendo indetectables en el análisis individual. En tanto *concepto viajero*, la interseccionalidad viene a permear este campo y habilita la posibilidad de pensar los sesgos que se dan cuando se cruzan múltiples categorías de opresión, complejizando los análisis hasta el momento centrado en un solo eje de opresión a la vez. Aquí identificamos y analizamos, a partir de una revisión sistemática de la bibliografía disponible sobre *justicia interseccional*, cómo la interseccionalidad "viaja" hacia la ciencia de datos, provocando nuevas preguntas en torno a la justicia. Destacamos, entre las conclusiones, la necesidad de conformar grupos interdisciplinarios de trabajo para generar diagnósticos y alternativas que se nutran de la interseccionalidad en toda su riqueza y complejidad.

**Palabras clave:** Sesgos; Justicia Interseccional; Ciencia de Datos.

[1] CICADA (Espacio Interdisciplinario - Universidad de la República). E-mail: noe.beltramelli1701@gmail.com
[2] CICADA (Espacio Interdisciplinario - Universidad de la República). E-mail: camif197@gmail.com
[3] Comisión Sectorial de Investigación Científica (Universidad de la República). E-mail: sadja27@gmail.com
[4] Instituto de Computación (Facultad de Ingeniería - Universidad de la República). E-mail: lorenae@fing.edu.uy
[5] Instituto de Ingeniería Eléctrica (Facultad de Ingeniería - Universidad de la República). E-mail: rocamora@fing.edu.uy

**Abstract**

The use of large volumes of data, the design of algorithms and the use of machine learning techniques seem to be more and more frequent for decision-making in different areas such as healthcare, education and public safety, among others. The strand of fairness in data science investigates how these systems can reproduce and intensify biases and inequalities that exist in our societies. Gender and racial biases have been detected in predictive systems for criminal recidivism, algorithms that assist medical decisions, among many others. Recently, deficiencies have been manifested in the detection of biases when the variables are taken independently, since inequalities can appear in their intersections, being undetectable in individual analysis. As a traveling concept, intersectionality permeates this field, enabling the analysis of biases that occur when multiple categories of oppression intersect, making the analyzes, so far focused on a single axis of oppression at a time, more complex. Here we identify and analyze, based on a systematic review of the available literature on intersectional justice, how intersectionality "travels" into data science, prompting new questions around justice. We highlight, among other conclusions, the need to form interdisciplinary work groups to create diagnoses and alternatives that are nourished by intersectionality in all its richness and complexity.

**Keywords:** Biases; Intersectional Fairness; Data Science

## Introducción

¿Qué sesgos están presentes en el desarrollo de sistemas que asisten procesos de toma de decisiones basados en datos? ¿Cómo inciden de forma interconectada las diferentes estructuras de desigualdad en estos procesos? ¿Qué innovaciones y nuevas preguntas implica la incorporación de un enfoque interseccional en estos análisis? Los avances generados recientemente en el campo de la ciencia de datos, la inteligencia artificial y el aprendizaje automático vienen de la mano de una creciente utilización de estas tecnologías para asistir procesos de toma de decisiones que anteriormente eran realizados enteramente por personas. La aplicación de estos sistemas influye cada vez más en la vida cotidiana de las personas, pero esto no siempre implica que las personas estén al tanto de cómo se toman algunas decisiones que pueden tener incidencia directa sobre su vida. La salud, la seguridad pública y la educación son algunas áreas que han sido permeadas por sistemas que intentan emular los procesos humanos de toma de decisiones basándose en datos de forma automatizada y aparentemente neutral. Recientemente, múltiples estudios desde el campo de la ciencia de datos han mostrado la presencia de sesgos y consecuencias desiguales que estos sistemas generan sobre diferentes grupos de la población al no tomar en cuenta los diferentes sistemas de desigualdad que

están presentes en la sociedad y que interactúan. En el marco de estos análisis, persiste un esfuerzo por visibilizar las desigualdades de género, raza, clase – entre otras – que producen, sin una intencionalidad aparente, la aplicación de diferentes sistemas automáticos. De este reconocimiento se han propuesto diferentes iniciativas que comienzan a delinear metodologías para auditar, mitigar y corregir los sesgos en las diferentes partes del ciclo de vida de los sistemas (DWORK et al., 2012; CHOULDECHOVA, 2017; BAROCAS; SELBST, 2016).

Desde la ciencia de datos, el concepto de justicia algorítmica busca conectar las nociones estadísticas de paridad y equidad con nociones morales y filosóficas (BAROCAS; HARDT; NARAYANAN, 2019). Los estudios sobre justicia en este campo suelen dividirse en dos conceptualizaciones y metodologías diferentes sobre cómo se entiende la misma. Por un lado, se encuentra la justicia individual y, por otro, la justicia grupal (DWORK et al., 2012). La primera implica asegurar que aquellos individuos que son similares sean tratados equitativamente respecto de la tarea de clasificación. La segunda identifica ciertas variables protegidas y define a su interior los grupos protegidos, es decir, aquellas poblaciones que comparten una característica por la que podrían eventualmente sufrir las consecuencias negativas de los sesgos si los hay. Ejemplos de grupos protegidos pueden ser las mujeres en la variable género, las personas no blancas en la variable raza, etc.

En este marco, se detectaron sesgos raciales en sistemas de predicción de reincidencia criminal aplicados en Estados Unidos (CHOULDECHOVA, 2017) así como en algoritmos que asisten procesos de toma de decisiones en el área de la salud (RAJKOMAR et al., 2018), entre otros.[6] Pero estos trabajos suelen enfocarse en una sola variable de análisis a la vez, es decir, en un solo eje de desigualdad de forma independiente. Esto hace que no se logre detectar las consecuencias para los sujetos que pertenecen a más de uno de los grupos protegidos a la vez, para los que

[6] Obermeyer y otros (2019) analizan los sesgos raciales presentes en los servicios de salud que utilizan algoritmos que parten de proxis aparentemente efectivos.

sabemos que la mera acumulación o suma de resultados no logra visualizar las desigualdades que se entrecruzan en la realidad.

Joy Buolamwini (2017) es una de las primeras en problematizar el análisis en un solo eje en este campo. Parte del marco interpretativo de las estructuras de dominación y desigualdad, que propone la interseccionalidad (CRENSHAW, 1989), para analizar el desempeño desigual en las intersecciones de las variables género y etnia-raza en algoritmos de clasificación a partir del reconocimiento facial. A este trabajo le siguen otros que proponen diferentes formas de incluir la mirada interseccional al análisis de grupos, ya sea de forma explícita o implícita.[7] Además, algunos trabajos recientes analizan cómo se utiliza la interseccionalidad y qué críticas surgen para poder seguir complejizando la justicia en la ciencia de datos (KONG, 2022).

El artículo se estructura de la siguiente forma. Comenzamos ubicando el concepto de justicia en la ciencia de datos y la división entre justicia individual y grupal, para luego concentrarnos en la interseccionalidad como concepto viajero que se traslada de las ciencias sociales para permear la ciencia de datos en busca de tener herramientas que den cuenta de las consecuencias desiguales de la aplicación de estas tecnologías. Luego de retomar una breve genealogía del término y las complejidades tanto teóricas como metodológicas que supone, pasaremos a los resultados de la revisión bibliográfica. Veremos las definiciones de interseccionalidad utilizadas, así como qué entienden por justicia interseccional y los casos de estudio a los que se aplica. Finalmente, concluimos con una serie de interrogantes que parten del planteo de Youjin Kong (2022) sobre las interpretaciones que se realizan de la interseccionalidad en la inteligencia artificial y posibles líneas de profundización local de estos temas.

## 1. *Fairness* e interseccionalidad: procesos iterativos de construcción conceptual para llegar a definiciones y soluciones más justas

### 1.1 Justicia en la ciencia de datos

[7]Buolamwini, 2017; Kearns *et al.*, 2018; Cabrera *et al.*, 2019; Foulds *et al.*, 2020; Mathioudakis *et al.*, 2021; Ghosh; Genuit; Reagan, 2021.

La dimensión de justicia o *fairness* en torno a la ciencia de datos, la inteligencia artificial y el aprendizaje automático es uno de los ejes centrales en los que se enmarcan los trabajos tanto teóricos como metodológicos que tienen entre sus objetivos visualizar, analizar y transformar las dinámicas que reproducen o acentúan desigualdades y sesgos en los sistemas de datos. Se entiende por sesgo el perjuicio sistemático que las salidas de estos sistemas de decisión basados en datos producen sobre ciertas poblaciones en comparación con otras. Dado que el abordaje de la justicia en este campo es un tema cuyo desarrollo es reciente – centrado en los últimos años de esta década – aún no hay un consenso establecido sobre las definiciones teóricas y los posibles abordajes prácticos. Como explican Buolamwini y Gebru (2018), las definiciones se basan en supuestos, dados por el contexto, y en la búsqueda de precisión de los sistemas desarrollados, por lo cual no hay una definición que se ajuste a todos los casos.

A pesar de esto, podemos acercarnos a una conceptualización general que es utilizada en la literatura especializada. *Fairness* puede ser definido como "*la ausencia de perjuicio o preferencia por un individuo o grupo basado en sus características*" (GHOSH; GENUIT; REAGAN, 2021, p. 2) por parte de un sistema al momento de realizar una tarea de clasificación, predicción y toma de decisiones basada en datos.

El desarrollo del concepto abarca desde cuestionamientos éticos y teóricos al momento de definir qué implica que ciertos procesos cumplan o no con parámetros de justicia (MITTELSTADT *et al.*, 2016; SKIRPAN; GORELICK, 2017; BIRD *et al.*, 2016; BINNS, 2018), el tratamiento y resultados desiguales de diferentes poblaciones (BOLUKBASI *et al.*, 2016), así como ejemplos de estudios de casos específicos (DRESSEL; FARID, 2018; RAJKOMAR *et al.*, 2018). Estos últimos toman diferentes definiciones de *fairness*, realizan un diagnóstico sobre el tratamiento desigual y eventualmente buscan diferentes métricas para evaluar si un sistema es justo o no. Por otro lado, en algunos casos, también se busca diseñar posibles mecanismos para mitigar los sesgos encontrados en las diferentes etapas del ciclo del sistema. Finalmente, algunos trabajos han analizado las diferentes definiciones disponibles hasta el momento, agrupándolas de

acuerdo a características comunes (VERMA; RUBIN, 2018) así como también explicitando las suposiciones de las que parte cada una y las posibles incompatibilidades entre definiciones (MITCHELL *et al.*, 2020).

### 1.2 Justicia individual, justicia grupal y la complejización del área de estudio

El desarrollo teórico-metodológico sobre justicia en este campo se divide generalmente en dos grupos de conceptualizaciones: justicia individual y justicia grupal (o paridad estadística). Las definiciones agrupadas dentro de justicia individual pueden entenderse como aquellas que garantizan que las personas que son "similares" con respecto a la tarea de clasificación o predicción, obtengan resultados similares (DWORK *et al.*, 2012). La similitud entre individuos se define a través de métricas de distancia, donde la distancia entre la distribución de resultados por individuo debe ser igual a la distancia entre individuos con respecto a la tarea de clasificación o predicción (VERMA; RUBIN, 2018).

Por su parte, las definiciones de justicia grupal (DWORK *et al.*, 2012; SIMOIU; CORBETT-DAVIES; GOEL, 2017) buscan garantizar alguna forma de paridad estadística (por ejemplo, entre resultados positivos/negativos, errores o valores positivos predictivos) para los miembros de diferentes grupos generados a partir de atributos protegidos (por ejemplo, género o raza). Sin embargo, este abordaje presenta algunas desventajas que han sido planteadas y discutidas a partir de la incorporación implícita o explícita de la noción de interseccionalidad. Trabajos como el de Kearns y otros (2018) proponen que existen casos en los que las métricas de *fairness* grupal utilizadas para evaluar ciertos clasificadores solo permiten evaluar su desempeño en un número reducido de grupos. Esto puede dejar pasar como justas aquellas clasificaciones injustas que realiza el sistema para ciertos subgrupos compuestos por más de un atributo protegido (por ejemplo, la intersección de un grupo protegido de la variable género y uno de la variable raza). Este trabajo habla de justicia por subgrupos, sin explicitar la idea de interseccionalidad de fondo. Otros trabajos como el de Foulds et al. (2019) van a diseñar definiciones y aplicaciones de justicia que incorporen

explícitamente la perspectiva interseccional, definiéndose como justicia interseccional.

## 2. Interseccionalidad

### 2.1 Breve genealogía de la interseccionalidad

Las reflexiones en torno a las múltiples estructuras de dominación actuando de forma conjunta y diferencial sobre algunos grupos (como las mujeres negras, los hombres negros pobres, las mujeres indígenas, etc.) y las carencias de los marcos interpretativos disponibles para visualizarlas se venían gestando en diversos contextos históricos previos a que Crenshaw (1989) acuñara el término interseccionalidad (VIVEROS VIGOYA, 2016).

Mara Viveros Vigoya rastrea la presencia de esta mirada crítica desde Olympia de Gouges en Francia, en 1791, con la declaración de los derechos de la mujer. En el contexto latinoamericano, Clorinda Matto de Turner (1899) en Perú denunció los abusos que sufrían particularmente las mujeres indígenas. Más recientemente, la declaración de la Colectiva del Río Combahee, los textos de María Lugones y otras feministas descoloniales son también antecedentes de la gestación de esta mirada crítica a los análisis que simplifican las experiencias en un solo eje de desigualdad y no complejizan la realidad diversa de los grupos oprimidos.

El concepto de interseccionalidad, formulado por Kimberlé Crenshaw (1989), permite reconocer la complejidad de los procesos formales e informales que generan las desigualdades sociales. Crenshaw propone hablar de "experiencias interseccionales" como aquellas que viven las mujeres negras que no son recogidas como grupo, cuando los ejes de representación de las relaciones de poder se estructuran en términos "sexuales" o en términos "raciales" (CRENSHAW, 1989). La interseccionalidad revela que las desigualdades son producidas por las interacciones entre los sistemas de subordinación de género, orientación sexual, etnia, discapacidad, situación socioeconómica, entre otras. Los ejes de subordinación social no generan experiencias de subordinación que deban entenderse una por añadidura de la otra, sino que la intersección es "constitutiva" – genera experiencias singulares y concretas de subordinación.

De esta forma, invitan a analizar en qué manera la interconexión inextricable de sexismo, racismo y clasismo – junto con otros sistemas de subordinación – contribuye en la creación, mantenimiento y refuerzo de las desigualdades formales e informales que sufren las mujeres (VIVEROS VIGOYA, 2016; CRENSHAW, 1989).

En 1989 Crenshaw define la interseccionalidad de forma práctica, intentando visualizar la multidimensionalidad de las experiencias de las mujeres afroamericanas que era invisibilizada por los marcos interpretativos legales y los análisis que consideran solamente un eje de desigualdad. La intención de la autora en ese momento era poder construir categorías legales concretas que habilitaran el análisis de múltiples dimensiones de desigualdad operando de forma unitaria, compleja e interconectada, a partir de un caso en que mujeres afroamericanas denunciaban a la compañía General Motors por excluirlas sistemáticamente y no considerarlas para los puestos de trabajo.

Una vez definido el concepto, este comenzó a difundirse y desbordar al análisis inicial. Hill Collins (2015) habla de la interseccionalidad como proyecto amplio de conocimiento que suele tener diferentes aplicaciones. En primer lugar, el estudiar el objeto de investigación, en segundo, como estrategia analítica de instituciones y estructuras y, por último, como praxis crítica para la justicia social. Sin embargo, más allá de estas oscilaciones, la interseccionalidad supone un avance en la complejización de las múltiples experiencias que se engloban en el paraguas de "mujer", así como en las formas de considerar las múltiples desigualdades más allá de la suma de las mismas (VIVEROS VIGOYA, 2016, a partir de Dorlin, 2009). Este avance trascendió las ciencias sociales y las humanidades y ha ido permeando diferentes campos de estudio, dando lugar a nuevas y diversas creaciones académicas.

## 2.2 Discusiones teóricas y metodológicas alrededor de la interseccionalidad

### 2.2.1 Dilemas teóricos

Dorlin (2009) plantea que la investigación en torno a la interseccionalidad ha transitado partiendo de dos aproximaciones: una

fenomenológica y otra analítica. La primera engloba el trabajo de Crenshaw y aquellos que parten de las experiencias concretas de dominación y la segunda implica partir de una mirada global que entiende que todas las estructuras de dominación – el género, la clase y la raza y otras – operan de forma interconectada e interdependiente. Estas dos formas polarizantes de acercarse al enfoque reducen su potencial teórico y político. Mientras la primera se enfoca en buscar categorías concretas que determinen los tipos de dominación cruzada de forma particular, la segunda asume que todos los sujetos son producidos por relaciones de clase, género y raza pero cabe destacar que los sujetos no vivencian las relaciones interconectadas de dominación de la misma forma según su posición en la estructura.

Por su parte, Hill Collins (2015) encuentra seis puntos focales de producción académica sobre interseccionalidad para intentar visualizar tanto los temas ampliamente desarrollados, como también las ausencias o carencias en el desarrollo de las últimas tres décadas en este campo y las suposiciones implícitas al momento de partir de la interseccionalidad en la academia. Una de las carencias de la literatura estadounidense es la falta de atención a la clase como categoría, frente a un énfasis en el género y la raza. Viveros Vigoya (2016) argumenta que esto responde al contexto de surgimiento del concepto, en el que se privilegia el análisis de las desigualdades raciales como factor diferenciador. Sumado a esto, hay una necesidad de visualizar el lugar de la interseccionalidad a nivel epistemológico en las investigaciones, ya que se define como perspectiva, en algunos casos, como tipo de análisis, como concepto, entre otras. Por último, en línea con lo anterior, desde las críticas latinoamericanas, debemos tener presentes las implicancias que tiene trasladar la interseccionalidad a un contexto fuera del que fue acuñada. La interseccionalidad se creó en el mismo contexto de dominación que busca transformar; por lo tanto, es necesario visualizar que el importar marcos de análisis no implique invisibilizar realidades particulares de ciertos territorios, ya que estaríamos utilizando un concepto con un potencial político transformador como una herramienta de reproducción colonial de subalternidades (RIVERA CUSICANQUI, 2010).

### 2.2.2 Dilemas metodológicos

Cuando los estudios se proponen abordar una temática desde la interseccionalidad, una de las cuestiones a problematizar es la búsqueda de una metodología que responda adecuadamente a las exigencias que esta implica. El uso de categorías preexistentes para definir y estudiar diferentes grupos de las sociedades o la búsqueda de nuevas formas de acercarse a estos complejizando su conformación (atravesada por múltiples identidades y desigualdades) es el punto de partida a partir del cual McCall (2005) clasifica tres enfoques metodológicos diferentes dentro de los estudios en interseccionalidad. En este trabajo utilizaremos la clasificación de McCall para visualizar de qué forma abordan los problemas aquellos estudios de justicia en la ciencia de datos que incorporan la perspectiva interseccional.

McCall define la metodología como "un conjunto coherente de ideas sobre la filosofía, los métodos y los datos que delinean el proceso de investigación y de producción del conocimiento" (2005, p. 1774). En este marco, define los enfoques como: *complejidad anticategórica, intracategórica e intercategórica*. El primer enfoque parte de cuestionar las categorías existentes, entendiendo que no responden a la complejidad que existe en la realidad y, por lo tanto, los estudios que de ellas surjan tendrán resultados parciales. Esta deconstrucción implica cuestionar los criterios de clasificación de los grupos, que parecen reflejos estáticos de la realidad cuando son construcciones del lenguaje que pueden estar reproduciendo jerarquías e invisibilizando experiencias particulares.

El segundo enfoque implica visibilizar y estudiar aquellos grupos que se encuentran en los límites de las categorías definidas, es decir, que bajo las categorías preexistentes, pertenecen a dos grupos dentro de la misma categoría y esto implica que seguramente tengan características particulares que no puedan reducirse a la suma de ambos grupos. Finalmente, el enfoque intercategórico toma los grupos de las categorías preexistentes y los utiliza como punto de partida para el estudio de la naturaleza de las relaciones de desigualdad existentes, entendiendo que los grupos no son estáticos, sino que son dinámicos y contextuales (MCCALL, 2005, a partir de Glenn, 2002).

### 2.3 Interseccionalidad: un concepto viajero

A partir de su formulación, el término interseccionalidad ha "viajado" hacia diferentes disciplinas y áreas de conocimientos. La noción de "concepto viajero" (BAL, 2002) sirve para comprender cómo la interseccionalidad, en este caso, ha sido útil para producir nuevas preguntas y profundizar en el análisis de diversos problemas que son abordados por diferentes áreas de conocimiento (LA BARBERA, 2016).

En el campo de la ciencia de datos, el concepto es utilizado para analizar los sesgos y las discriminaciones producto de los falsos binarios que benefician a una sola de las perspectivas parciales. Uno de los aportes de la incorporación de este concepto a este campo implica reconocer las carencias y las consecuencias adversas de los análisis que solo consideran un eje de opresión a la vez. No considerar la existencia interconectada e interdependiente de las desigualdades y las estructuras sociales que les dan origen y las sustentan tiene como consecuencia la reproducción y el reforzamiento de las normas sociales discriminatorias y los estereotipos existentes. A pesar de que han surgido intentos en Estados Unidos y algunos países de Europa de regular la inteligencia artificial, las normas hasta ahora formuladas no parten desde la incorporación de la perspectiva de género, siendo solo algunas las que mencionan al pasar la necesidad de atender a esta categoría en particular. Las recomendaciones desarrollan de forma amplia la noción de sesgos,[8] sin explicitar los ejes fundamentales de desigualdad a los que se debe atender, ni cómo estos ejes generan consecuencias particulares según la intersección en la que se encuentren los individuos afectados, en contextos y territorios particulares.

En este campo de investigación, Buolamwini y Gebru (2018) son pioneras en exponer, a través de un caso de estudio específico, los sesgos y las consecuencias desiguales que se generan en la aplicación de las

[8] Declaración de Montreal para un Desarrollo Responsable de la Inteligencia Artificial: https://www.canasean.com/the-montreal-declaration-for-the-responsible-development-of-artificial-intelligence-launched/
Recomendaciones del Consejo de la OCDE sobre la IA https://legalinstruments.oecd.org/en/instruments/OECD-LEGAL-0449

tecnologías a la vida cotidiana, a partir de una mirada interseccional.[9] Las autoras analizan tecnologías comerciales de reconocimiento facial y sus resultados notoriamente más deficientes en mujeres de piel oscura que en hombres de piel clara. Este ejemplo es un claro recordatorio de que la equidad es inherentemente interseccional. La construcción de un mundo equitativo, por definición, debe incluir a todos, por lo que no puede considerarse la equidad de un algoritmo solo en función del género. Se deben considerar los grupos interseccionales a través de varios rasgos protegidos: género, raza, nacionalidad, orientación sexual y más. De esta manera, el uso del concepto de interseccionalidad contribuye a complejizar, por un lado, el análisis de los problemas y, por otro, la búsqueda de soluciones que trascienden las estructuras y razonamientos binarios y unilaterales que imperan en este campo.[10] La interseccionalidad viene a poner en entredicho los análisis que buscan reducir a los sujetos de estudio a variables estáticas e independientes. Su incorporación a las investigaciones desafía tanto los abordajes teóricos como la aplicación metodológica de los estudios. Incluso dentro de aquellos que la incorporan, podemos ver diferentes grados de complejización del análisis de las desigualdades. Además, el hecho de que el concepto haya "viajado" de las ciencias sociales hacia la ciencia de datos obliga a que los estudios deban realizarse partiendo de una interdisciplina amplia (HUUTONIEMI, 2010) para poder cuestionar profundamente las estructuras de dominación, cómo operan en el desarrollo de estas nuevas tecnologías y buscar posibles soluciones a las desigualdades reproducidas y/o reforzadas.

Entendiendo la potencialidad de este reciente campo de estudio, la presente revisión bibliográfica tuvo como cometido responder a la pregunta sobre qué definiciones de interseccionalidad se incorporan a la definición de justicia en la literatura del concepto, y qué problemas han sido y están siendo abordados a partir de la conjunción de ambos conceptos.

[9] Buolamwini (2017) habla de *the coded gaze*, o la "mirada codificada", como concepto que busca explicitar que *"*cualquier tecnología creada por humanos reflejará valores individuales o colectivos, prioridades y, sin control, prejuicios*"* (BUOLAMWINI, 2017, p. 17)

[10] Buolamwini y Gebru (2018) y Ryu, Mitchell y Adam (2017) mencionan el reduccionismo que conlleva considerar binariamente la categoría género, sin considerar el espectro de identidades que existen. Sin embargo, al operacionalizar las variables utilizadas, las autoras se ciñen a esta categorización porque es la que utilizan los sets de datos que pretenden auditar.

## 3. Metodología: una revisión sobre la bibliografía actual en justicia interseccional en la ciencia de datos

El grupo de investigación que llevó adelante esta revisión se enmarca dentro del Centro Interdisciplinario en Ciencia de Datos y Aprendizaje Automático de la Universidad de la República[11] y está conformado por integrantes provenientes de las ciencias sociales y de la ingeniería. Particularmente, esta línea de investigación se centra en una mirada crítica de estos sistemas, intentando visualizar las posibles consecuencias desiguales para las personas a las que afectan. Esto no implica desestimar la potencialidad de estas nuevas tecnologías, sino resaltar la necesidad de ampliar los estudios sobre cómo nos afectan cotidianamente y qué alternativas se pueden generar para que todas las personas puedan acceder de forma igualitaria a los beneficios de la aplicación de estas tecnologías. Consideramos pertinente explicitar desde dónde partimos, entendiendo que nuestra mirada sobre la temática está permeada por nuestro contexto, afecta cómo construimos nuestro trabajo, las preguntas que nos hacemos y es, al igual que todas, parcial (HARAWAY, 1991). Entendemos que este grupo ha logrado trabajar la interdisciplina de forma amplia (HUUTONIEMI et al., 2010), trascendiendo áreas de conocimiento que parten de miradas diferentes sobre qué es investigar y cómo hacerlo. Esto nos permite generar una investigación orientada a la instrumentalización, es decir, a responder de forma innovadora a algunos problemas como los que hemos mencionado, pero además implica construir colectivamente una investigación que ponga en juego las diferentes epistemologías de las que partimos como investigadores/as. Esta revisión se enmarca como primer acercamiento para determinar el estado del arte de la literatura sobre justicia interseccional en la ciencia de datos.

Para llevar adelante la revisión sistemática de la literatura nos basamos en los criterios propuestos por Okoli (2015). La revisión se realizó entre los meses de setiembre y noviembre del 2021 y se seleccionaron los años de publicación entre 2016 y 2021 (noviembre inclusive). Los criterios

[11] Espacio Interdisciplinario, Universidad de la República, Montevideo, Uruguay.

que orientaron la búsqueda tuvieron en cuenta los criterios de exclusión e inclusión de la bibliografía que se presentan en la Tabla 1. En cuanto a los criterios de inclusión, se seleccionaron artículos, reportes y libros que incluyeran los conceptos de AI (inteligencia artificial), *fairness* y *machine learning*, de todas las instituciones y locaciones, y que estuvieran en idioma inglés o español. Además, también se incluyeron noticias relevantes sobre la temática. Por su parte, se excluyó la bibliografía previa a 2016, a excepción de los textos "Fairness through awareness" (DWORK et al., 2012) y "Learning fair representations" (ZEMEL et al., 2013) ya que se citan en varios de los textos posteriores a 2016 y, por lo tanto, resultan relevantes para comprender el campo de estudio, que tiene un desarrollo sustantivo en los años delimitados. Para la definición de las palabras claves se partió de los operadores booleanos *"and"* y *"or"*, utilizando los buscadores de Google Scholar, Timbó,[12] Dialnet, Scielo y Latindex. La búsqueda se estructuró en base a las siguientes palabras clave en inglés y sus equivalentes en español: *fairness* AND (*algorithmic* OR *machine learning* OR *data science*) AND (*gender* OR *race* OR *social class*). También se excluyeron aquellos que incluyen los conceptos AI y *machine learning*, pero no *fairness*, y viceversa. El motor de búsqueda con mayor volumen de bibliografía fue Google Scholar, seguido de Timbó. En los restantes no se encontraron resultados con las combinaciones mencionadas.

Cabe mencionar que una de las principales limitantes con respecto a esta búsqueda es que las palabras claves utilizadas se limitaron al español e inglés, que son los idiomas manejados por quienes llevamos adelante la revisión. Sobre este punto, es necesario remarcar que toda la bibliografía encontrada está en inglés y proviene mayoritariamente de instituciones y autores de Estados Unidos, algunos países de Europa y Asia. Esto es relevante al momento de considerar desde dónde partimos para analizar problemas locales sobre consecuencias desiguales del uso de estas tecnologías. Esta característica de la bibliografía no implica necesariamente una limitante, pero sí es un aspecto a considerar para no extrapolar diseños metodológicos o posibles resultados sin antes tener en cuenta, por ejemplo,

[12] https://foco.timbo.org.uy/home

categorías de opresión particulares de ciertos territorios como pueden ser la nacionalidad, la identidad de género, entre otras.

TABLA 1: CRITERIOS DE INCLUSIÓN Y EXCLUSIÓN DE LA BIBLIOGRAFÍA RELEVADA

| Criterios de Inclusión, Exclusión y Evaluación de Calidad | |
|---|---|
| **Criterios de Inclusión** | **Criterios de Exclusión** |
| Artículos, Reportes, Libros que incluyan AI, *machine learning*, *fairness* | Bibliografía previa a 2016 (con excepción de las que se consideran fundamentales para el área de investigación y han sido más citados) |
| Todas las instituciones /empresas | Investigación que incluye AI y *machine learning* pero no *fairness* |
| Todas las localizaciones | Investigación que incluye *fairness* pero no AI y *machine learning* |
| Noticias relevantes basadas en bibliografía del tema | |
| Idiomas: inglés y español | |

Fuente: Elaboración propia.

Luego de una primera búsqueda orientada a textos relacionados a justicia *(fairness)* en el campo de estudio de la inteligencia artificial, aplicando los criterios de inclusión/exclusión propuestos, se seleccionaron 216 artículos. Sobre estos, se incluyeron únicamente aquellos que incorporan el concepto de interseccionalidad. Luego, se realizó una segunda búsqueda en los motores de búsqueda mencionados sumando a las palabras claves ya seleccionadas la palabra "*intersectionality*". En total fueron incluidos 30 artículos que serán analizados en este trabajo. Se destaca que encontramos numerosos textos de revisión del campo de estudio de la justicia (MITCHELL et al., 2020; CATON; HAAS, 2020; BIRD et al., 2016), pero hasta ese momento ninguno abordaba una revisión específica del concepto justicia interseccional. De esta forma, realizar una revisión sistemática del concepto de justicia interseccional contribuyó tanto a

nuestros objetivos específicos como al campo de estudio de justicia en la ciencia de datos.[13]

## 4. Análisis

### 4.1. Definiciones de interseccionalidad y justicia interseccional

Una de las definiciones de justicia interseccional más citadas es la de Foulds et al. (2019), que desarrollan el concepto de *differential fairness*. Este se basa en una definición de justicia extendida en el ámbito del derecho regulatorio.[14] Según los autores, la definición debe satisfacer los siguientes criterios: deben considerarse múltiples atributos protegidos; todos los valores de intersección de los atributos protegidos deberían estar protegidos; se debe garantizar la protección de los valores individuales de los atributos protegidos (por ejemplo "las mujeres", en el caso de "mujeres afro"); la definición debe proteger a los grupos minoritarios; y la definición debería garantizar que las diferencias sistemáticas entre los grupos protegidos, se rectifiquen, en lugar de codificarse (FOULDS et al., 2019).

De esta manera, se propone la creación de subgrupos a partir de la intersección de dos o más variables protegidas, asegurando la paridad estadística entre los mismos. Así, un algoritmo de IA es justo si las probabilidades de los resultados (por ejemplo, ser contratado/a, ser aprobado/a para un préstamo, etc.) son iguales o similares, independientemente de la combinación de intersección de atributos de un grupo, como el género y la raza. Es decir, si las probabilidades son iguales entre todos los subgrupos con diferentes combinaciones de estos atributos. La mayor parte de los textos revisados se concentran en el desarrollo metodológico, generando métricas para auditar e intervenir, y en las cuales

[13] Recientemente, en 2022, con motivo de la realización de la última ACM Fairness, Accountability and Transparency Conference (ACM FAccT) han surgido nuevos trabajos que incluyen el abordaje de justicia interseccional en este campo. En particular, recogemos los aportes críticos de Youjin Kong (2022) sobre cómo se ha utilizado el concepto en este campo de estudio para complejizar la discusión sobre una mirada verdaderamente amplia sobre la justicia en la ciencia de datos y lo que podemos hacer para que esto suceda.

[14] Los autores parten de la regla del 80%, establecida en el Código de Regulaciones Federales (EEUU) como pauta para establecer impactos dispares en violación de leyes antidiscriminatorias. Esta plantea que si la tasa de selección para un determinado grupo, por ejemplo, para la contratación para empleos, es inferior al 80 por ciento de la del grupo con la tasa de selección más alta, hay un impacto adverso en ese grupo.

dicha paridad entre subgrupos se cumple (FITZSIMONS; OSBORNE; ROBERTS, 2018; MATHIOUDAKIS et al., 2021).

En cuanto al término "interseccionalidad", cerca de la mitad de los textos revisados no lo definen explícitamente, sino que lo abordan de forma práctica en la generación de subgrupos según atributos protegidos. Entre quienes sí lo abordan teóricamente, encontramos algunos que retoman a Crenshaw (1989) como referencia desde las ciencias sociales en la definición del concepto y a Buolamwini (2017) como antecedente pionero. El argumento a favor de utilizar este lente teórico es que, según los autores, la combinación de categorías puede dar lugar tanto a diferentes intensificaciones de los prejuicios y sentimientos negativos, como a formas cualitativamente nuevas de marginación y estigmatización (MATHIOUDAKIS et al., 2021).

Retomando el planteo de Dorlin (2009), podemos argumentar que el trabajo de Buolamwini y Gebru (2018) parte de una aproximación fenomenológica, ya que, como relatan las autoras, su estudio surge a partir de darse cuenta de que un sistema de clasificación de rasgos faciales no era capaz de reconocer su rostro (siendo ella una mujer negra). Sin embargo, los estudios posteriores se aproximan a la justicia interseccional desde una perspectiva analítica. Es decir que parten del supuesto de que los sistemas que van a auditar y los sujetos afectados por los mismos son generados en una estructura en la cual se entrecruzan desigualdades de género, raza y clase, entre otras.

Por otra parte, si pensamos en la clasificación de McCall (2005) sobre los abordajes metodológicos de los estudios de interseccionalidad, podríamos decir que todos los revisados en este trabajo se engloban en la categoría de *complejidad intercategórica*. Esto significa que se parte de categorías preexistentes sobre los grupos atravesados por desigualdades interconectadas. Esto permite utilizar los datos de los sistemas sin necesidad de plantear nuevas categorías de clasificación, simplemente realizando las intersecciones consideradas. Veremos que esto puede facilitar el proceso, pero podemos pensar que se están invisibilizando otras categorías también existentes. Además, como va a criticar Kong (2022), hay

un foco generalizado sobre estas categorías identitarias y una ausencia del análisis de las opresiones particulares de esas identidades, es decir, se privilegia la mirada de las identidades y no de las estructuras que las generan y reproducen las relaciones desiguales. Podemos pensar que una de las alternativas que habilite la demanda de complejidad de la que habla McCall (2005) sería cuestionar los criterios de las categorías existentes, así como también encontrar grupos en los que sus miembros pertenezcan a más de una categoría de la misma variable a la vez y requieran, por lo tanto, de una nueva categorización que dé cuenta de su realidad concreta.[15]

### 4.2 Casos de estudio presentes en la bibliografía

En la literatura se encuentran una diversidad de "casos" que sirven a modo de ejemplos para dar cuenta de los sesgos producidos al no tomar en cuenta las diferentes dimensiones que atraviesan a las personas. Podemos clasificar los trabajos en dos subgrupos: aquellos que se basan en sistemas de clasificación, y aquellos que se basan en sistemas de predicción. Existen múltiples trabajos que incluyen la interseccionalidad en la auditoría de resultados desiguales de la aplicación de estas tecnologías. En este artículo mencionamos tres de los textos revisados que dan cuenta de la diversidad de campos en los que se puede aplicar esta mirada.

En cuanto a los sistemas de clasificación, el caso más paradigmático es el del artículo "Gender Shades" (BUOLAMWINI; GEBRU, 2018) ya que es uno de los primeros en abordar la cuestión del sesgo desde un punto de vista interseccional y es uno de los textos más citados por la restante bibliografía. El estudio propone evaluar el desempeño de tres sets de datos comerciales utilizados para entrenar algoritmos de reconocimiento facial teniendo en cuenta las diferencias en cuatro subgrupos interseccionales. Las autoras enfatizan la necesidad de estudiar los sesgos de forma interseccional, ya que el estudio del desempeño a partir de las variables de forma individual no lleva a los mismos resultados. De esta forma, muestran cómo varios sistemas de clasificación de imágenes faciales por género tienen

[15] Ghosh et al. (2021) critican que la creación de subgrupos interseccionales no toma en consideración la pertenencia parcial a un grupo (por ejemplo, una persona que se identifica como multirracial); y tampoco toma en cuenta las variables continuas.

disparidades importantes en su desempeño cuando se consideran de forma conjunta el género y la raza (para diferenciar esta última utilizan una clasificación de tipos de piel de Fitzpatrick). En los resultados de este trabajo, el desempeño de los sistemas para las mujeres de piel oscura es un 30% menos preciso en comparación al desempeño en hombres de piel clara.

Otro trabajo es el de Trinh y Liu (2021), en que evalúan el sesgo en los conjuntos de datos de *deepfake*[16] y en los modelos de detección entre subgrupos protegidos. Utilizando conjuntos de datos faciales equilibrados por raza y género, examinan tres detectores de *deepfake* populares encontrando grandes disparidades en el rendimiento predictivo en la variable etnia-raza, con una diferencia de hasta el 10,7% en la tasa de error entre los subgrupos (TRINH et al., 2021).

Finalmente, el trabajo de Mathioudakis et al. (2021) indaga sobre las políticas de selección *top-k*[17] incorporando la mirada interseccional a una política de acción afirmativa en la educación terciaria. La noción de justicia utilizada se presenta a través de la probabilidad de un candidato ser seleccionado independientemente del conjunto de atributos protegidos, lo que significa que las personas de todos los subgrupos poblacionales definidos tienen la misma probabilidad de ser admitidas. En el caso de las políticas de selección *top-k*, el objetivo es tanto seleccionar a los candidatos que tienen un alto rendimiento esperado, como garantizar que los candidatos procedentes de entornos desfavorecidos estén bien representados. Los autores entienden que este desarrollo presenta un reto computacional debido a la explosión combinatoria de subgrupos potenciales a medida que se consideran más atributos. En el desarrollo del trabajo proponen dos algoritmos para resolver este problema, evaluándose en un conjunto de datos sobre las puntuaciones de las solicitudes universitarias y las admisiones a las licenciaturas en un país de la OCDE. Su conclusión es que es posible reducir significativamente las disparidades en las tasas de

[16] Es una herramienta de inteligencia artificial que, a través del aprendizaje profundo, genera imágenes o videos "que retratan sujetos humanos con identidades alteradas o acciones maliciosas/vergonzosas. Esto ha surgido como un vehículo para la desinformación" (TRINH; LIU, 2021).

[17] Se parte de un predictor entrenado con datos históricos (con indicadores de desempeño laboral, educativo, etc.) para generar una clasificación en la que solo se seleccionan las personas ubicadas en los puestos superiores del ranking obtenido. (MATHIOUDAKIS et al., 2021)

admisión que afectan a los subgrupos poblacionales con tan solo una pequeña pérdida en términos de aptitud de los candidatos seleccionados.

### 4.3 Nuevas preguntas que aporta la interseccionalidad a la ciencia de datos

Como se mencionó previamente, el concepto de justicia interseccional se ubica dentro del paradigma de la justicia grupal, al tiempo que dialoga y problematiza la manera binaria en la cual se entiende la discriminación dentro de las medidas de esta conceptualización. Resulta interesante abordar los textos revisados desde la óptica de Youjin Kong (2022), quien encuentra cuatro problemas en el desarrollo de la mirada interseccional en la búsqueda de justicia en la ciencia de datos. Las críticas de Kong (2022) contribuyen a un intento de complejizar la producción académica y las implicancias políticas de la misma cuando el campo de la ciencia de datos busca conformar el proyecto amplio de conocimiento (HILL COLLINS, 2015) que abarca la interseccionalidad.

El abordaje dominante en la literatura de la justicia interseccional se encuentra preocupado por enfocarse en la problemática de discriminación desde la visión de la intersección de categorías identitarias, olvidando la mirada desde la intersección de las opresiones, es decir, privilegian la mirada identitaria sobre la estructural (KONG, 2022). En este sentido, cerca de la mitad de los textos revisados no proponen una definición precisa del término interseccional, sino que se concentran en explicar que la discriminación puede ocurrir en la intersección de dos o más variables protegidas en los abordajes metodológicos. Estos resultados coinciden con la revisión de Kong (2022). Además, no se suelen explicitar los objetivos de la intervención algorítmica, lo cual puede producir una falta de asidero sobre qué tipo de opresiones se pretende paliar, o cómo estas opresiones se desarrollan en los contextos específicos. El abordaje meramente metodológico enfocado en categorías predefinidas como posiblemente vulneradas no puede perder de vista la aplicación concreta de esos sistemas y el contexto territorial e histórico de las poblaciones afectadas.[18]

[18] Mhasawade y Chunara (2021) proponen una definición en la cual se discute con el concepto de justicia interseccional, en tanto entienden que estos abordajes no toman en cuenta las interacciones causales entre los atributos protegidos. Así, los autores incluyen en sus modelos atributos tanto a

El segundo problema que ubica Kong (2022) es el entendimiento del *fairness* como igualdad mecánica en lugar de la equidad o justicia. Esto es consistente con lo encontrado por Mitchell et al. (2020), los cuales, en una revisión de la literatura sobre justicia, explican que la mayor parte de las evaluaciones de las intervenciones algorítmicas asumen que todos los individuos pueden ser considerados de manera simétrica.

Los mismos autores encuentran que gran parte del debate técnico sobre la equidad algorítmica da por sentado el objetivo social de la aplicación de un modelo y el conjunto de individuos sujetos a la decisión, olvidando que en muchos casos existen objetivos contrapuestos de estas aplicaciones que no pueden ser resueltos con datos más completos o modelados matemáticos, así como tampoco pueden resolver cuestiones éticas sobre qué acciones son aceptables en primer lugar (Mitchell et al., 2020). En nuestra revisión bibliográfica encontramos que esta falta de un objetivo específico y contextual está presente en varios de los textos que abordan la justicia interseccional. Asimismo, esta problemática es mencionada por Sánchez-Monedero, Dencik y Edwards (2020), quienes exponen que confiarle a las matemáticas el significado de conceptos sociales como la equidad puede conllevar el riesgo del *solucionismo tecnológico* (SÁNCHEZ-MONEDERO et al. 2020).

Dentro de los textos de corte metodológico o teórico/metodológico que auditan o intervienen sobre el sesgo (que, a su vez, son la mayor parte dentro de los textos revisados), encontramos que se concentran en la cuestión estadística del problema por sobre la cuestión política. Es decir, se centran en la construcción y problematización de métodos estadísticos, tanto para auditar como para intervenir en sesgos, y no abordan la reflexión sobre los objetivos y alcances de la intervención algorítmica en cuestión.

Al igual que en la investigación de Xivuri y Twinomurinzi (2021), encontramos que hay una necesidad de pensar el problema de justicia en la ciencia de datos en sectores particulares, ya que en su gran mayoría se aborda el problema desde una mirada teórica y metodológica general, que

nivel individual (género, raza) como a nivel macro (nivel socioeconómico, lugar de residencia), y buscan mecanismos causales entre el sesgo y la interacción entre ambos niveles.

pretende contribuir a áreas diversas. Abordar el problema de justicia desde un punto de vista contextualizado es imprescindible si entendemos que los problemas de justicia no son únicamente una cuestión de tipo tecnológica, sino, fundamentalmente, un problema ético y moral, y por lo tanto las intervenciones dependen de los objetivos, del sector y los actores que las lleven adelante. Un ejemplo claro es que, si bien encontramos que son ampliamente utilizadas las bases de datos de COMPAS[19] (FITZSIMONS et al., 2018; FOULDS et al., 2019; CABRERA et al., 2019; YANG; STOYANOVICH, 2020; KOBAYASHI; NAKAO, 2020; JIN et al., 2020) y UCI[20] (CABRERA et al., 2019; FOULDS et al., 2019), las mismas se utilizan para testear las métricas desarrolladas, y no se aborda desde una pregunta de investigación concreta, lo cual entendemos puede resultar en un alcance teórico limitado.

El tercero y el cuarto problemas se encuentran íntimamente ligados. Por un lado, Kong (2021) se pregunta hasta dónde es posible dividir en subgrupos poblacionales sin que se llegue a un nivel individual. Este mismo problema exponen Ghosh et al. (2021) como limitaciones al abordaje interseccional, aunque desde un punto de vista metodológico, ya que entienden que la creación de un número combinatoriamente grande de subgrupos conduce inevitablemente a subgrupos que tienen un número muy pequeño de miembros. Sumado a lo anterior se encuentra el dilema de qué punto de vista adoptamos para seleccionar las variables que vamos a considerar como protegidas sin que el mismo sea arbitrario. Esto es consistente con la investigación realizada, ya que observamos que las variables protegidas se utilizan muchas veces sin una problematización de cómo se ha seleccionado a las mismas, bajo qué criterios, ni qué mecanismos se hallan detrás de los sesgos observados. Estas variables son, en general, el género y la raza. Sin embargo, en ningún caso se problematiza sobre cómo se decide utilizar estas variables y no otras, ni cómo estas variables están construidas. Consideramos que esto se debe

[19] Es la base de datos de esta herramienta de gestión de casos y apoyo a la toma de decisiones utilizada por los tribunales de Estados Unidos para evaluar la probabilidad de que un acusado de un crimen se convierta en reincidente.

[20] El repositorio de aprendizaje automático de UCI es una colección de bases de datos, teorías de dominio y generadores de datos que utiliza la comunidad de aprendizaje automático para el análisis empírico de algoritmos de aprendizaje automático.

fundamentalmente a que los textos proponen, en su mayoría, métricas con las cuales auditar o intervenir sobre el *fairness* en modelos y ambientes hipotéticos, en lugar de realizar experimentos en contextos reales y orientados por objetivos específicos.

Complejizando esta última crítica, y en el espíritu de que este trabajo también busca ser un esfuerzo por poner este tema sobre la mesa en el contexto latinoamericano, no debemos dejar pasar que, dentro de los trabajos revisados, todos provienen del norte global. Los estudios de caso y ejemplos situados se encuentran en contextos que presumiblemente difieren bastante de las realidades de esta región. Esto no pretende desestimar la bibliografía existente, sino utilizarla como base y motivación para continuar indagando, en primer lugar, dónde están presentes los sistemas descritos en nuestras realidades locales y qué consecuencias desiguales están teniendo o podrían tener cuando se aplican acríticamente algoritmos producidos en cierto contexto a otros con características diferentes. Desde la aplicación de la perspectiva interseccional a estos temas, nos toca pensar cuáles serían los ejes relevantes para el análisis tanto de estudios de caso de poblaciones afectadas negativamente, como para la generación de planes o instructivos para la aplicación de estos sistemas de la forma más justa posible para todos los grupos involucrados.

Finalmente, si partimos de que los sistemas desarrollados por la ciencia de datos son desarrollados por personas que viven en sociedades marcadas por estructuras de desigualdad, cabe cuestionarnos quiénes están pensando hoy en día sobre estos temas. La mayoría de los autores de los textos revisados provienen de diferentes ramas de la ingeniería y la ciencia de datos. Sin embargo, los cuestionamientos, las auditorías y las soluciones alternativas que surgen en la línea de investigación de justicia interseccional ponen en juego temas ampliamente desarrollados por las ciencias humanas y sociales como son la ética, la justicia, la equidad y los múltiples ejes de desigualdad y opresión y que no son referenciados (NOBLE; ROBERTS, 2020).

En esta revisión mencionamos que una gran parte de la bibliografía no hace referencia de forma explícita a la definición de interseccionalidad,

además de que suelen centrarse en aspectos identitarios de las categorías de opresión y no en las estructuras que las generan y reproducen. Esto puede deberse a la falta de cientistas sociales en estos equipos de investigación, donde mayoritariamente se realizan abordajes metodológicos sin antes cuestionar realmente el trasfondo estructural de las desigualdades que se busca analizar. Para lograr una incorporación compleja del concepto de interseccionalidad a la ciencia de datos y poder pensar soluciones más justas que tengan en cuenta los aspectos y consecuencias sociales de estos procesos, es fundamental que se conformen equipos interdisciplinarios en los que se piensen estos problemas desde diversos puntos de partida epistemológicos. Esto habilita a nuevas formas de construcción de conocimiento, orientadas a pensar nuevas preguntas y también nuevas soluciones a los sesgos y desigualdades que hoy en día se codifican y reproducen en sistemas de toma de decisiones basados en datos que permanecen ocultos bajo un velo de aparente precisión y neutralidad.

## Conclusiones

Este artículo se propuso analizar la bibliografía que resulta del viaje del concepto de interseccionalidad desde las ciencias sociales hacia el campo de estudio de la justicia en los sistemas de toma de decisiones basados en datos y asistidos por aprendizaje automático en la ciencia de datos. Pudimos observar que gran parte de la bibliografía disponible no define la interseccionalidad ni parte de las estructuras que generan las desigualdades que se buscan visualizar, sino que abordan la justicia interseccional de forma práctica, elaborando metodologías para auditar y corregir los sesgos que afectan diferencialmente a los grupos conformados por más de una categoría protegida a la vez. Además, la interseccionalidad se aborda con experimentos que buscan probar la justicia o injusticia de los algoritmos o la eficiencia de las soluciones, pero no necesariamente se parte de casos concretos u objetivos contextualizados en los que se visualicen variables protegidas específicas que puedan arrojar luz sobre realidades de grupos afectados negativamente por estos sistemas de forma concreta.

Consideramos que partir de casos concretos en los que se detecten inequidades entre grupos, como el que da origen a la investigación de Buolamwini (2017) abre la posibilidad a considerar nuevas variables y nuevos atributos protegidos (como la nacionalidad, situación de discapacidad, etc.) que no hayan sido definidos de antemano. Sumado a esto, todos los textos parten de un abordaje metodológico de complejidad intercategórica (MCCALL, 2005); esto permite utilizar las bases de datos tal cual se encuentran disponibles, con variables definidas de forma binaria. Sin embargo, pensar en aplicar abordajes intracategóricos o incluso anticategóricos implica cuestionar las fronteras de las variables históricas y los procesos de construcción de los datos disponibles y los sesgos presentes también en esos procesos. Poner en entredicho estas construcciones nos lleva a relativizar las decisiones que toman los equipos de investigación cuando crean las bases de datos que, como muestra la bibliografía, sobrerrepresenta a ciertas poblaciones y subrepresenta otras, pudiendo sesgar los resultados que se basen en esta información. Estos sesgos no son intencionales, sino que son resultados de acciones de personas que viven en sociedades en las que las estructuras de desigualdad favorecen a ciertos grupos y oprimen a otros. Como primer paso para poner estos procesos en cuestión y buscar formas más justas de llevarlos adelante, visualizamos la interdisciplina y la construcción de equipos demográficamente diversos como condición para multiplicar los puntos de vista a partir de los cuales se generan estos sistemas de toma de decisiones que tienen efectos sobre cada vez más aspectos de nuestra vida cotidiana.

La interseccionalidad como punto de partida crítico tiene el potencial de desafiar las nociones actuales de justicia en torno a la ciencia de datos y sus consecuencias, pero esto también implica hacernos preguntas y hacer más igualitarios los pasos previos a esos resultados. Desde quiénes son las personas que desarrollan estos sistemas (quiénes en lo que respecta a: de qué territorios provienen, cuáles son sus realidades socioeconómicas, culturales y académicas), con qué fines concretos, qué datos históricos se tomarán en cuenta, cuáles serán las variables a definir como protegidas y cómo se evaluarán los resultados.

**Referencias**

BAL, Mieke. **Conceptos viajeros en las humanidades**: una guía de viaje. Murcia: Cendeac, 2002.

BAROCAS, Solon; SELBST, Andrew. Big Data's Disparate Impact. **California Law Review**, v. 104, n. 3, p. 671-732, 2016.

BAROCAS, Solon; HARDT, Moritz; NARAYANAN, Arvind. **Fairness and Machine Learning**: Limitations and Opportunities. fairmlbook.org, 2019. Disponible en: https://fairmlbook.org/. Acceso 12/11/2022.

BINNS, Reuben. Fairness in Machine Learning: Lessons from Political Philosophy. Conference on Fairness, Accountability, and Transparency. **Proceedings of Machine Learning Research**, v. 81, p. 1-11, 2018.

BIRD, Sarah. et al. Exploring or Exploiting? Social and Ethical Implications of Autonomous Experimentation in AI. 3rd **WORKSHOP ON FAIRNESS, ACCOUNTABILITY, AND TRANSPARENCY IN MACHINE LEARNING**. Nueva York, nov. 2016.

BOLUKBASI, Tolga. Man is to Computer Programmer as Woman is to Homemaker? Debiasing Word Embeddings. **Proceedings** of the 30th Conference on Neural Information Processing Systems, p. 4356-4364. Barcelona, España, dic. 2016.

BUOLAMWINI, Joy. **Gender Shades:** Intersectional Phenotypic and Demographic Evaluation of Face Datasets and Gender Classifiers. Tesis (Maestría en Media Arts and Sciences), Massachusetts Institute of Technology, Cambridge, MA, EUA, 2017.

BUOLAMWINI, Joy; GEBRU, Timnit. **Gender Shades: Intersectional Accuracy Disparities in Commercial Gender Classification**. Conference on Fairness, Accountability, and Transparency. **Proceedings of Machine Learning Research**, v. 81, p. 1–15, 2018.

CABRERA, Angel Alexander et al. Discovery of intersectional bias in Machine Learning using automatic subgroup generation. Debugging Machine Learning Models Workshop. **INTERNATIONAL CONFERENCE ON LEARNING REPRESENTATIONS**, Nueva Orleans, 2019.

CATON, Simon; HAAS, Christian**. Fairness in machine learning**: A survey. arXiv: Learning, 2020. Disponible en: https://arxiv.org/abs/2010.04053. Acceso 12/11/2022.

CHOULDECHOVA, Alexandra. Fair Prediction with Disparate Impact: A Study of Bias in Recidivism Prediction Instruments. **Big Data**, v. 5, n. 2, p. 153-163, 2017.

CRENSHAW, Kimberle. Demarginalizing the Intersection of Race and Sex: A Black Feminist Critique of Antidiscrimination Doctrine, Feminist Theory and Antiracist Politics. **University of Chicago Legal Forum**, n. 1, Article 8, 1989.

DORLIN, Elsa. Introduction: Vers une épistémologie des résistances. *In* : DORLIN, E. (ed.), **Sexe, race, classe, pour une épistémologie de la domination**. París PUF, 2009. p. 5-20.

DRESSEL, Julia; FARID, Hany. The accuracy, fairness, and limits of predicting recidivism. **Science Advances**, v. 4, n. 1, eaao5580, 2018.

DWORK, Cynthia et al. Fairness through awareness. **Proceedings** of the 3rd Innovations in Theoretical Computer Science Conference, p. 214-226, 2012.

FITZSIMONS, Jack; OSBORNE, Michael; ROBERTS, Stephen. Intersectionality: Multiple Group Fairness in Expectation Constraints. Workshop on Ethical, Social and Governance Issues in AI. 32nd **CONFERENCE ON NEURAL INFORMATION PROCESSING SYSTEMS** (NIPS), 2018.

FOULDS, James et al. **An Intersectional Definition of Fairness**. arXiv:1807.08362, 2019. Disponible en: https://arxiv.org/pdf/1807.08362.pdf. Acceso 12/11/2022.

GHOSH, Avijit; GENUIT, Lea; REAGAN, Marry. Characterizing Intersectional Group Fairness with Worst-Case Comparisons. 2nd WORKSHOP ON DIVERSITY IN ARTIFICIAL INTELLIGENCE (AIDBEI). **Proceedings of Machine Learning Research**, n.142, p. 22-34, 2021.

HARAWAY, Donna. **Ciencia, Cyborgs y Mujeres**: La reinvención de la naturaleza. Valencia: Cátedra, 1991.

HILL COLLINS, Patricia. Intersectionality`s Definitional Dilemmas. **The Annual Review of Sociology**, n. 41, p. 1-20, 2015.

HUUTONIEMI, Katri. Analyzing interdisciplinarity: Typology and indicators. **Research Policy**, v. 39, n. 1, p. 79–88, 2010.

JIN, Zhongjun et al. MithraCoverage: a system for investigating population bias for intersectional fairness. **Proceedings** of Sigmod - International Conference on Management of Data, p. 2721-2724, 2020.

KEARNS, Michael et al. Preventing Fairness Gerrymandering: Auditing and Learning for Subgroup Fairness. **Proceedings** of the 35th International Conference on Machine Learning, v. 80, p. 2564-2572, 2018.

KOBAYASHI, Kenji; NAKAO. Yuri**. One-vs.-One Mitigation of Intersectional Bias:** A General Method to Extend Fairness-Aware Binary Classification. arXiv:2010.13494, 2020. Disponible en: https://arxiv.org/pdf/2010.13494.pdf. Acceso 12/11/2022.

KONG, Youjin. Are “Intersectionally Fair” AI Algorithms Really Fair toWomen of Color? A Philosophical Analysis. **Proceedings** of the 2022 ACM Conference on Fairness, Accountability, and Transparency, p. 485–494, 2022.

LA BARBERA, Maria Caterina. Interseccionalidad, un “concepto viajero”: orígenes, desarrollo e implementación en la Unión Europea. **Interdisciplina**, v. 4, n. 8, p. 105-122, 2016.

MATHIOUDAKIS, Michael et al. Intersectional Affirmative Action Policies for Top-k Candidates Selection. **ACM CONFERENCE ON INFORMATION AND KNOWLEDGE MANAGEMENT**, October 19–23, 2020, Galway, Ireland. arXiv:2007.14775. Disponible en: https://arxiv.org/pdf/2007.14775.pdf. Acceso 12/11/2022.

MCCALL, Leslie. The complexity of intersectionality. **Signs: Journal of Women Culture and Society**, v. 30, n. 3, p. 1771-1800, 2005.

MHASAWADE, Vishwali; CHUNARA, Rumi. Causal Multi-Level Fairness. **Proceedings** of the 2021 AAAI/ACM Conference on AI, Ethics, and Society, p. 784.794.

MITCHELL, Shira et al. **Prediction-Based Decisions and Fairness**: A Catalogue of Choices, Assumptions, and Definitions. arXiv:1811.07867v3 , 2020. Disponible en: https://arxiv.org/pdf/1811.07867.pdf. Acceso 12/11/2022.

MITTELSTADT, Brent D. et al. The ethics of Algorithms. **Big Data & Society**, v. 3, n. 2. 2016.

NOBLE, Safiya; ROBERTS, Sarah. Transforming the Culture: Internet Research at the Crossroads. **Bulletin of the IEEE Computer Society Technical Committee on Data Engineering**, v. 43, n. 4, p. 3-10, 2020.

OBERMEYER, Ziad et al. Dissecting racial bias in an algorithm used to manage the health of populations. **Science**, v. 366, n. 6464, p. 447–453, 2019.

OKOLI, Chitu. A Guide to Conducting a Standalone Systematic Literature Review. **Communications of the Association for Information Systems**, v. 37, Article 43, 2015.

RAJKOMAR, Alvin et al. Ensuring fairness in machine learning to advance health equity. **Annals of Internal Medicine**, v. 169, n. 12, p. 866-872, 2018.

RIVERA CUSICANQUI, Silvia. **Ch’ixinakax utxiwa:** Una reflexión sobre prácticas y discursos descolonizadores. Buenos Aires: Tinta Limón, 2010.

RYU, Hee Jung; MITCHELL, Margaret; ADAM, Hartwig. **Improving Smiling Detection with Race and Gender Diversity**. ArXiv, abs/1712.00193, 2017. Disponible en: https://arxiv.org/abs/1712.00193. Acceso 12/11/2022.

SÁNCHEZ-MONEDERO, Javier; DENCIK, Lina; EDWARDS, Lilian What does it mean to 'solve' the problem of discrimination in hiring? Social, technical and legal perspectives from the UK on automated hiring systems. **Proceedings** of the 2020 Conference on Fairness, Accountability, and Transparency, Jan. 2020, p. 458–468, 2020.

SIMOIU, Camelia; CORBETT-DAVIES, Sam; GOEL, Sharad. The problem of intra-marginality in outcome tests for discrimination. **The Annals of Applied Statistics**, v. 11, n. 3, p. 1193-1216, 2017.

SKIRPAN, Michael; GORELICK, Micha. The Authority of "Fair" in Machine Learning. 2017 **Workshop on Fairness, Accountability, and Transparency in Machine Learning**, arXiv:1706.09976, 2017.

TRINH, Loc; LIU, Yan. **An Examination of Fairness of AI Models for Deepfake Detection**. ArXiv, abs/2105.00558, 2021. Disponible en: https://arxiv.org/pdf/2105.00558.pdf. Acceso 12/11/2022.

VERMA, Sahil; RUBIN, Julia. Fairness Definitions Explained. 2018 ACM/IEEE **INTERNATIONAL WORKSHOP ON SOFTWARE FAIRNESS**, 2018.

VIVEROS VIGOYA, Mara. La interseccionalidad: una aproximación situada a la dominación. **Debate Feminista**, v. 52. p. 1-17, 2016.

XIVURI, Khensani; TWINOMURINZI, Hossana. A Systematic Review of Fairness in Artificial Intelligence Algorithms. *In*: DENNEHY, Denis et al. (eds). **Responsible AI and Analytics for an Ethical and Inclusive Digitized Society**. 20th IFIP WG 6.11 Conference on e-Business, e-Services and e-Society, I3E 2021, Galway, Ireland, September 1–3, 2021, Proceedings. Springer International Publishing, LNCS-12896, 2021. p. 271-284.

YANG, Ke; STOYANOVICH, Julia. Measuring Fairness in Ranked Outputs. **Proceedings** of the 29th International Conference on Scientific and Statistical Database Management, Article n. 22, 2020.

ZEMEL, Richard et al. Learning Fair Representations. **Proceedings** of the 30th International Conference on Machine Learning, PMLR, v. 28, n. 3, p. 325-333, 2013.