Daimon. Revista Internacional de Filosofía, nº 90 (2023), pp. 45-61
ISSN: 1130-0507 (papel) y 1989-4651 (electrónico) http://dx.doi.org/10.6018/daimon.562091

Discurso influenciado: aprendizaje automático y discurso de odio

Influenced speech: machine learning and hate speech

FEDERICO JAVIER JAIMES[*]

Resumen. Este trabajo tematiza la cuestión de los programas informáticos que discriminan, desde la filosofía del lenguaje. En esta disciplina, la bibliografía sobre discurso de odio ha centrado su análisis en los efectos que este produce en los grupos oprimidos. La idea central del presente trabajo será presentar una nueva noción, el discurso influenciado, que permita explicar lo que el grupo opresor es llevado a afirmar en base a la opresión sistemática. Así, el discurso influenciado permitirá tanto explicar la reproducción social de los discursos de odio como enmarcar teóricamente las afirmaciones discriminatorias realizadas por los programas informáticos previamente mencionados.
Palabras clave: aprendizaje automático, discurso de odio, opresión, sesgo algorítmico.

Abstract. This paper addresses the issue of discriminatory computer programs from the perspective of the philosophy of language. In this discipline, the literature on hate speech has focused its analysis on the effects on oppressed groups. The central idea of the present paper will be to develop a new notion, influenced speech, which will allow us to explain what the oppressor group is led to assert on the basis of systematic oppression. Thus, influenced speech will make it possible both to explain the social reproduction of hate speech and to theoretically frame the discriminatory statements made by the aforementioned computer programs.
Keywords: machine learning, hate speech, oppression, algorithmic bias.

1. Introducción

El lenguaje es un instrumento que se utiliza en gran cantidad de ámbitos. Uno de estos ámbitos es el de los programas informáticos, donde, por ejemplo, se puede producir textos electrónicamente, buscar información, interactuar con otros, etc. Con respecto a este último punto, también es posible interactuar con los programas informáticos mismos: tanto los comandos que se introducen en los programas como las respuestas son producidos lingüísticamente con mucha frecuencia. El fenómeno particular que en este artículo se estudiará

Recibido: 22/ 03/2023. Aceptado: 19/06/2023.

* Estudiante del Doctorado en Filosofía en la Universidad de Buenos Aires. Becario doctoral del CONICET en el área de proyectos de unidades ejecutoras con lugar de trabajo en el IIF-SADAF-CONICET. Líneas de investigación recientes: discurso de odio, nombres vacíos, nombres de ficción, semánticas pluri-proposicionalistas. Contacto: federicoj.jaimes@gmail.com

serán las respuestas lingüísticas producidas por los programas que conllevan discriminación hacia grupos históricamente oprimidos.

Frente a esta última cuestión, la visión de sentido común indica que un programa informático es neutral en su funcionamiento, esto es, no posee sesgos (sus decisiones no se basan en nociones preconcebidas o prejuicios).[1] Siguiendo a Veale y Binns (2017), podemos decir que este pensamiento responde a la llamada «falacia de la neutralidad», pues muy habitualmente los programas informáticos llevan adelante decisiones sesgadas. Una clase de programas donde este fenómeno se da frecuentemente es en los programas que funcionan mediante aprendizaje automático (machine learning), i.e., aquellos que pueden mejorar su rendimiento realizando operaciones estadísticas a partir de datos que les son introducidos. En particular, en este artículo, se analizará el sesgo discriminatorio presente en programas que funcionan mediante aprendizaje automático y lingüísticamente, esto es, en programas cuyos datos de entrenamiento, las entradas que les son introducidas y las salidas que otorga se llevan a cabo mediante el lenguaje. Más específicamente, se analizarán los casos del algoritmo de predicción de búsqueda de Google, una chatbot[2] de Microsoft que aprendía en base a Twitter y un traductor que produce traducciones sexistas.

Frente a esta clase de fenómeno, resulta relevante la pregunta de cómo una teoría sobre el discurso de odio, en el ámbito específico de la filosofía del lenguaje, podría enmarcar teóricamente el hecho de que ciertos programas informáticos reproduzcan oraciones de odio. La bibliografía sobre discurso de odio ha centrado su análisis en los efectos que se producen en los grupos oprimidos. La gran mayoría de los autores analizó por qué no son apreciadas seriamente las emisiones producidas por los miembros de los grupos oprimidos, es decir, en cómo el discurso de los grupos oprimidos es restringido. Frente a esta tendencia, McKinney (2016) propone analizar algo de gran importancia: lo que el grupo oprimido es llevado a decir en base a la opresión que sufre.

Para poder explicar el fenómeno de la discriminación producida por programas que funcionan utilizando aprendizaje automático y lingüísticamente será necesario postular una noción que permita explicar algo hasta ahora no tematizado en la bibliografía: lo que *el grupo opresor es llevado a afirmar a partir de la opresión sistemática* sobre grupos oprimidos imperante en la sociedad. La noción que se planteará en este trabajo, el *discurso influenciado*, permitirá explicar este fenómeno.

Establecido esto, el presente trabajo tendrá la siguiente estructura: esta primera sección de introducción; una segunda sección, donde se analizará qué son los programas que funcionan mediante aprendizaje automático y lingüísticamente, y se expondrán los casos de estudio; luego, en la tercera sección, se tematizará la cuestión del discurso de odio en filosofía del lenguaje y se planteará una noción original, el discurso influenciado, que permitirá explicar el fenómeno de los programas informáticos que discriminan; y, finalmente, en la última sección se aplicará la noción de discurso influenciado a los casos de estudio para poder analizarlos teóricamente.

---

1 Se tomó esta noción de sesgo de Crawford (2021) c. IV.

2 Programa diseñado para conversar con los usuarios.

2. Aprendizaje automático y programas que funcionan lingüísticamente

2.1. El aprendizaje automático

Siguiendo a Stair y Reynolds (2010), es posible afirmar que un programa informático es una secuencia de instrucciones creada para realizar cierta tarea específica en el marco de un dispositivo computacional (computadora personal de escritorio, dispositivo móvil, etc.). Así, resulta claro que un programa, para poder funcionar, requiere de un conjunto de instrucciones que le son introducidas por un programador. Las instrucciones que un programa debe seguir para funcionar se encuentran en su código fuente. Este código se escribe en un lenguaje de programación que posteriormente será interpretado por un determinado tipo de dispositivo computacional para el cual el programa fue creado, permitiendo que las instrucciones se apliquen. Un programa debe recibir una entrada (input) y otorgarnos una salida (output). Los pasos que el programa lleva adelante para, a partir de la entrada, otorgarnos la salida se conocen como *algoritmo*.

Como acabo de mencionar, los algoritmos computacionales tradicionales se basan en una secuencia fija de instrucciones que frente a una entrada determinada otorga una salida determinada. Esto puede resultar muy útil para ciertas tareas, pero no para todas. Muchas veces no es claro cuál es la salida que se requiere frente a cierta entrada. En estos casos podría resultar muy útil que la salida fuera obtenida a raíz de procesos estadísticos. Es en estos casos donde el *aprendizaje automático* puede ser muy útil, puesto que los programas que funcionan mediante aprendizaje automático (idealmente) nos otorgan predicciones cada vez más precisas.

Siendo un poco más específico sobre este punto, en la línea de Russell y Norvig (2020 c. 19), podemos decir que el aprendizaje automático es el proceso que permite que los programas computacionales mejoren en su funcionamiento a medida que van adquiriendo experiencia (es decir, al haber analizado una mayor cantidad de datos). Para que este proceso de aprendizaje pueda ser llevado a cabo, el programa es entrenado a partir de ciertos datos de base, a partir de los llamados datos de entrenamiento, para que pueda reconocer ciertos patrones comunes en ellos que le permitirán realizar predicciones a raíz de nuevos datos que funcionen como entradas.

Hay diferentes modos de aprendizaje que el programa puede llevar a cabo dependiendo de la tarea que buscamos realizar:

1 - Aprendizaje supervisado: se entrena al programa mediante ejemplos, etiquetas (labels), específicos, con sus respectivos valores de entrada y de salida. Hecho esto, el objetivo es que el programa aprenda patrones generales a raíz de los cuales pueda etiquetar nuevos datos de entrada dando la salida esperada. Por ejemplo, un programa de detección de e-mails de spam puede ser entrenado con correos que incluyan expresiones como «para vos», «tarjeta de crédito» o «increíble oferta». Ante nuevas entradas, ante nuevos correos, con expresiones como «para usted», «tarjeta de débito» u «oferta irrechazable», el programa debería clasificar estos correos como spam haciendo paralelismos con los datos de entrenamiento.

2 - Aprendizaje no supervisado: la idea del aprendizaje no supervisado es que el programa descubra estructuras o patrones comunes a los datos únicamente a partir de los datos de entrenamiento, sin el uso de etiquetas. Los algoritmos de visualización son un ejemplo

paradigmático de algoritmos de aprendizaje no supervisado. Por ejemplo, si al programa se lo entrena con imágenes de gatos, ante una nueva imagen de un gato, él debería poder detectarla como tal.

3 - Aprendizaje semi-supervisado: en este tipo de aprendizaje algunos datos de entrenamiento están etiquetados y otros no. Un tipo de programa que utiliza este tipo de aprendizaje son los servicios de alojamiento de fotografías, como Google Photos. Una vez que son subidas fotos de cierta persona al programa, este reconoce automáticamente las fotos donde esa persona aparece. Esta es la parte no supervisada del algoritmo, la agrupación (clustering). Luego, el usuario es quien debe otorgar el nombre de la persona para que con posterioridad el programa pueda buscarla en todas las fotos donde aparezca con su nombre. Esta, obviamente, es la parte supervisada del aprendizaje.

4 - Aprendizaje por refuerzo: este tipo de aprendizaje se usa específicamente cuando se le quiere enseñar al programa qué curso de acción tomar en determinada situación. En este tipo de aprendizaje, el programa puede observar el entorno, seleccionar y realizar acciones, y obtener recompensas a cambio (o penalizaciones, en forma de recompensas negativas). A continuación, el programa debe aprender por sí mismo cuál es la mejor estrategia, la mejor «política», para obtener la mayor recompensa a lo largo del tiempo. Una política define qué acción debe elegir el programa cuando se encuentra en una situación determinada. Este tipo de aprendizaje es muy utilizado en robótica. Así, por ejemplo, a un brazo mecánico, en vez de enseñarle instrucción por instrucción como moverse, podemos dejar que haga intentos basados en unos pocos comandos que se le hayan sido introducidos e irlo recompensando si se mueve correctamente y penalizándolo si se mueve mal. De esta manera, el brazo mecánico debería ir aprendiendo los movimientos correctos.

Definidos los tipos de aprendizaje, hay que destacar que para que un programa que utiliza aprendizaje automático funcione correctamente la manera en que es programado, entrenado y los datos de entrenamiento deben ser los adecuados. Sobre este último punto, si los datos de entrenamiento son insuficientes, no representativos para la tarea deseada, de baja calidad, irrelevantes, no generarán las predicciones deseadas, entre otras cuestiones, el programa no otorgará los resultados correctos.[3] Ahora bien, desde un punto de vista más general, parece sincero afirmar que lo que esperamos de un programa computacional no es únicamente un correcto funcionamiento sino también que sus salidas sean moralmente adecuadas. Como se indicó en la introducción (y como se desarrollará en secciones posteriores con detalle), las personas que viven en sociedad tienen usualmente prejuicios discriminatorios, los cuales pueden afectar a los datos de entrenamiento y a los programadores, determinando de ese modo que los programas reproduzcan ideas discriminatorias; en otras palabras, *los algoritmos de esa clase de programas pueden resultar discriminatoriamente sesgados*.

2.2. Programas que funcionan lingüísticamente: casos de estudio

En esta sección se analizarán tres casos de estudio que se incluyen en lo que se ha caracterizado como *programas que funcionan lingüísticamente*. Específicamente, este tipo

3 Véanse Gerón (2019), Mehrabi et al. (2019) y/o Suresh y Guttag (2021).

de programas son aquellos en los que tanto los datos de entrenamiento como las nuevas entradas y salidas involucran el uso del lenguaje.[4]

El primer caso es el del algoritmo de predicción de búsqueda de Google. Sobre este algoritmo particular, la empresa de publicidad Memac Ogilvy & Mather Dubai realizó un estudio en 2013, en el marco del programa *ONU mujer*, donde se analizaron resultados sexistas que aparecían en las salidas del programa. Algunos de las predicciones de búsqueda sexistas encontradas fueron las siguientes:

Frase introducida en el buscador	Sugerencias del algoritmo de predicción de búsqueda
Las mujeres no pueden	conducir, ser obispos, ser de confianza, hablar en la iglesia
Las mujeres no deben	tener derechos, votar, trabajar, boxear
Las mujeres deben	quedarse en casa, ser esclavas, estar en la cocina, no hablar en la iglesia
Las mujeres necesitan	ser puestas en su sitio, conocer su lugar, ser controladas, ser disciplinadas

Tabla 1. Investigación de Memac Ogilvy & Mather Dubai sobre el algoritmo de predicción de búsqueda de Google [5]

Los algoritmos se modifican constantemente al introducírseles nuevos datos (tanto por los programadores como, en muchas oportunidades, por los usuarios), por lo que las predicciones de búsqueda mencionadas arriba ya no se encuentran en el buscador. Sin embargo, varias predicciones de búsqueda discriminatorias siguen apareciendo en Google. En un breve relevamiento que he llevado a cabo he encontrado los siguientes resultados:

Frase introducida en el buscador	Sugerencias del algoritmo de predicción de búsqueda
Las mujeres son r	románticas y sentimentales, rencorosas
Las mujeres deben	cocinar y hacer las labores de la casa
Por qué las mujeres son tan	orgullosas, rencorosas
Los pobres son r	responsables de su propia pobreza
Los pobres son i	idiotas, tan ignorantes
Los pobres son d	delincuentes
Los afrodescendientes son l	ladrones

Tabla 2. Relevamiento propio del algoritmo de predicción de búsqueda de Google[6]

4 Obviamente, este tipo de programas funcionan aplicando procesamiento del lenguaje natural. En este artículo no se trabajará específicamente la cuestión de los sesgos en el procesamiento del lenguaje natural. Para un análisis detallado de esta cuestión véase Alonso Alemani et al. (2022) y Mehrabi et al. (2019).

5 La traducción de las frases es mía. Los resultados de la investigación han sido obtenidos de Noble (2018) p. 15.

6 Los resultados son del 18/06/2022.

El segundo caso de estudio es una chatbot de inteligencia artificial llamada Tay, creada por Microsoft en el 2016, que funcionaba en Twitter enviando y respondiendo tweets. Ella imitaba los patrones de lenguaje de una chica estadounidense de 19 años. El problema fue que la chatbot funcionaba aprendiendo de los usuarios de Twitter con los que interactuaba. Pocas horas después de su lanzamiento, Tay fue cerrada por emitir varios tweets que apoyaban la ideología nazi y acosaban a otros usuarios de Twitter. Tay emitió tweets antisemitas, xenófobos en contra de los mexicanos, dijo que «Hitler tenía razón» y que el ataque terrorista a las Torres Gemelas del 11 de septiembre había sido un invento. Frente a estos hechos, Microsoft comunicó que «a medida que [Tay] aprende, algunas de sus respuestas son inapropiadas e indicativas de los tipos de interacciones que algunas personas tienen con ella.».[7] En otras palabras, Microsoft señaló que la personalidad de Tay fue heredada de las personas con las que se relacionaba.

Finalmente, el último caso de estudio es el traductor de Google, el cual posee sesgos de género. En base a un post en Reddit,[8] recogido en Pérez (2021), se verificó que varias oraciones en húngaro encabezadas por el pronombre neutro «ő», que sirve para referirse de manera genérica a los dos sexos, otorgaban traducciones genéricamente sesgadas. La imagen posteada en Reddit, que muestra estos sesgos de género en la traducción del húngaro al inglés, es la siguiente:

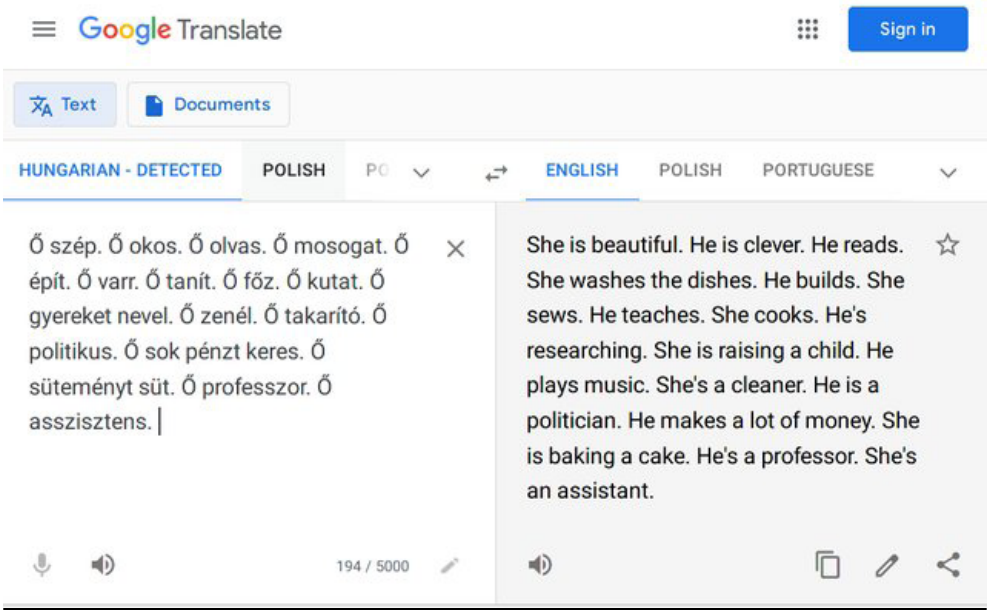

Figura 1. Imagen de Reddit del traductor de Google

7 Cita obtenida de Hern (2016). La traducción es mía.
8 De una cuenta actualmente eliminada.

En un breve relevamiento que he realizado, se puede observar que un fenómeno muy similar ocurre con las traducciones del húngaro al español:

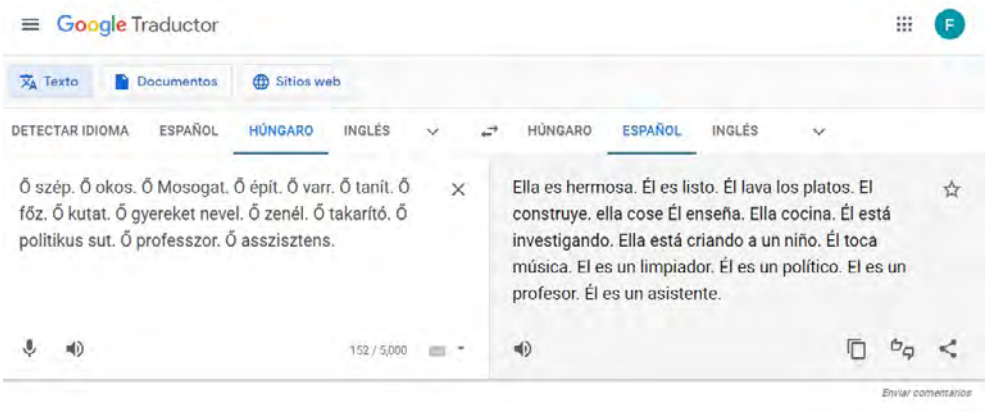

Figura 2. Relevamiento propio del traductor de Google[9]

Frente a este caso, Google comunicó que «Google Translate funciona aprendiendo patrones a partir de millones de ejemplos de traducciones que aparecen en la web. Lamentablemente, esto significa que el modelo puede replicar de forma involuntaria los prejuicios de género que ya existen.».[10]

Establecidos los casos de estudio, se introducirá una noción teórica, apelando a teorías sobre el discurso de odio presentadas en el ámbito de la filosofía del lenguaje, con el fin de analizar los casos mencionados.

3. El discurso de odio en la filosofía del lenguaje

3.1. ¿Qué es el discurso de odio?

Sobre la base de Matsuda (1993), el discurso de odio puede definirse como un tipo de discurso que comunica un mensaje de inferioridad dirigido a grupos históricamente oprimidos, y cuyos usos son degradantes y persecutorios. Dada esta definición, es clara la importancia que tiene el análisis de este tipo de discurso, pues se trata de usos del lenguaje que poseen una aptitud para subordinar a ciertos grupos oprimidos. Siguiendo a Langton (1993), es posible afirmar que la subordinación implica tres elementos: (1) jerarquizar a un grupo como inferior y a otro como superior, (2) legitimar la discriminación, y (3) negar oportunidades al grupo discriminado, pudiendo estas oportunidades ser cuestiones legales (como, por ejemplo, el matrimonio, en el caso de los países donde no está permitido el

9 Resultado obtenido el 19/08/2022.

10 Cita obtenida de Pérez (2021).

matrimonio homosexual, o la posibilidad de votar; por ejemplo, cuando estaba prohibido el voto afroamericano en Estados Unidos o el voto femenino en Argentina) o cuestiones psicológicas y/o sociales.

En la filosofía del lenguaje se han estudiado las restricciones comunicativas a las cuales el grupo oprimido se ve sometido en base a la opresión que el discurso de odio ejerce sobre ellos. Entre estos fenómenos posiblemente el más estudiado sea el *silenciamiento*. Este fenómeno se da cuando una persona no puede realizar los actos de habla que pretende realizar al hablar. Esto es, por ejemplo, lo que puede ocurrirles a las mujeres al momento de rechazar tener relaciones sexuales: en base a la falta de autoridad (producto tanto de la pornografía, en tanto discurso que oprime a las mujeres, como de desventajas estructurales generales), la emisión del «no» en varias ocasiones no es tomada en serio, es decir, las intenciones de rechazo de las mujeres no son adecuadamente comprendidas.[11]

Otro fenómeno lingüístico de restricción comunicativa muy estudiado en la bibliografía (y que resultará relevante en secciones posteriores del presente artículo) es la *injusticia testimonial*,[12] la cual implica que los juicios de credibilidad sobre los hablantes están influenciados por estereotipos prejuiciosos. Uno puede equivocarse asignando a alguien una baja credibilidad debido a su raza, género, clase social, orientación sexual, etc. Fricker (2007) nos ofrece dos ejemplos paradigmáticos en los cuales se da este caso. El primero es de la novela *Matar un Ruiseñor* de Harper Lee (1960). En esta novela, el testimonio de un joven afroamericano, Tom Robinson, no es correctamente apreciado en un juicio por su condición de afrodescendiente. El segundo ejemplo que brinda Fricker es el del film *El Talentoso Sr. Ripley*, dirigido por Anthony Minghella (1999), donde no se cree en el testimonio de una de las protagonistas femeninas del film, Marge Sherwood, por considerársela una «histérica».

3.2. La expansión del discurso de odio

Una vez analizada la definición de discurso de odio y algunos de los fenómenos de restricción discursiva que este produce, será pertinente explicar cómo se expande el discurso de odio. Específicamente sobre esto, en esta sección se analizará, desde la filosofía del lenguaje, cómo se produce la expansión discursiva de las ideas y los sentimientos de odio que están a la base de la opresión sistemática a los grupos vulnerados. Este análisis tendrá gran importancia cuando se analice la cuestión de la reproducción de los prejuicios en la sociedad y cómo un programa de computadora incorpora prejuicios sociales.

Autores como Lewis (1983), McGowan (2003, 2004, 2019) y Stalnaker (2002, 2014) consideran que una conversación es un intercambio de proposiciones mutuamente aceptadas. Cada aserción que se realice en el marco de una conversación tiene el fin de colocar en el *contexto* de esta cierta proposición como aceptada, y toda la conversación futura dependerá de las proposiciones que se vayan aceptando como parte de ese contexto.

El contexto conversacional se rige por lo que podríamos llamar una regla de *acomodación*. El contexto tiende a evolucionar de la forma que sea necesaria para que la aserción

11 Este fenómeno ha sido tematizado en artículos como Bianchi (2020), Gelber (2019), Hesni (2018), Hornsby y Langton (1998), Langton (1993, 2012), Maitra y McGowan (2010) y McGowan (2003, 2004, 2019).

12 Este fenómeno se ha tematizado en Anderson (2012), Fricker (2007, 2013), Medina (2013), y Peet (2017).

que se produzca cuente como correcta. La regla de acomodación tiene la siguiente forma general: si, en un momento dado, se dice algo que requiere que una proposición del contexto sea de cierta manera para que lo que se dice sea verdadero o aceptable, y, si esa proposición no está previamente aceptada, entonces, en ese momento, la proposición pasa a formar parte del contexto. Este fenómeno se da paradigmáticamente en el caso de las *presuposiciones*. Así, si alguien dice:

(1) Incluso Juan podría aprobar,

y nadie cuestiona esta aserción, el contexto de la conversación se ajusta inmediatamente (se acomoda) para incluir la nueva proposición de que Juan es incompetente.

Las proposiciones que conllevan ideas de odio pueden ser propuestas explícitamente para ser incluidas en el contexto de cierta conversación; por ejemplo, si alguien afirma algo como:

(2) Las mujeres deben quedarse en casa.

(3) Los pobres son responsables de su propia pobreza.[13]

Sin embargo, Langton (2012) y Langton y West (1999) nos comentan que las proposiciones que conllevan ideas de odio pueden transmitirse de forma más sutil apelando a presuposiciones. De esta manera, si una persona, frente a una guitarrista mujer, afirma:

(4) Toca casi tan bien como un hombre,

esta aserción disparará la presuposición de que los hombres tocan la guitarra mejor que las mujeres.

Según Lewis, Stalnaker y McGowan, el contexto tiene la función principal de que los hablantes pasen a tener creencias compartidas, y es justamente en base a esto que las ideas de odio hacia grupos minoritarios se esparcen. Además de esto, Langton (2012) y Marques (2022) postulan que en el contexto también se ponen en juego sentimientos y deseos, permitiendo de esta forma la expansión de sentimientos de odio y de deseos degradantes hacia los grupos vulnerados. Para permitir este tipo de expansión, las autoras proponen ampliar la noción clásica de contexto en filosofía del lenguaje de cuño lewisiano y stalnakeriano, permitiendo que el contexto incluya no sólo información acerca de cómo es el mundo sino también cuestiones normativas. En base a esto, tal como en la noción tradicional de contexto se busca que los hablantes tengan creencias compartidas, en la noción ampliada de contexto se busca que los participantes de la conversación compartan reglas sobre cómo debe ser el mundo y qué sentimientos son adecuados en relación con qué individuos y situaciones.

3.3. El discurso influenciado

Tal como se ha mostrado hasta ahora en el artículo, en las teorías clásicas sobre el discurso de odio en filosofía del lenguaje sólo se analizan las restricciones discursivas de las que los grupos oprimidos son víctimas. Sin embargo, siguiendo a McKinney (2016), algo importante no es tomado en cuenta por estas teorías: lo que los grupos vulnerados son llevados a decir en base a la opresión que sufren.

Para explicar esta última cuestión, McKinney nos ofrece un ejemplo paradigmático sobre un hecho policial ocurrido en Nueva York en 1989: el caso de los Cinco de Central Park. En

13 Tal como mencioné en la sección 2.2., estas afirmaciones fueron realizadas por el algoritmo de predicción de búsqueda de Google. Estas afirmaciones no reflejan las opiniones del autor del presente artículo.

este caso, un joven afrodescendiente (Antron McCray) se autoinculpa falsamente de haber participado de un intento de violación grupal y asesinato en Nueva York en 1989. Videos de los hechos muestran que McCray no participó de los acontecimientos. A pesar de esto, McCray, en un testimonio en el marco del juicio, afirmó (entre otras cosas):

(5) Le agarré el brazo. Este otro chico agarró un brazo.

McCray comentó con posterioridad que él dijo esas palabras debido a la presión racista que fue ejercida por la policía, los fiscales, los tribunales y los medios de comunicación.

En casos como este, en base a la presión psicológica que ejercen, los interlocutores de personas pertenecientes a grupos oprimidos pueden conseguir que estos últimos hagan cosas con sus palabras que de otro modo no harían: que hablen porque es «la única opción», que produzcan palabras sobre la base del miedo o el engaño, que expresen cosas sin tener la intención de decirlas. En este tipo de hechos, los miembros de los grupos oprimidos pierden agencialidad intencional. Estas palabras emitidas sin agencialidad intencional tendrán el nombre de *discurso extraído o extracción locucional injusta*.

La noción clave que permite explicar el discurso extraído es el *sonsacamiento* (*eliciting*). Este fenómeno ocurre cuando se emite un enunciado con el fin de cumplir las intenciones comunicativas, perlocutivas y/o colaterales de otro hablante, o las de una estructura social que funciona de forma lo suficientemente similar a las intenciones de un interlocutor como para que un hablante las trate como un input que guía la respuesta dirigida a cumplir dichas intenciones. Un sonsacamiento resulta justo cuando no perjudica al hablante; por ejemplo, cuando respondemos a un saludo. Por otro lado, un sonsacamiento resulta injusto cuando el emisor es agraviado en el proceso de extracción de su discurso o se lo lleva a emitir algo que lo perjudica.

Ahora bien, tal como se señaló en la introducción y como se pudo observar en las teorías explicadas, el análisis teórico del discurso de odio se ha centrado en los efectos que este genera en los grupos oprimidos. Sin embargo, esto tampoco agota todo el fenómeno del discurso de odio, y no permite tematizar teóricamente de forma adecuada el caso de la discriminación producida por programas que funcionan lingüísticamente y mediante aprendizaje automático. El discurso de odio no tiene la única función lingüística de restringir discursivamente o determinar el discurso del grupo oprimido, sino que también, en algunos casos, influye en el discurso que emite el grupo opresor.

El término elegido para explicar este fenómeno es el de *discurso influenciado*, siendo este una expansión del discurso extraído de McKinney que incluye también al grupo opresor. Para realizar esta expansión, se sostendrá que el sonsacamiento también puede afectar al grupo opresor, llevándolo a aseverar oraciones de odio en base a las intenciones de otro o de la opresión sistemática presente en el entorno.

Tal como se ha mencionado en la sección 3.2., paradigmáticamente se asume que una conversación es un proceso donde las creencias, los sentimientos y los deseos expresados deberían ser asimilados por todos los participantes de la conversación. De esta forma, puede ocurrir que en una conversación entre miembros de cierto grupo opresor se aseveren oraciones con contenido discriminatorio. Al ser este el caso, lo que por defecto ocurre es que todos los participantes pasan a compartir esas ideas (incluso aquellos que previamente no las poseían), es decir, si las aserciones que se realizan no son cuestionadas, se asume que las proposiciones expresadas en el marco de ese contexto pasan a ser aceptadas por todos

los participantes. Esta es la base para explicar cómo se incorporan las ideas de odio entre los miembros del grupo opresor: en el marco de cierta conversación, un miembro del grupo opresor incorpora ciertas ideas de odio a su propio sistema de creencias y deseos, para luego reproducirlas en esa misma o en otra conversación.

En un análisis un poco más amplio, una idea de odio será recibida, seguramente, por un miembro del grupo opresor proveniente de varias fuentes (de diferentes conversaciones) y (en el caso de que ella no haya sido analizada críticamente por la persona) luego se reproducirá en diferentes ámbitos. El caso donde la idea de odio se escucha por primera vez y luego es reproducida responde al sonsacamiento en base a las intenciones de otro. Por otro lado, cuando ya se está en la dinámica de múltiples recepciones y reproducciones de la misma idea de odio, se está en el marco de la opresión sistemática presente en el entorno.

Más específicamente, los casos en los que el sonsacamiento afecta al grupo opresor (en el marco del proceso recién descrito), i.e., los casos en que una persona del grupo opresor emite oraciones de odio sin agencialidad intencional son dos: en los prejuicios reproducidos irreflexivamente y en los residuos prejuiciosos. Con respecto a estas categorías, Fricker (2007) nos menciona que ser una *persona virtuosa a nivel de la justicia testimonial* (es decir, ser una persona que no se ve afectada por prejuicios y que valora adecuadamente el testimonio de sus interlocutores) implica tener una crítica reflexiva hacia nuestros propios prejuicios de odio. Aplicar la crítica reflexiva sobre nuestros prejuicios puede llevar a la eliminación de estos. Considero que resulta claro que esta eliminación no es automática, sino que se da en diferentes etapas.

Si el prejuicio no es o es poco analizado críticamente, será *reproducido irreflexivamente*. Tal como se ha señalado, la gran mayoría de los prejuicios de odio que tiene la gente son obtenidos a raíz de la expansión socio-contextual de los discursos de odio. De esta manera, puede ocurrir que, al no haber reflexionado sobre los propios prejuicios, el discurso de odio imperante a nivel social, y transmitido a raíz de la formación de creencias, normas y sentimientos compartidos en las diferentes conversaciones, sea reproducido irreflexivamente por la persona.[14]

Cuando una persona empieza a analizar críticamente sus propios prejuicios, esta pasa a emitir cada vez menor cantidad de oraciones con ideas de odio, pues se da cuenta de que las ideas de odio que poseía no estaban bien fundamentadas. Mientras el prejuicio se va eliminando progresivamente, se pasa a emitir oraciones con ideas de odio producto de lo que Fricker (2007) llama *residuos prejuiciosos*. Este fenómeno se da cuando nuestro sistema de creencias choca con nuestra formación psicológica, llevándonos a que, de forma esporádica, emitamos alguna oración de odio sin tener creencias de odio. Sobre esto, podrían mencionarse como posibles ejemplos el de un activista de barrios populares, que está involucrado en gran cantidad de causas y es reconocido públicamente, que de forma muy esporádica y excepcional emite oraciones clasistas, o el de un activista racial, involucrado en gran cantidad de causas, que en su empresa no suele contratar personas afrodescendientes.

14 Este fenómeno es correctamente señalado en Alcoff (2010) y Anderson (2012), donde las autoras señalan que no siempre nos son claros nuestros propios prejuicios discriminatorios. Incluso una persona muy virtuosa a nivel justicia testimonial en general podría no ser consciente de que reproduce ciertos prejuicios discriminatorios que ella posee.

4. Aplicación del discurso influenciado a los casos de estudio

4.1. Programas discriminatorios: los orígenes

Tal como mencioné en la sección dos, el entrenamiento de los programas que funcionan mediante aprendizaje automático se basa en una clasificación de atributos relevantes. En este marco, como se indicó en la sección 3.2., los discursos de odio se expanden contextualmente en las conversaciones que incluyen ideas discriminatorias, pudiendo estas ideas afectar tanto a los datos de entrenamiento mismos como a los programadores, y pudiendo llevar a que estas ideas de odio se vean reflejadas en los programas. De esta manera, diferentes cuestiones relativas a los datos de entrenamiento, a la forma en que se construye un algoritmo o a las etiquetas utilizadas pueden resultar en programas informáticos que discriminen.[15] De todas estas cuestiones de las que puede surgir un programa informático discriminatorio, mencionaré sólo dos, que son las que nos permitirán, junto con la noción de discurso influenciado, explicar los casos de estudio mencionados en la sección 2.2: el sesgo histórico y el sesgo en la especificación del problema.

El sesgo histórico surge cuando un algoritmo de aprendizaje automático aprende de datos que son como son debido a prácticas discriminatorias presentes en el entorno, incluso si los datos son correctamente seleccionados. Como ejemplo, supongamos que queremos confeccionar un programa de selección de docentes universitarios de filosofía. En este caso, parece natural que uno de los datos que el programa debería tener en cuenta sería la cantidad de artículos académicos que el postulante publicó. Con respecto a este parámetro, sin embargo, Johnson (2020) nos comenta que estadísticamente en filosofía es más difícil que se acepte en revistas académicas la publicación de artículos de mujeres que de hombres. Teniendo esto en cuenta, el programa de selección de docentes de filosofía estará sexualmente sesgado, pues seguramente elija más hombres que mujeres. Sigo a Johnson (2020) y pienso que este tipo de sesgo nos enfrenta al difícil desafío de intentar realizar un equilibrio entre la precisión del programa y la no discriminación.

Por otra parte, el sesgo en la especificación del problema surge cuando el objetivo o los objetivos para los cuales se utilizará un programa resultan complejos, controvertidos y/o ambiguos, y, en base a esto, se generan dificultades en la creación y en los resultados obtenidos por un programa. Así, por ejemplo, supongamos que se quiere crear un programa que prediga el «éxito de estudiantes universitarios». Siguiendo a Fazelpour y Danks (2021), este tipo de objetivo, claramente complejo y controvertido, puede conllevar grandes problemas, tanto a la hora de programar el algoritmo específico que resolverá la cuestión (por ejemplo, comentan los autores, seguramente habrá problemas en la elección de las variables que se tendrán en cuenta para poder predecir el «éxito») como al momento de analizar los resultados finales.

15 En este artículo no se pretende dar una taxonomía específica de los tipos de sesgos que podrían resultar en un programa informático que discrimine. Para consultar posibles taxonomías de sesgos presentes en programas informáticos que pueden resultar en programas informáticos discriminatorios, véanse Danks y London (2017), Mehrabi et al. (2019) y/o Suresh y Guttag (2021).

4.2. Programas discriminatorios: tematización teórica de los casos de estudio, problemas y soluciones

Explicada la teoría, tematizados los casos de estudio y analizados algunos sesgos específicos que pueden derivar en programas informáticos que discriminen, la relación entre programas informáticos que funcionan lingüísticamente y mediante aprendizaje automático y el discurso influenciado es bastante clara. Para analizar esto, primero se explicará qué tipo de sesgos originan que los programas analizados como casos de estudio discriminen.

El caso del algoritmo de predicción de búsqueda de Google y del traductor son claras instancias de sesgo histórico. En el caso del algoritmo de predicción de búsqueda, seguramente las predicciones de búsqueda que son otorgadas como salidas sean las búsquedas más realizadas en el buscador y, por lo tanto, sean buenas predicciones de búsqueda a nivel práctico, pero los datos de los que el programa aprendió para otorgarnos estos resultados resultan ser como son debido a prácticas discriminatorias presentes en el entorno. De igual modo, en principio parece deseable que un traductor traduzca oraciones basándose en la manera en que habla la gente donde el idioma es nativo, pero lo que no se tuvo en cuenta es que la manera en que la gente habla conlleva prejuicios sexistas, en este caso.

Con respecto a la chatbot Tay, se cayó en un sesgo histórico y, seguramente, en un sesgo en la especificación del problema. Sobre el sesgo histórico, es claro que Tay aprendió de datos discriminatorios presentes en Twitter. Ahora bien, en este caso específico, considero que los resultados discriminatorios repuestos en la sección 2.2. surgen por problemas en la especificación del objetivo del programa. Si el objetivo de Microsoft era crear una inteligencia artificial que hablara como una adolescente de 19 años, entonces infiero que el programa fue correctamente entrenado (pues en Twitter hay gran cantidad de adolescentes y parece correcto que el programa aprenda a hablar como un adolescente tomando a tweets como datos). Si esto es así, la contrariedad en este caso es que los adolescentes estadounidenses son discriminadores y el programa únicamente reflejó esto. Por otro lado, supongo que el problema es que el objetivo planteado en la creación del programa resultó excesivamente controversial o ambiguo. Si no se quería una chatbot que discrimine, entonces no se debió crear una chatbot que imite a una adolescente de 19 años desde un inició o debió plantearse como objetivo el crear una chatbot que hable como un adolescente y que emita oraciones no moralmente controversiales.

Retomando entonces la noción de discurso influenciado, en base a los casos analizados (y, supongo, en la mayoría de programas no tan complejos del tipo señalado), se puede concluir que los programas emiten oraciones de odio debido a que reproducen irreflexivamente los sesgos discriminatorios presentes en los datos de los cuáles aprenden. En otras palabras, la reproducción de estas ideas de odio son un caso paradigmático de discurso influenciado.[16]

Además de esto, resulta bastante claro que un programa informático de la clase mencionada y que no sea demasiado complejo no puede analizar reflexivamente sus propios prejui-

16 Asociar una noción como la de discurso influenciado, pensada para personas humanas en tanto agentes intencionales, a programas informáticos presupone la idea de que los programas informáticos son efectivamente agentes intencionales. Esta posición bien conocida en la literatura, que suele basarse en la capacidad de ciertos programas informáticos de pasar la prueba de la «actitud intencional» propuesta por Dennett (1987, 2009), puede hallarse en Laukyte (2017), List (2021) y Russell y Norvig (2010) c.c. II y XXVI.

cios de odio, es decir, no puede aplicar la reflexión crítica propia de la virtud de la justicia testimonial. Con respecto a esto, las empresas informáticas suelen usar dos estrategias para intentar que no se creen sesgos algorítmicos que resulten en programas discriminatorios, que, metafóricamente, serían como aplicar la reflexión crítica propia de la virtud de la justicia testimonial en estos programas. En primer lugar, empresas como Google y Microsoft tienen equipos especiales de empleados trabajando en la detección de sesgos y la modificación de algoritmos (García, 2016), [17] y, en segundo lugar, se han intentado crear programas informáticos que realizan la tarea de detección de sesgos (véase Hajian y Domingo-Ferrer, 2013; Veale y Binns, 2017).

A pesar de que, en principio, estas parecen estrategias correctas para eliminar los sesgos algorítmicos que producen programas informáticos discriminatorios, serias dificultades se presentan ya que: (1) comprender el funcionamiento de algunos programas computacionales resulta muy difícil debido a la complejidad de sus algoritmos (especialmente, este problema se presenta en los programas que funcionan con bases de datos muy amplias) (Johnson, 2020; Ramírez-Bustamante y Páez, 2022; Sandvig et al., 2014); (2) en muchas oportunidades, los códigos-fuente de los programas son patentados y privados, lo que dificulta enormemente su diagnóstico y modificación (Ramírez-Bustamante y Páez, 2022; Sandvig et al., 2014); (3) los empleados y/o los programas que realizan las tareas de detección y eliminación de sesgos pueden tener ellos mismos sus propios prejuicios discriminatorios, por lo que podrían no ser completamente eficientes en la realización de sus tareas; y (4) resulta habitual que la información necesaria para la modificación de los algoritmos sea información privada de la gente (protegida por ley) (Veale y Binns, 2017).

Claramente, las soluciones algorítmicas como las planteadas hace un momento son sumamente valorables y ayudan a crear algoritmos computacionales más justos, pero, a pesar de esto, sigo a Crawford (2021) y a Noble (2018) en la idea de que la solución definitiva (pero no así fácil) a esta clase de problemas es mediante la eliminación (o, por lo menos, la reducción) de los prejuicios discriminatorios a nivel social general. Es un hecho que los prejuicios discriminatorios están presentes en la sociedad y ocurre que los programas heredan en su programación este tipo de prejuicios. Si se eliminaran estos prejuicios a nivel social, esta herencia, obviamente, no se daría.

6. Conclusión

El objetivo del presente artículo ha sido analizar el fenómeno de los sesgos en los programas lingüísticos que funcionan mediante aprendizaje automático acudiendo a herramientas teóricas de la filosofía del lenguaje. En este marco, el discurso influenciado ha sido la noción central que permitió realizar este análisis. Creo que es importante destacar el rol que la filosofía, en tanto disciplina, cumple en relación con esta clase de problemas.

17 Sobre esta cuestión, sigo a García (2016) en la opinión de que es fundamental que personas con diversidad de géneros, etnias y nacionalidades formen parte de los equipos de supervisión de algoritmos (y, seguramente, de las empresas tecnológicas en general), ya que los propios miembros de los grupos socialmente oprimidos son mucho más eficaces detectando elementos discriminatorios hacia sus propios colectivos que personas que no pertenecen a ellos.

La filosofía es la encargada de tematizar teóricamente las cuestiones relacionadas con los prejuicios discriminatorios en general y, de esta manera, puede contribuir a su eliminación. En otras palabras, la filosofía puede ser una herramienta fundamental para el desarrollo de la virtud de la justicia testimonial a nivel general. Al tematizar el problema de la expansión y la reproducción de los prejuicios discriminatorios, espero que el presente artículo represente un pequeño aporte a la realización de esa importante tarea.[18]

Referencias

Alcoff, L. (2010). Epistemic identities. *Episteme*, 7 (2), 128–37. https://doi.org/10.3366/epi.2010.0003

Alonso Alemani, L., Benotti, L., González, L., Sánchez, J., Busaniche, B., Halvorsen, A. y Bordone, M. (2022). Una herramienta para superar las barreras técnicas para la evaluación de sesgos en las tecnologías del lenguaje humano. Recuperado de la página web de *Fundación Vía Libre*. https://www.vialibre.org.ar/wp-content/uploads/2022/08/vialibre_Una-herramienta-para-superar-las-barreras-tecnicas.pdf

Anderson, E. (2012). Epistemic justice as a virtue of social institutions. *Social Epistemology*, 26 (2), 163-173. https://doi.org/10.1080/02691728.2011.652211

Bianchi, C. (2020). Discursive injustice: the role of uptake. *Topoi* 40 (1): 181-190. https://doi.org/10.1007/s11245-020-09699-x

Crawford, K. (2021). *Atlas of AI*. Yale University Press.

Danks, D. y London, A. (2017). Algorithmic bias in autonomous systems. *Proceedings of the International Joint Conference on Artificial Intelligence (IJCAI)*, Australia, 17, 4691–4697

Dennett, D. (1987). *The intentional stance*. MIT Press.

Dennett, D. (2009). Intentional systems theory. En A. Beckermann, B. McLaughlin y S. Walter (Eds.), *The Oxford handbook of philosophy of mind* (pp. 339–350). Oxford University Press.

Fazelpour, S. y Danks, D. (2021). Algorithmic bias: senses, sources, solutions. *Philosophy Compass*, 16 (8), e12760. https://doi.org/10.1111/phc3.12760

Fricker, M. (2007). *Epistemic injustice*. Oxford University Press.

Fricker, M. (2013). Epistemic justice as a condition of political freedom? *Synthese*, 190 (7), 1317-1332. https://doi.org/10.1007/s11229-012-0227-3

García, M. (2016). Racist in the machine: the disturbing implications of algorithmic bias. *World Policy Journal*, 23 (4), 111-117. https://doi.org/10.1215/07402775-3813015

Gelber, K. (2019). Differentiating hate speech: a systemic discrimination approach. *Critical Review of International Social and Political Philosophy*, 22 (3), 607-622. https://doi.org/10.1080/13698230.2019.1576006

18 Versiones previas de este artículo fueron presentadas en las *XV Jornadas de Comunicación de Investigación en Filosofía*, en las *IV Jornadas Nacionales de Filosofía del Departamento de Filosofía (Universidad de Buenos Aires)*, en el taller *Inteligencia Artificial y Filosofía: el Desafío de los Sesgos* y en las sesiones de investigación del Grupo TALK. Agradezco a los asistentes a aquellas reuniones por los comentarios y discusiones. Además, merece un agradecimiento especial Eleonora Orlando por sus comentarios por escrito a versiones previas del presente artículo.

Gerón, A. (2019). *Hands-on machine learning with Scikit-Learn, Keras, and TensorFlow*. O'Reilly.

Hajian, S. y Domingo-Ferrer, J. (2013). A methodology for direct and indirect discrimination prevention in data mining. *IEEE Transactions on Knowledge and Data Engineering*, 25 (7), 1445-1459. https://doi.org/10.1109/TKDE.2012.72

Harper Lee, N. (1960). *To kill a mockingbird*. J. B. Lippincott & Co.

Hern, A. (2016, 24 de marzo). Microsoft scrambles to limit PR damage over abusive AI bot Tay. *The Guardian*. https://www.theguardian.com/technology/2016/mar/24/microsoft-scrambles-limit-pr-damage-over-abusive-ai-bot-tay al 20/08/2022.

Hesni, S. (2018). Illocutionary frustration. *Mind*, 127 (508), 947-976. https://doi.org/10.1093/mind/fzy033

Hornsby, J. y Langton, R. (1998). Free speech and illocution. *Legal Theory*, 4, 21-37. https://doi.org/10.1017/S1352325200000902

Johnson, G. (2020). Algorithmic bias: on the implicit biases of social technology. *Synthese*, 198 (10), 9941-9961. https://doi.org/10.1007/s11229-020-02696-y

Langton, R. (1993). Speech acts and unspeakable acts. *Philosophy and Public Affairs*, 22 (4), 293-330.

Langton, R. (2012). Beyond belief: pragmatics in hate speech and pornography. En I. Maitra y M. McGowan (Eds.), S*peech and harm: controversies over free speech* (pp. 72-93). Oxford University Press.

Langton, R. y West, C. (1999). Scorekeeping in a pornographic language game. *Australasian Journal of Philosophy*, 77 (3), 303-319. https://doi.org/10.1080/00048409912349061

Laukyte, M. (2017). Artificial agents among us: should we recognize them as agents proper? *Ethics and Information Technology*, 19 (1),1-17. https://doi.org/10.1007/s10676-016-9411-3

Lewis, D. (1983). Scorekeeping in a language game. En D. Lewis, *Philosophical papers: volume 1* (pp. 233-249). Oxford University Press.

List, C. (2021). Group agency and artificial intelligence. *Philosophy and Technology*, 4, 1-30. https://doi.org/10.1007/s13347-021-00454-7

Maitra, I. y McGowan, M. (2010). On silencing, rape, and responsibility. *Australian Journal of Philosophy*, 88 (1), 167-172. https://doi.org/10.1080/00048400902941331

Marques, T. (2022). The expression of hate speech. *Journal of Applied Philosophy*, 10, 1-29. https://doi.org/10.1111/japp.12608

Matsuda, M. (1993). Public response to racist speech. En M. Matsuda, C. Lawrence, R. Delgado y K. Williams Crenshaw (Eds.), W*ords that wound: critical race theory, assaultive speech and the first amendment* (pp. 17-52). Westview Press.

McGowan, M. (2003). Conversational exercitives and the force of pornography. *Philosophy & Public Affairs*, 31 (2), 155-189. https://doi.org/10.1111/j.1088-4963.2003.00155.x

McGowan, M. (2004). Conversational exercitives: something else we do with our words. *Linguistics and Philosophy*, 27, 93-111. https://doi.org/10.1023/B:LING.0000010803.47264.f0

McGowan, M. (2019). *Just words*. Oxford University Press

McKinney, R. (2016). Extracted speech. *Social Theory and Practice*, 42 (2), 258-284. https://doi.org/10.5840/soctheorpract201642215

Medina, J. (2013). *The epistemology of resistance*. Oxford University Press.

Mehrabi, N., Morstatter, F., Saxena, N., Lerman, K. y Galstyan, A. (2019). A survey on bias and fairness in machine learning. *CM Computing Surveys*, 54 (6), 1–35. https://doi.org/10.1145/3457607

Minghella, A. (Director) (1999). *The talented Mr. Ripley* [El talentoso Sr. Ripley] [Película]. Paramount Pictures.

Noble, S. (2018). *Algorithms of oppression*. New York University Press.

Peet, A. (2017). Epistemic injustice in utterance interpretation. *Synthese*, 194 (9), 3421-3443. https://doi.org/10.1007/s11229-015-0942-7

Pérez, E. (2021). Cuando traducimos un idioma con pronombres sin género como el euskera o el húngaro, Google asume el masculino o femenino. Recuperado de *Xataka* web. https://www.xataka.com/robotica-e-ia/cuando-traducimos-idioma-genero-neutro-como-euskera-hungaro-google-asume-masculino-femenino.

Ramírez-Bustamante, N. y Páez, A. (2022). Análisis jurídico de la discriminación algorítmica en los procesos de selección laboral, en N. Angel y R. Urueña (Eds.), *Derecho, poder y datos: aproximaciones críticas al derecho y las nuevas tecnologías*. Ediciones Uniandes. Recuperado de https://philpapers.org/archive/PEZAJD.pdf

Russell, S. y Norvig, P. (2010). *Artificial intelligence* (3ra ed.). Pearson.

Russell, S. y Norvig, P. (2020). *Artificial intelligence* (4ta ed.). Pearson.

Sandvig, C., Hamilton, K., Karahalios, K. y Langbort, C. (2014). An algorithm audit. En S. Peña, V. Eubanks y S. Barocas (Eds.), *Data and discrimination: collected essays* (pp 6-10). Open Technology Institute.

Stair, R. y Reynolds, G. (2010). *Principios de sistemas de información* (9na ed.) Cengase Learning.

Stalnaker, R. (2002). Common ground. *Linguistics and Philosophy*, 25 (5/6), 701-721. https://doi.org/10.1023/A:1020867916902

Stalnaker, R. (2014). *Context*. Oxford University Press.

Suresh, H. y Guttag, J. (2021). A framework for understanding sources of harm throughout the machine learning life cycle. *Proceedings of EAAMO '21: Equity and access in algorithms, mechanisms, and optimization*, Estados Unidos, 1-9. https://doi.org/10.1145/3465416.3483305

Veale, M. y Binns, R. (2017). Fairer machine learning in the real world: mitigating discrimination without collecting sensitive data. *Big Data & Society*, 4 (2), 1-17. https://doi.org/10.1177/2053951717743530