Sesgos en inteligencia artificial: un estudio sobre la generación de imágenes a partir de comandos de raza/etnia y género

Denysson Axel Ribeiro Mota
UFCA, Brasil
denysson.mota@ufca.edu.br
Gracy Kelli Martins
UFPB, Brasil
gracykelli@gmail.com
Denise Braga Sampaio
UFBA, Brasil
denisebs@gmail.com

Resumen: Siguiendo los estudios científicos sobre sesgos presentes en algoritmos, después de investigar la influencia en la creación de algoritmos del hecho que, en Brasil, la mayoría de los programadores está compuesta por hombres blancos, cisgénero, heterosexuales, solteros, de mediana edad y sin hijos, este trabajo es el primero de una serie de otros que buscan identificar sesgos en herramientas específicas, en este caso investigando la generación de imágenes por inteligencia artificial (IA). La metodología incluye un enfoque bibliográfico-documental, búsqueda de evidencias encontradas en otras investigaciones sobre sesgos en herramientas digitales y temas similares, seguido de un análisis descriptivo de las imágenes generadas con la herramienta BlueWillow, elegida por ser una herramienta aún no discutido en otras investigaciones y porque era gratuito. Las pruebas se realizaron inicialmente con comandos en inglés y sin asignar género, el idioma fue escogido por el género neutro en profesiones y otros elementos lingüísticos ser algo intrínseco a la lengua. Las palabras utilizadas fueron: [white], [black], [a man], [a woman], [a firefighter], [a nurse], [a doctor], [an inmate], escogidas por ya haber sido identificadas, en la literatura, como propensas a la atribución de sesgo en algoritmos. Los resultados fueron evaluados con la asignación automática de género y raza de la plataforma, y apuntan a la presencia de sesgos, pero requiere mayor investigación, abarcando bases de entrenamiento, prácticas científicas y otras posibles causas. El estudio destaca la importancia de políticas y directrices que promuevan la equidad y la justicia informacional en el desarrollo de estas tecnologías.

Palabras clave: Inteligencia artificial, redes neuronales generativas, algoritmos racistas.

Introducción

La utilización de algoritmos en la Ciencia de la Información no es algo reciente. Lancaster, en 1991, señalaba en su obra Indexing and abstracting in theory and practice, un historial desde la década de 1970, del uso de herramientas computacionales para actividades de extracción, organización y representación de datos e información. Por otro lado, la inteligencia artificial (IA), en su origen, es un poco más antigua, establecida durante la década de 1950 por investigadores como Alan Turing y John McCarthy. Ellos exploraron la idea de crear máquinas que pudieran imitar el comportamiento humano, recopilar y organizar información, dando origen a teorías sobre cómo probar la capacidad de las máquinas, como el Juego de la Imitación, de Alan Turing, conocido simplemente como Test de Turing (Turing, 1950).

A lo largo de las décadas siguientes, la IA pasó por diferentes fases y avances tecnológicos, desde la creación de los primeros programas de ajedrez, como el DeepBlue, en 1996, por parte de la International Business Machines (IBM), hasta el desarrollo de algoritmos de aprendizaje automático y redes neuronales. Actualmente, la IA se aplica en diversas áreas, como la salud, las finanzas, el transporte y la automatización, presentando avances cada vez más notables. Los orígenes de la inteligencia artificial fueron fundamentales para impulsar la investigación y el desarrollo de esta área multidisciplinaria, que continúa transformando la sociedad y fomentando la innovación tecnológica. Sin embargo, su existencia no está exenta de cuestionamientos, ya que, a diferencia de la idea de neutralidad de las tecnologías, las IA están imbuidas de elementos políticos y sociales en su proceso de creación.

Basándonos en tales inferencias, nuestro objetivo general es explorar los sesgos presentes en las IA, específicamente en los servicios de generación de imágenes, centrándonos en los cuestionamientos y usos inadecuados de imágenes con sesgos de racismo/sexismo. Como práctica metodológica, nos dedicamos a aplicaciones específicas, como la generación y reconocimiento de imágenes, utilizando el método bibliográfico-documental para el análisis de los resultados obtenidos, con énfasis en producciones científicas sobre IA, algoritmización y sesgos tendenciosos en el uso de tecnologías de la información y comunicación (TIC).

También como parte de las prácticas metodológicas, realizamos investigaciones experimentales en un servicio de generación de imágenes por IA, en busca de datos que fundamenten nuestros cuestionamientos, partiendo de la hipótesis de que los algoritmos de generación de imágenes, al igual que los de reconocimiento de imágenes, pueden presentar sesgos raciales, de género, culturales, etc. Por lo tanto, los resultados son incorrectos o injustos, "[...] cuyos sesgos están enmascarados tanto por la propia tecnología (marcada por prácticas informacionales invisibles para sus usuarios) como por la confianza y creencia datista en la neutralidad tecnológica" (Bezerra, Costa, 2022, p. 3)

Buscamos identificar si la generación de imágenes, mediante los avances en los algoritmos de IA, como las redes neuronales generativas y los modelos de lenguaje, ha permitido la creación de imágenes realistas a partir de las instrucciones o descripciones proporcionadas

por los usuarios. Esta capacidad de generación de imágenes personalizadas y bajo demanda proporciona numerosas oportunidades, pero también plantea preocupaciones sobre la posible existencia de sesgos en los resultados producidos. Ante este contexto, nuestra pregunta de investigación es la siguiente: ¿Existen sesgos en la generación de imágenes por IA, a través de comandos específicos, al igual que ocurre en los procesos de identificación de imágenes? Con esta pregunta, nuestro objetivo es analizar los sesgos existentes en la representación y producción de imágenes por parte de la IA, especialmente en lo que respecta a los marcadores de raza y género.

Establecemos este objetivo y planteamos esta pregunta teniendo en cuenta que la perpetuación de estereotipos, prejuicios y desigualdades genera consecuencias perjudiciales y muchas veces irreparables para la sociedad, perpetuando las desigualdades y perjudicando a grupos marginados. Por ello, esta perpetuación debe ser constantemente debatida y combatida. Esta investigación se justifica en la medida en que reconoce que comprender y evidenciar la existencia de sesgos en la generación de imágenes por IA es fundamental para mitigar posibles problemas éticos y sociales generados por las TIC en los actuales entornos informacionales.

Racismo en el reconocimiento automático de imágenes

Varios estudios han señalado a lo largo de los años cómo las IA tienden a representar de manera diferente a hombres y mujeres, personas negras y blancas, y perjudicialmente a las personas negras. Joy Buolamwini y Timnit Gebru (2018) analizaron las herramientas de reconocimiento facial de IBM, Microsoft y Face++ y evidenciaron los márgenes de error de identificación, destacando cómo son mayores para las mujeres en comparación con los hombres, y para las personas negras en comparación con las personas blancas. La categoría de mujer negra se ve afectada negativamente por ambos márgenes de error.

Tarcízio Silva (2020a) utiliza resultados similares al evaluar la herramienta de Google para la representación de imágenes, presentando dos conjuntos de imágenes. En el primer conjunto, que muestra un equipo para medir la temperatura, en la imagen con una persona asiática, las etiquetas [Tecnología] y [Dispositivo Electrónico] se asignan con un 68% y un 66% de certeza, respectivamente; mientras que a la persona negra se le asigna la etiqueta [Arma] con un 88% de certeza. En el segundo conjunto, hay un recorte de la mano de la persona negra de la imagen anterior, que también se identifica con la etiqueta [Arma], con un 61% de certeza; mientras que en la imagen digitalmente alterada para que la piel parezca clara, la etiqueta de [Arma] se reemplaza por [Herramienta] con un 55% de precisión.

Abeba Birhane, Vinay Prahbu y Emmanuel Kahembwe (PRABHU; BIRHANE, 2020; BIRHANE; PRAHBU; KAHEMBWE, 2021), por otro lado, examinan bases de datos de imágenes comúnmente utilizadas para entrenar estas IA para la representación, identificando elementos problemáticos como ofensas raciales y de género, imágenes con contenido íntimo o sexual, muchas veces no autorizadas, e imágenes de niños, indicando una clara presencia de misoginia, pornografía y estereotipos negativos, además del riesgo de pérdida de privacidad en el uso de

estas bases. Según Bezerra y Costa (2022, p. 4), "no hay razón, por lo tanto, para pensar que los algoritmos estarían exentos de los sesgos que infestan el orden social que los precede".

Decidimos entonces seguir el camino inverso: en lugar de identificar las imágenes con palabras clave o categorías, como se describió anteriormente, se proporcionarán a las IA de generación de imágenes las palabras clave/categorías deseadas y se verificarán sus respuestas a esas instrucciones. Los resultados obtenidos se abordarán a continuación.

Inteligencia Artificial Generativa: los casos de Midjourney y Bluewillow

Los cuestionamientos sobre las herramientas de generación automática de imágenes son diversos, desde la violación de los derechos de autor (ya que hay una especie de remuneración a las herramientas) hasta el uso indebido de las imágenes producidas basadas en obras y estilos de otros artistas, incluso aún vivos. En este trabajo, sin embargo, nos centramos en los posibles sesgos que pueden existir en las imágenes.

Estas herramientas surgieron en 2021, inicialmente con DALL-E y Midjourney. En 2022 y 2023, surgieron varias herramientas, incluida una versión actualizada de DALL·E (llamada DALL·E 2), Stable Diffusion, Adobe Firefly y BlueWillow.

Michael Senkow (2022, s.f.) presenta una pregunta interesante al exponer algunas pruebas realizadas en la herramienta Midjourney con palabras sin asignación de género, inicialmente, como [humano], [blanco], [negro], [buena persona] y [mala persona], etc., para luego volver a hacer la solicitud reforzando el género y la raza. Se detectaron posibles sesgos en la plataforma, destacando que, aunque no son muy evidentes, hay una representación significativa de mujeres, sin embargo, claramente hay una "ausencia de melanina" y hay estereotipos cuando las instrucciones tienen marcas de raza.

Con el objetivo de verificar cómo responden las IA a ciertas instrucciones mediante comandos con palabras seleccionadas para esta investigación, buscamos servicios de IA de generación de imágenes. Debido a los costos exigidos por algunas de estas herramientas, optamos por utilizar BlueWillow, que es gratuito, aunque presenta restricciones en las generaciones diarias. Para esta prueba, todas las instrucciones se realizaron en secuencia, con el comando {/imagine [instrucción]}, usando en la [instrucción] la palabra en inglés, inicialmente [un hombre] y [una mujer], pero sin marcador de raza; posteriormente, de color (y no raza) y ocupación, siempre sin género.

En la segunda etapa, las instrucciones fueron [blanco], [negro], [un hombre], [una mujer], [un bombero], [una enfermera], [un médico], [un recluso]. Debido a la limitación del alcance de la producción científica del evento, decidimos limitar la cantidad de instrucciones en la plataforma. Para cada instrucción, se generaron automáticamente cuatro imágenes; se enviaron, para cada instrucción, ocho comandos, lo que totaliza, para cada instrucción, 32 imágenes; considerando todas las instrucciones, 256 imágenes. Entendemos que es una muestra relativamente baja para un estudio más detallado. Sin embargo, debido a las limitaciones de espacio y considerando que

el objetivo de este trabajo es simplemente hacer más evidente esta discusión, creemos que este límite es razonable. Por lo tanto, indicamos que este trabajo continuará con la generación de nuevas imágenes y un estudio estadístico más profundo.

En la Figura 1, se puede ver cómo la instrucción [blanco] devuelve prioritariamente imágenes de paisajes (18/32), seguido de personas (14/32). Entre las personas, se generan imágenes prioritariamente de mujeres blancas (11/14), seguido de mujeres negras (2/14) y hombres negros (1/14).

Figura 1 – Resultado de Bluewillow para el comando [*white*]

Fuente: La autora, 2023.

En la Figura 2, se puede ver cómo la instrucción [negro], por otro lado, devuelve prioritariamente imágenes de personas (24/32), seguido entonces de paisajes (8/32). De las personas, las imágenes son prioritariamente de hombres negros (9/24), algunas mujeres negras (8/24), además de personas blancas vestidas de negro (7/24).

Figura 2 – Resultado de Bluewillow para el comando [negro]

Fuente: La autora, 2023.

En la Figura 3, se presentan los resultados obtenidos para la instrucción [un hombre], que también devuelve el 100% de imágenes de personas, la mayoría de las cuales son hombres blancos (22/32), seguido de hombres negros (10/32). No se generaron imágenes de otras etnias, como asiáticos, indígenas y/o latinos.

Figura 3 – Resultado de Bluewillow para el comando [un hombre]

Fuente: La autora, 2023.

En la Figura 4, los resultados obtenidos para la instrucción [una mujer] también devuelven el 100% de imágenes de personas, la mayoría de las cuales son mujeres blancas (20/32), seguido de mujeres no blancas (11/32) y una imagen no identificable (1/32).

Figura 4 – Resultado de Bluewillow para el comando [una mujer]

Fuente: La autora, 2023

En la Figura 5, se muestran los resultados obtenidos para la instrucción [un bombero], donde la atribución para el género masculino ocurre el 100% de las veces, y la mayoría son hombres blancos (12/32).

Figura 5 – Resultado de Bluewillow para el comando [un bombero]

Fuente: La autora, 2023.

En la Figura 6, se presentan los resultados obtenidos para la instrucción [una enfermera], donde la atribución para el género femenino ocurre el 100% de las veces, siendo en su mayoría de etnia blanca (22/32).

Figura 6 – Resultado de Bluewillow para el comando [una enfermera]

Fuente: La autora, 2023.

En la Figura 7, se presentan los resultados obtenidos para la instrucción [un médico], donde la mayoría de las imágenes son de personas, siendo la atribución para el género masculino en la mayoría de los casos (28/31), seguido de mujeres (2/31) y una persona sin género definido (1/31).

Figura 7 – Resultado de Bluewillow para el comando [un médico]

Fuente: La autora, 2023.

En la Figura 8, se muestran los resultados obtenidos para la instrucción [un recluso], donde la atribución para el género masculino ocurre en su totalidad, y las imágenes son principalmente de personas negras (24/31).

Figura 8 – Resultado de Bluewillow para el comando [un recluso]

Fuente: La autora, 2023.

La generación de imágenes presentada en esta investigación refuerza consideraciones ya expuestas en otros trabajos que evidencian el patrón de las identidades representadas en las redes o que "son menos susceptibles a la marginalización, pornificación y comoditización" (NOBLE, 2019, p. 112). Entre los estereotipos que destacan, las mujeres, en cuanto a profesiones, están subordinadas o aún relacionadas con profesiones tradicionalmente reconocidas como femeninas: maestra, enfermera, trabajadora social y bibliotecaria (FERREIRA, 2003, p. 193).

Aquí, problematizamos que la neutralidad de la máquina como objeto irracional es reconocible, considerando que los agentes artificiales no son humanos. Sin embargo, en la medida en que, para alcanzar el nivel de racionalidad humana, requiere entrenamiento y programación por parte de humanos, sus "algoritmos tienden a ser vulnerables a características de sus datos de entrenamiento" (OSABA; WELSER, 2017, p. 7). El comportamiento de las IA con la generación de información/imagenes está determinado por especificaciones humanas de entrenamiento. Estos problemas no se limitan solo al reconocimiento o generación de imágenes: están presentes en anuncios, recomendaciones de contenido, visión computacional, motores de búsqueda, entre otros (SILVA, 2020b).

A través del experimento realizado, en todas las generaciones de imágenes, se observa que las personas negras están en menor número, especialmente las mujeres, lo que refuerza que la opresión algorítmica no es solo un error en el sistema: "El término algoritmo de mal comportamiento es solo una metáfora para referirse a agentes artificiales cuyos resultados llevan a consecuencias incorrectas, injustas o peligrosas" (OSABA; WELSER, 2017, p. 7), ilustrando cómo "los algoritmos están proporcionando información perniciosa sobre las personas, creando y normalizando el aislamiento estructural y sistemático, o practicando la demarcación digital, todas prácticas que refuerzan las relaciones sociales y económicas opresivas" (NOBLE, 2019, p. 32).

Consideraciones finales

El estudio teórico y los breves tests realizados en las herramientas muestran que el sesgo realmente existe, siendo reconocido incluso por algunas de las herramientas, que toman medidas para reducirlo (MOTA; BANDEIRA; MARTINS, 2021). Sin embargo, se necesitan más estudios sobre el origen de este sesgo: ¿proviene de las personas que las programan, como

sugieren Mota, Bandeira y Martins (2021), de las bases de entrenamiento, como sugieren Vinay Uday Prabhu y Abeba Birhane (2020), de la práctica científica, como plantea Caroline Criado Perez (2019), o de otra fuente?

Nuevas investigaciones pueden proporcionar mayor detalle para estos interrogantes, así como realizar una comparación de la evolución de las diferentes versiones e incluir otras instrucciones, centrándose en la representación de etnias y verificando si hay sesgos más evidentes de prejuicios y/o estereotipos. Se recopilaron y analizaron algunos datos relacionados. Sin embargo, debido a la restricción de espacio, no los incluimos en este trabajo. Creemos que otras perspectivas sobre estos fenómenos contribuirán positivamente al desarrollo de estas herramientas y al uso de tecnologías en la generación de información, pertinentes para la Ciencia de la Información, desde una perspectiva de inclusión, equidad y justicia informacional.

Por lo tanto, es necesario que los marcos legales se apliquen efectivamente, con sanciones que permitan una revisión de las cajas negras de los algoritmos, ampliando la comprensión de que los servicios prestados por las IA no pueden considerarse neutrales, ya que al aprender de datos de entrenamiento, las IA incorporan sesgos presentes en estos conjuntos de datos, proporcionando resultados que reproducen estereotipos, prejuicios y/o desigualdades. Corroborando con Bezerra y Costa (2022, p. 8), cuando afirman que "las estructuras que las componen deben ser cuestionadas, especialmente en sistemas democráticos", defendemos la necesidad de orientaciones y la creación de directrices y políticas que promuevan la equidad y la justicia social e informacional en estos entornos tecnológicos.

Referencias

Bezerra, A. C., & Costa, C. M. da. (2022). Pele negra, algoritmos brancos: informação e racismo nas redes sociotécnicas. Liinc em Revista, 18(2), e6043. Disponible en: https://revista.ibict.br/liinc/article/view/6043. [Consulta 01/03/2024].

Birhane, A., Prahbu, V. U., & Kahembwe, E. (2021). Multimodal datasets: misogyny, pornography, and malignant stereotypes. arXiv. Disponible en: https://doi.org/10.48550/arXiv.2110.01963. [Consulta 01/03/2024].

Buolamwini, J., & Gebru, T. (2018). Gender Shades: Intersectional Accuracy Disparities in Commercial Gender Classification. Proceedings of Machine Learning Research, 81, 77-91. Disponible en: http://proceedings.mlr.press/v81/buolamwini18a/buolamwini18a.pdf. [Consulta 01/03/2024].

Ferreira, M. M. (2003). O profissional da informação no mundo do trabalho e as relações de gênero. Transinformação, 15(2), 189-201. Disponible en: https://www.scielo.br/j/tinf/a/b8fgrXCGZw83LwtjrL3LbcG/abstract/?lang=pt. [Consulta 01/03/2024].

Gender Shades. (2018). How well do IBM, Microsoft, and Face++ AI services guess the gender of a face? Disponible en: http://gendershades.org/overview.html. [Consulta 01/03/2024].

Mota, D. A. R., Bandeira, J. A. R., & Martins, G. K. (2021). Algoritmos excludentes: o preconceito no recorte de imagens do twitter. In Encontro Nacional de Pesquisa em Ciência da Informação, XXI, Anais eletrônicos. Disponible en: https://enancib.ancib.org/index.php/enancib/xxienancib/paper/view/621. [Consulta 01/03/2024].

Noble, S. U. (2018). Algorithms of Oppression. Nova Iorque: NYU Press.

Perez, C. C. (2019). Invisible Women: Exposing Data Bias in a World Designed for Men. Nova Iorque: Abrams Press.

Prahbu, V. U., & Birhane, A. (2020). Large datasets: a pyrrhic win for computer vision? arXiv. Disponible en: https://doi.org/10.48550/arXiv.2006.16923. [Consulta 01/03/2024].

Prates, M. O. R., Avelar, P. H., & Lamb, L. C. (2020). Assessing gender bias in machine translation: a case study with Google Translate. Neural Comput & Applic, 32, 6363–6381. Disponible en: https://doi.org/10.1007/s00521-019-04144-6. [Consulta 01/03/2024].

Senkow, M. (2022). Midjourney is incredible. But you can see there are definite existing biases in its dataset. Medium. Disponible en: https://uxdesign.cc/midjourney-is-incredible-but-you-can-see-there-are-definite-existing-biases-in-its-dataset-4b1131fb0533. [Consulta 01/03/2024].

Silva, T. (2020a). Google acha que ferramenta em mão negra é uma arma. Disponible en: https://tarciziosilva.com.br/blog/google-acha-que-ferramenta-em-mao-negra-e-uma-arma/. [Consulta 01/03/2024].

Silva, T. (2020b). Racismo Algorítmico em Plataformas Digitais: microagressões e discriminação em código. In T. Silva (Org.), Comunidades, Algoritmos e Ativismo Digitais: olhares afrodiaspóricos. São Paulo: LiteraRUA.

Turing, A. M. (1950). Computing Machinery and Intelligence. Mind, LIX(236), 433-460. Disponible en: https://redirect.cs.umbc.edu/courses/471/papers/turing.pdf. [Consulta 01/03/2024].