XVI Congreso Internacional de Investigación y Práctica Profesional en Psicología. XXXI Jornadas de Investigación. XX Encuentro de Investigadores en Psicología del MERCOSUR. VI Encuentro de Investigación de Terapia Ocupacional. VI Encuentro de Musicoterapia. Facultad de Psicología - Universidad de Buenos Aires, Buenos Aires, 2024.

Sesgo de género en asistentes virtuales: una aproximación desde la psicología social.

Travnik, Cecilia, Mandelbaum, Matias y Marchiano, Federico Agustín.

SESGO DE GÉNERO EN ASISTENTES VIRTUALES: UNA APROXIMACIÓN DESDE LA PSICOLOGÍA SOCIAL

Travnik, Cecilia; Mandelbaum, Matias; Marchiano, Federico Agustín
Universidad de Buenos Aires. Facultad de Psicología. Buenos Aires, Argentina.

RESUMEN

El presente estudio explora el sesgo de género en asistentes virtuales y su impacto, revelando que una mayoría de los participantes percibe estas tecnologías como femeninas, reflejando una tendencia a posicionar a la mujer como asistente. Un 87.2% percibe un género femenino en los asistentes, mientras que el 74.4% cree que el género de la voz no afecta la eficiencia del asistente. Sin embargo, un 76.9% de los encuestados no ha escuchado hablar del sesgo de género algorítmico, destacando la necesidad de mayor sensibilización y educación sobre el tema. Se muestra que los prejuicios implícitos y explícitos de los programadores se integran en los algoritmos, perpetuando estereotipos de género y reflejando roles tradicionales asignados a las mujeres. Estos sesgos, presentes en la configuración y el desarrollo de las tecnologías, pueden llevar a la discriminación algorítmica y amplificar desigualdades sociales. Aunque el estudio proporciona valiosa información, tiene limitaciones debido al muestreo no probabilístico por conveniencia y la administración en línea del cuestionario. La integración de perspectivas de la psicología social y el análisis de sesgos cognitivos es crucial para desarrollar tecnologías más equitativas y justas, reconociendo y mitigando los prejuicios inherentes en los sistemas informáticos.

Palabras clave

Sesgo de género - Prejuicio - Asistentes virtuales - Sesgo algorítmico

ABSTRACT

GENDER BIAS IN VIRTUAL ASSISTANTS: AN APPROACH FROM SOCIAL PSYCHOLOGY

The present study explores gender bias in virtual assistants and its impact, revealing that the majority of participants perceive these technologies as female, reflecting a tendency to position women as assistants. 87.2% perceive a female gender in the assistants, while 74.4% believe that the gender of the voice does not affect the assistant's efficiency. However, 76.9% of respondents have not heard about algorithmic gender bias, highlighting the need for greater awareness and education on the subject. It is shown that the implicit and explicit biases of programmers are integrated into algorithms, perpetuating gender stereotypes and reflecting traditional roles assigned to women. These biases, present in the configuration and development of technologies, can lead to algorithmic discrimination and amplify social inequalities. Although the study provides valuable information, it has limitations due to non-probability convenience sampling and the online administration of the questionnaire. Integrating perspectives from social psychology and cognitive bias analysis is crucial to develop more equitable and fair technologies, recognizing and mitigating the inherent biases in computer systems.

Keywords

Gender bias - Virtual assistants - Algorithmic bias - Prejudice

Introducción

En la última década, los avances en inteligencia artificial (IA) han transformado numerosos sectores, desde la medicina y el transporte hasta las finanzas y la educación. Los algoritmos, como parte fundamental de la IA, son utilizados para analizar grandes volúmenes de datos y tomar decisiones de manera eficiente en diversas áreas. Por ejemplo: sistemas de recomendación en plataformas de streaming, motores de búsqueda, contratación de personal, calificación crediticia para concesión de préstamos, evaluación del riesgo para otorgar libertad condicional, entre otras. Sin embargo, existe una creciente preocupación sobre los sesgos incorporados en estos algoritmos y cómo estos reflejan y perpetúan los prejuicios sociales existentes. Este fenómeno, conocido como sesgo algorítmico, puede tener consecuencias significativas, exacerbando desigualdades y discriminaciones.

Desde la Psicología Social, es crucial entender los mecanismos subyacentes que contribuyen a la formación y perpetuación de estos sesgos. Este escrito tiene como objetivo principal explorar cómo los prejuicios implícitos y explícitos se integran en los algoritmos y reconocer las implicaciones sociales de estos sesgos. Estos sesgos pueden surgir debido a diversas razones, como datos de entrenamiento que reflejan prejuicios humanos existentes, errores en el diseño del algoritmo o la falta de diversidad en los equipos de desarrollo. Como resultado, los algoritmos pueden perpetuar y, en algunos casos, amplificar las desigualdades sociales y económicas.

Estudiar los prejuicios incorporados en los algoritmos es relevante porque a través de ellos se están tomando decisiones que afectan la vida de millones de personas y porque la falta de transparencia y la equidad en la IA son esenciales para mantener la confianza pública en estas tecnologías emergentes. De lo contrario, puede conllevar a una menor adopción de tecnologías beneficiosas.

Sesgos cognitivos, sesgos cognitivos digitales y sesgos algorítmicos

La psicología y las ciencias cognitivas hace tiempo que estudian los sesgos cognitivos, atajos mentales o heurísticos que nuestro cerebro utiliza para tomar decisiones rápidas, pero que pueden llevar a errores de juicio o razonamiento. Kahneman (2011) junto con su colega Amos Tversky, identificaron y describieron numerosos sesgos cognitivos a través de su investigación, desafiando la noción de que los seres humanos son agentes racionales que siempre toman decisiones basadas en una lógica clara y objetiva. Los autores describen dos sistemas de pensamiento: el Sistema 1, que es rápido, automático y emocional; y el Sistema 2, que es más lento, deliberativo y lógico. Los sesgos cognitivos suelen originarse en el Sistema 1, donde las decisiones rápidas y basadas en heurísticas pueden ser eficientes pero a menudo son propensas a errores sistemáticos.

Los sesgos cognitivos son efectos psicológicos que se producen por recortar y categorizar la información a la que estamos expuestos/as. Se utilizan atajos mentales para disminuir el costoso proceso cognitivo que significa conceptualizar cada información que el entorno brinda; por ello, el cerebro resume la información. Esta economía cognitiva implica que, en el proceso de resumir información, se estereotipa y agrupa según determinadas características aprendidas en el proceso de socialización (Kahneman, 2011; Holroyd et al., 2017).

Décadas anteriores Weizenbaum (1976) en "Computer Power y Human Reason", destacó cómo los sesgos podrían surgir tanto de los datos utilizados en los programas como de la manera en que estos programas son codificados. Señaló que los algoritmos, siendo conjuntos de reglas creadas por humanos, pueden reflejar los prejuicios y expectativas de sus diseñadores. Esta crítica temprana sentó las bases para comprender cómo los sesgos pueden estar incrustados en los procesos de toma de decisiones algorítmicas.

Así, el sesgo algorítmico se ha convertido en un área significativa de estudio, especialmente con el advenimiento de las tecnologías de IA. O'Neil (2016) ha popularizado aún más el concepto, enfatizando cómo estos sesgos pueden llevar a resultados injustos en diversos sectores, como la justicia penal, la salud y las finanzas?. Estudios recientes muestran que los algoritmos de aprendizaje automático pueden discriminar por raza y género. Las tecnologías de reconocimiento facial reproducen desigualdades existentes. Buolamwini y Gebru (2018) presentan un enfoque para evaluar el sesgo en los algoritmos y conjuntos de datos de análisis facial automatizado respecto a subgrupos fenotípicos.

Utilizando el sistema de clasificación de tipos de piel de Fitzpatrick, analizamos la distribución de género y tipo de piel en los conjuntos de datos IJB-A y Adience, encontrando una predominancia de sujetos de piel clara (79.6% en IJB-A y 86.2% en Adience). Introducimos un nuevo conjunto de datos equilibrado por género y tipo de piel y evaluamos tres sistemas comerciales de clasificación de género. Descubrimos que las mujeres de piel oscura son el grupo más mal clasificado (hasta un 34.7% de error), mientras que los hombres de piel clara tienen la menor tasa de error (0.8%). Las diferencias en la precisión de clasificación requieren atención urgente para desarrollar algoritmos de análisis facial justos, transparentes y responsables (Buolamwini y Gebru, 2018, p. 1).

Estas tecnologías están creciendo exponencialmente y se utilizan cada vez más para distintas áreas de la vida cotidiana. Del mismo modo se desarrollan nuevas formas de prejuicio y de discriminación, lo que se intenta poner de relieve es la relación entre la tecnología y la sociedad como una nueva forma de interacción que reaviva antiguos debates.

El estudio del sesgo algorítmico no solo se centra en la tecnología, sino que también profundiza en el entendimiento psicológico del prejuicio y su manifestación en la sociedad. La psicología social y el análisis de los sesgos cognitivos nos permiten comprender cómo las decisiones rápidas y heurísticas pueden perpetuar y amplificar prejuicios existentes en los sistemas algorítmicos. Por tanto, es crucial integrar estas perspectivas para desarrollar tecnologías más justas y equitativas, reconociendo que los algoritmos, al igual que los humanos, no están exentos de sesgos inherentes.

Sesgo de género en los sistemas informáticos

El estudio del sesgo de género en los sistemas informáticos es una cuestión crítica en la era digital, ya que estos sistemas reflejan y, a menudo, perpetúan las desigualdades. La psicología social, siguiendo a Allport (1954) es una disciplina que examina cómo los pensamientos, sentimientos y comportamientos de las personas son influenciados por la presencia real o imaginada de otros, es fundamental para comprender el origen y la perpetuación del prejuicio hacia las mujeres.

El sexismo y los estereotipos de género son fenómenos profundamente arraigados en la estructura social y cultural de muchas sociedades. El comportamiento y los roles de género están influenciados por estas normas, donde las diferencias de género reflejan la asignación de roles en grupos y la sociedad. Esta asignación de roles es frecuentemente determinada y perpetuada por el grupo social que ostenta más poder, típicamente los hombres. La persistencia de los estereotipos de género se ve reforzada por la continua asignación de roles basados en el género (Hogg y Vaughan, 2018).

Una de las fuerzas más poderosas en la transmisión y mantenimiento de estos estereotipos de género tradicionales son los medios de comunicación. Las formas evidentes de esto incluyen la representación de mujeres semidesnudas en publicidad y su rol decorativo en programas de televisión, donde a menudo están ajenas a la trama principal y presentadas solo como entretenimiento sexual o romántico (Hogg y Vaughan, 2018). La feminización de las IA es considerada una forma sútil y poderosa de perpetuar estos estereotipos.

Ramírez Autrán (2023) plantea que se ha encontrado que asistentes virtuales como Siri, Cortana y Alexa, todas con nombres de mujeres o diosas, simbolizan un proceso de feminización de las IA. Sutko (2019) señala que en la división del trabajo por género se normaliza que las mujeres desempeñan un trabajo comunicativo y simbólico. Perspectivas como esta ayudan a entender la "domesticación de la IA", que resulta de asociar la feminidad con características como docilidad, receptividad y afectividad. Investigaciones internacionales (UNESCO, 2019) demuestran, sin ambigüedad, que los sesgos de género presentes en los conjuntos de datos, algoritmos y dispositivos de capacitación de IA tienen el potencial de propagar y reforzar estereotipos de género perjudiciales.

Estos sesgos pueden manifestarse durante el desarrollo del algoritmo, el entrenamiento de los conjuntos de datos o mediante la toma de decisiones generada por la IA (Manasi et al., 2022). Esto podría estigmatizar aún más a las mujeres, relegándolas en diversos ámbitos de la vida económica, política y social, y retrasar el progreso en materia de igualdad de género.

Procedimiento

Con el objetivo de explorar en la temática de la feminización de los asistentes virtuales y su relación con el sesgo algorítmico de género, se diseñó un cuestionario compuesto por 5 preguntas de opción de respuesta múltiple, 6 ítems utilizando una escala Likert de 5 puntos (1 = Totalmente en desacuerdo, 5 = Totalmente de acuerdo) para medir las actitudes sobre el género de los asistentes virtuales y su impacto percibido y 1 pregunta de respuesta abierta acerca de la definición de sesgo algorítmico.

La muestra estuvo compuesta por 286 individuos, se realizó un muestreo no probabilístico por conveniencia. El cuestionario fue administrado en línea a través de googleforms, lo que permitió la participación remota y la recolección eficiente de datos. Las personas recibieron un enlace al cuestionario junto con una breve explicación del propósito del estudio y una garantía de confidencialidad.

Resultados

Los participantes, en base a la edad, se distribuyeron de la siguiente manera: el 12.8% de los participantes se encontraron en el rango etario de los 18 a 25 años, el 33.33% de los participantes entre los 25 y 35 años, el 17,9% entre los 35 y 45 años, el 20,5% entre los 45 y 55 años, el 10,3% entre los 55 años y los 65 años y el 5.1% de más de 65 años. A su vez, en cuanto a la distribución por género de los/las individuos, el 53,8% se percibió del género femenino y el 46,2% del género masculino, no habiéndose registrado otras identidades de género o negarse a responder. En base a la distribución sobre nivel de estudios alcanzado, una mayoría representada por el 53% se encuentra con estudios universitarios completos, el 20,5% terciario y el 23,1% con secundario completo. Respecto al lugar de residencia, el 71,8% se encuentra en la Ciudad Autónoma de Buenos Aires, el 25,6% en Provincia de Buenos Aires y 2,6% en otro país. En cuanto al nivel de ocupación, la mayoría, representada por el 4.6%, refirió encontrarse empleado en relación de dependencia, el 30,8% como estudiante y trabajador y un 10,3% autónomo. En la presente distribución por ocupación, no se registraron respuestas referidas a encontrarse sin trabajo ni estudio.

En cuanto al uso de asistentes virtuales, los resultados indicaron que el 35.9% de los participantes usa estos asistentes de manera ocasional, un 30.8% los utiliza casi todos los días, y el 7.7% reportó usarlos todos los días. Una amplia mayoría del 87.2% de los encuestados percibe que los asistentes virtuales tienen un género femenino. Sobre la cuestión acerca si el género de la voz afecta la eficiencia del asistente virtual, el 74.4% de los participantes creen que el género de la voz no afecta la eficiencia, un 17.9% considera que sí afecta, y el 7.7% respondió que tal vez podría tener un impacto.

En relación al conocimiento sobre el sesgo de género algorítmico, un 76.9% de los encuestados no ha escuchado hablar del tema, mientras que el 23.1% sí tiene conocimiento sobre el sesgo de género algorítmico.

A la pregunta abierta "Si escuchaste hablar de "sesgo de género algorítmico" ¿Qué entendés por eso?" La tendencia es las respuestas recolectadas es relacionarlo con la utilización de voces femeninas en los asistentes virtuales, lo cual se percibe como una forma de segregación basada en el género. Muchos participantes entienden que este sesgo de género proviene de los programadores, quienes configuran las bases de datos y desarrollan los algoritmos, transfiriendo así sus propios sesgos a los sistemas. Las respuestas indican que las aplicaciones y la inteligencia artificial reproducen estos roles de género, lo que lleva a la reproducción de estereotipos y sesgos presentes en la sociedad. Algunos participantes expresaron que los errores en los sistemas informáticos, incluyendo el sexismo algorítmico, reflejan los prejuicios de sus creadores.

Los participantes señalaron que la preferencia por usar voces femeninas o aspectos humanos femeninos para interactuar con los sistemas refleja una tendencia a posicionar a la mujer como asistente, orientando estos roles a estereotipos de género femenino. Esta configuración no es neutral y se ve como una forma de discriminación algorítmica basada en el género, perpetuando los roles tradicionales que la sociedad asigna a las mujeres.

Finalmente, al abordar la neutralidad de los sistemas informáticos en relación al género, la mayoría de los participantes, el 80% cree que estos sistemas no son neutrales respecto a la problemática de género.

Discusión

Los resultados indican que una gran mayoría de los participantes percibe a los asistentes virtuales como femeninos, reflejando una feminización de estas tecnologías. Este hallazgo es consistente con estudios previos (UNESCO, 2019; Manasi et al., 2022) que demuestran cómo los sesgos de género en los

conjuntos de datos y algoritmos pueden reforzar estereotipos perjudiciales. La falta de conocimiento general sobre el sesgo de género algorítmico sugiere una necesidad de sensibilización y/o de educación en esta área.

La psicología social nos ofrece un marco para entender cómo estos prejuicios se perpetúan y amplifican en los sistemas algorítmicos. Los estereotipos de género, mantenidos por normas culturales y sociales, se ven reflejados y perpetuados en la tecnología que se utiliza diariamente. La domesticación de la IA, donde las voces femeninas se asocian con características de docilidad y receptividad, no solo refuerza los roles de género tradicionales, sino que también puede colaborar en la estigmatización hacia las mujeres en diversos ámbitos de la vida.

Este estudio, aunque proporciona una visión sobre la percepción del sesgo de género en los asistentes virtuales, presenta ciertas limitaciones. La muestra fue obtenida mediante un muestreo no probabilístico por conveniencia, lo que podría no representar adecuadamente a la población general. Asimismo, la administración del cuestionario en línea puede haber excluido a individuos con acceso limitado a internet. Estas limitaciones sugieren la necesidad de estudios futuros con muestras más diversas y representativas para obtener una comprensión más completa del sesgo de género en los sistemas informáticos.

Para desarrollar tecnologías más justas y equitativas, es crucial integrar perspectivas de la psicología social y el análisis de los sesgos cognitivos en el diseño y entrenamiento de algoritmos. Reconociendo que los algoritmos, al igual que los humanos, no están exentos de sesgos inherentes, se puede trabajar hacia la mitigación de estos prejuicios. La educación y la sensibilización sobre el sesgo algorítmico de género son pasos esenciales hacia la igualdad de género en la era digital.

REFERENCIAS BIBLIOGRÁFICAS

Allport, G. W. (1954). *La naturaleza del prejuicio*. Addison-Wesley.

Buolamwini, J., & Gebru, T. (2018). Gender shades: Intersectional accuracy disparities in commercial gender classification. *Proceedings of Machine Learning Research* (Vol. 81, pp. 1-15). http://proceedings.mlr.press/v81/buolamwini18a.html

Hogg, M. A., & Vaughan, G. M. (2018). Prejudice and discrimination. En *Social Psychology* (pp. 366-411). United Kingdom: Pearson Education.

Holroyd, J., Scaife, R., & Stafford, T. (2017). What is implicit bias?. *Philosophy Compass,* 12(10), 18-23.

Kahneman, D. (2011). *Pensar rápido, pensar despacio*. Debate.

Manasi, A., Panchanadeswaran, S., Sours, E. y Ju, S. (2022, 8 de noviembre). Mirroring the bias: gender and Artificial Intelligence. *Gender, Technology and Development*, 3(26), 295-305. https://doi.org/10.1080/09718524.2022.2128254

O'Neil, C. (2016). *Weapons of Math Destruction. How Big Data Increases Inequality and Threatens Democracy*. Crown Publishers.

Ramirez Autrán, R. (2023). Sesgos y discriminaciones sociales de los algoritmos en Inteligencia Artificial: una revisión documental. *Entretextos, 15(39*), 1-17. https://doi.org/10.59057/iberoleon.200753 16.202339664

Sutko, D. (2019). Theorizing femininity in Artificial Intelligence: a framework for undoing technology's gender troubles. *Cultural Studies,* 34(4), 567-592. https://doi.org/10.1080/09502386.2019.1671469

Weizenbaum, J. (1976). *Computer power and human reason: From judgment to calculation*. W. H. Freeman.