10.52712/issn.1850-0013-555

Injusticia epistémica y reproducción de sesgos de género en la inteligencia artificial *

Injustiça epistêmica e reprodução do preconceito de gênero na inteligência artificial

Epistemic Injustice and Reproduction of Gender Bias in Artificial Intelligence

Inmaculada Perdomo Reyes **

La inteligencia artificial (IA) generativa reifica y pone en circulación las brechas y los sesgos de género ya existentes, pero otorgándoles un barniz de objetividad y neutralidad a pesar de la opacidad de los procesos y su capacidad para reproducir e incrementar las situaciones de desigualdad y exclusión. La situación es de clara injusticia algorítmica y epistémica y nos enfrenta con retos de gran envergadura en nuestras modernas democracias. Con ejemplos de casos concretos y con la revisión crítica de importantes textos que ofrecen claves interpretativas para comprender el impacto del rápido desarrollo e implantación de estas herramientas, trazaremos algunas líneas maestras que requerirán de estudios con mayor profundidad pero que pretenden recoger, desde la perspectiva de los estudios de ciencia, tecnología y género, nuevos desafíos para el desarrollo de la disciplina y avistar las posibilidades de una IA feminista.

Palabras clave: tecnologías; sesgos de género; feminismo; inteligencia artificial

* Recepción del artículo: 30/01/2024. Entrega del dictamen: 27/03/2024. Recepción del artículo final: 08/04/2024.

** Doctora en filosofía de la ciencia. Profesora titular del Área de Lógica y Filosofía de la Ciencia, Facultad de Humanidades, Sección de Filosofía de la Universidad de La Laguna (ULL), España. Investigadora en el Instituto Universitario de Estudios de las Mujeres de la ULL. Correo electrónico: mperdomo@ull.edu.es. ORCID: https://orcid.org/0000-0003-4838-7278.

10.52712/issn.1850-0013-555

As IA generativas reificam e fazem circular as lacunas e preconceitos de género existentes, mas conferem-lhes um verniz de objetividade e neutralidade, apesar da opacidade dos processos e da sua capacidade de reproduzir e aumentar situações de desigualdade e exclusão. A situação é de clara injustiça algorítmica e epistémica e confronta-nos com grandes desafios nas nossas democracias modernas. Com exemplos de casos específicos e com a revisão crítica de textos importantes que oferecem chaves interpretativas para compreender o impacto do rápido desenvolvimento e implementação dessas ferramentas, traçaremos algumas diretrizes que exigirão estudos mais aprofundados, mas que visam coletar, na perspetiva da ciência, tecnologia e estudos de gênero, novos desafios para o desenvolvimento da disciplina e ver as possibilidades de uma IA feminista.

Palavras-chave: tecnologias; preconceitos de gênero; feminismo; inteligência artificial

Generative AIs reify and circulate existing gender gaps and biases, but give them a veneer of objectivity and neutrality despite the opacity of their processes and ability to reproduce and increase situations of inequality and exclusion. The situation is one of clear algorithmic and epistemic injustice that confronts us with major challenges in our modern democracies. With examples of specific cases and with the critical review of important texts that offer interpretative keys to understand the impact of the rapid development and implementation of these tools, we will outline some guidelines that will require more in-depth studies, but that aim to collect, from the perspective of science, technology and gender studies, new challenges for the development of the discipline and to envision the possibilities of a feminist AI.

Keywords: *technologies; gender biases; feminism; artificial intelligence*

10.52712/issn.1850-0013-555

Introducción

Las inteligencias artificiales (IA) con aprendizaje autónomo basado en redes neuronales son capaces de realizar correlaciones e inferencias a partir de los millones de datos que utilizan (bases de datos, de imágenes, resultados de las interacciones humanas con *smartphones*, etc.) que reflejan, como si de un espejo se tratara (Coeckelbergh, 2021), el cúmulo de rasgos y características humanas. La palabra latina *datum*, que viene de *dare* (dar), significa literalmente "lo dado". Como sugiere el filósofo Byung Chul Han, "el conocimiento en el régimen de la información se esfuerza por lograr un conocimiento total a través de la operación algorítmica, sustituyendo lo narrativo por lo numérico. El dataísmo quiere calcular todo lo que es y será" (2022, p. 21).

Pero reflejan también en esa operación los sesgos, prejuicios y estereotipos que siguen estructurando nuestras sociedades. Los ejemplos de sesgo algorítmico adquirido a través de los datos son múltiples y muy variados y tienen que ver con el carácter mismo de las bases de datos que, en realidad, no representan al conjunto de la población; por el contrario, son el reflejo de las concepciones dominantes y las experiencias de aquellos que han ocupado los centros de privilegio epistémico en nuestra cultura (Criado, 2019). Las IA generativas reifican y ponen en circulación las brechas y sesgos ya existentes, pero otorgándoles un barniz de objetividad y neutralidad a pesar de la opacidad de la mayoría de estos procesos. Pueden ser verdaderos mecanismos automatizados de reproducción y generación de desigualdades y exclusión (Eubanks, 2018). Las tecnologías son, al tiempo, un reflejo y una cristalización de los procesos sociales, y aún hay pocos estudios que se centren en cómo las relaciones de poder y género acaban integradas en la tecnociencia, desde el diseño de programas y herramientas al trazado de objetivos pragmáticos (Wajcman y Young, 2023, p. 48), pero es ampliamente reconocido que tanto la fuerza de trabajo como las culturas dominantes en la tecnociencia son claramente representativas de los grupos de personas que forman su núcleo y que las mujeres, por ejemplo, no representan más de un 18% en los puestos especializados de la industria tecnológica (Young, Wajcman y Sprejer, 2023, p.16). La crítica feminista de la ciencia y de la tecnología tiene ya una larga tradición de estudios que muestran cómo los valores e intereses están presentes en diferente grado en todas las fases del proceso de investigación e innovación. Hoy, desde perspectivas plurales, los estudios feministas de la ciencia y la tecnología asumen la tarea del análisis empírico y contextual de los imaginarios sociotécnicos y las relaciones de conformidad mutua tecnología-sociedad (Jasanoff, 2004, 2016; Wajcman, 2010; Wacjman y Young, 2023) en una nueva época de definición de las estrategias feministas ante el desarrollo de las tecnologías más disruptivas que impactan directamente en nuestras vidas en sociedad y en los procesos mismos de tomas de decisión en las sociedades democráticas.

Así, delegar en herramientas algorítmicas determinados procesos de tomas de decisión en las sociedades como: la elegibilidad de personas candidatas para un trabajo, para acceder a seguros de vida, créditos o hipotecas, y la resolución de solicitudes administrativas, etc., ofrecen a los gestores públicos y privados la distancia ética necesaria para tomar decisiones que están incrementando la vulnerabilidad, desigualdad y exclusión de las personas menos favorecidas. Pertenecer a los perfiles estadísticamente relevantes identificados por los modelos, algoritmos y sistemas de

10.52712/issn.1850-0013-555

IA generativa puede suponer la diferencia entre ser elegible o no para ser beneficiario de todos los recursos que una sociedad pone en circulación para facilitar la vida de la ciudadanía. La línea que separa la clasificación basada en cálculo algorítmico (aparentemente objetiva) y la evaluación de las condiciones concretas de casos que requieren reflexión, racionalidad y comunicación humanas no es tan delgada si supone el incremento de la vulnerabilidad y exclusión social de esas personas. Multitud de ejemplos son recogidos por Virginia Eubanks en *Automating Inequality*. La situación es de clara injusticia algorítmica y epistémica y nos enfrenta con retos de gran envergadura en el ámbito del trabajo y, especialmente, en los procesos de tomas de decisión públicas y en los procesos de generación y transferencia de los conocimientos. En este artículo abordaré algunos de los retos a que nos enfrenta el avance imparable de la IA. Con ejemplos de casos concretos y con la revisión crítica de importantes textos que ofrecen claves interpretativas para comprender el impacto del rápido desarrollo e implantación de estas herramientas basadas en algoritmos cada vez más complejos y de las IA generativas, trazaremos algunas líneas maestras que requerirán de estudios con mayor profundidad pero que pretenden recoger, desde la perspectiva de los estudios de ciencia, tecnología y género, nuevos desafíos para el desarrollo de la misma disciplina y avistar las posibilidades de una IA feminista.

1. Los riesgos de la elegibilidad automatizada

Las personas más afectadas por las innovaciones basadas en IA no participan en el diseño, desarrollo y despliegue de esas tecnologías y sus intereses no se tienen necesariamente en cuenta. Todo lo contrario. Esta distancia entre las personas más afectadas por los algoritmos y quienes diseñan y desarrollan los programas tiene consecuencias. Veamos algunos ejemplos.

1.1. El caso Amazon

A su nuevo motor de reclutamiento no le gustaban las mujeres. La noticia saltó a la prensa y a los juzgados, aunque Amazon argumentó que la herramienta no había sido utilizada más que en fase de prueba. La automatización ha sido clave para el dominio del comercio electrónico de Amazon, ya sea dentro de los almacenes o impulsando las decisiones de precios. La herramienta de contratación experimental de la compañía utilizó inteligencia artificial para adjudicar a las personas candidatas unas puntuaciones de una a cinco estrellas, de la misma forma que los compradores califican los productos en Amazon. Pero en 2015 la compañía se dio cuenta de que su nuevo sistema no estaba calificando a los candidatos para trabajos de desarrollador de *software* y otros puestos técnicos de una manera neutral en cuanto al género. Esto se debía a que el modelo informático de Amazon fue entrenado para examinar a los solicitantes mediante la observación de patrones en los currículos presentados a la empresa durante un período de diez años. La mayoría provenía de hombres, un reflejo del dominio masculino en la industria tecnológica. En efecto, el sistema de Amazon aprendió, ya que eran los estadísticamente relevantes, que los candidatos masculinos eran preferibles y, como consecuencia, en su proceso de selección penalizaba los currículos que incluían las palabras “femenino” o “mujeres” como en "capitana del club de ajedrez femenino". En cambio, la tecnología favoreció a los candidatos que

se describían a sí mismos usando verbos que se encuentran más comúnmente en los currículos de los ingenieros masculinos, como “ejecutar” y “capturar” (Dastin, 2022, p. 296). El algoritmo aprendió a infravalorar sistemáticamente los CV de las mujeres para trabajos técnicos como el de desarrollador de *software* y, aunque Amazon está a la vanguardia de la tecnología de IA, la empresa no pudo encontrar una manera de hacer que su algoritmo fuera neutral en cuanto al género. Los algoritmos de IA están entrenados para observar patrones en grandes conjuntos de datos para ayudar a predecir resultados. En el caso de Amazon, su algoritmo utilizó todos los CV enviados a la empresa durante un período de diez años para aprender a detectar a los mejores candidatos. Dada la baja proporción de mujeres que trabajaban en la empresa, como en la mayoría de las empresas tecnológicas, el algoritmo detectó rápidamente la prevalencia masculina y registró que era un factor de éxito.

Como expone Caroline Criado en su premiado libro *Invisible Women* (2019), las vidas de los hombres representan en nuestra cultura a las de los seres humanos en general; cuando se trata de las vidas de la otra mitad de la humanidad, a menudo no hay nada más que silencio. Las experiencias de las mujeres y las de otros colectivos que sufren exclusión y desigualdades en nuestra sociedad están excluidas de los datos, y la brecha de datos de género, en clave interseccional, no implica solo silencio, sino que estos silencios, estas lagunas, tienen consecuencias: el “*Big Data* se transforma en Grandes Verdades a través de los Grandes Algoritmos, utilizando Grandes Ordenadores y esto afecta a nuestra vida cotidiana, desde el transporte público hasta la política, pasando por el lugar de trabajo y la consulta médica” (Criado, 2019, p. 6). Las IA, en definitiva, se trate de aprendizaje supervisado o autónomo, se entrenan con conjuntos de datos que están plagados de brechas, sesgos y estereotipos, y los algoritmos reflejan y amplifican esa situación a la perfección en sus procesos de tomas de decisión. Unos procesos que, sin embargo, tienden a ser valorados como más objetivos al eliminar el factor de la subjetividad humana.

> “*An underlying problem is that AI systems are presented as objective and neutral in decision making rather than as inscribed with masculine, and other, preferences and values. Machines trained using datasets generated in an unequal society tend to magnify existing inequities, turning human prejudices into seemingly objective facts*” (Wajcman y Young, 2023, p. 58).

1.2. El caso Bosco

Un algoritmo nada complejo de árbol de decisión nos ofrece otras claves de vulnerabilidad de la ciudadanía cuando esta ni siquiera es consciente de que son utilizados para procesos de tomas decisión que analizan sus solicitudes de acceso a ayudas sociales para enfrentar la carestía de la vida. Civio,[1] una fundación de

1. Civio se define como la primera organización en España especializada en vigilar a los poderes públicos. Su misión es “lograr transparencia de verdad en los asuntos públicos y dotar a toda la sociedad de la información que necesita para exigir transparencia, responsabilidad y eficacia a las administraciones. Para defender mejor sus derechos e intereses. Y para actuar en consecuencia”. Más información en: https://civio.es/nosotros/.

periodismo comprometido dio la alerta hace unos años en España. El gobierno español aprobó en 2009 la opción de descuentos en la factura de electricidad. Los posibles beneficiarios debían solicitar el bono social a través de su compañía eléctrica proveedora; esta solicitaba su aprobación o denegación al gobierno y se aplicaban los descuentos correspondientes en la factura si los solicitantes cumplían los requisitos para obtenerlo. En 2017 son publicados unos nuevos criterios de acceso al bono social eléctrico en el Boletín Oficial del Estado (BOE), que implicaba una redefinición de los perfiles de los beneficiarios: niveles de rentas bajas, pensionistas, pero solo aquellos con pensión mínima, y las familias numerosas que ahora podían beneficiarse todas independientemente del nivel de renta. Además, tal como comprobó Civio, las solicitudes a la compañía y enviadas a la administración correspondiente eran analizadas por un algoritmo llamado Bosco, algo totalmente desconocido por las personas solicitantes y por la ciudadanía en general. Civio crea una aplicación informática en 2018 que cualquier persona podía utilizar para comprobar si podía optar a ser beneficiaria, y para ayudar a utilizar el formulario que debía rellenarse y enviar a la compañía eléctrica. Un número importante de usuarios llama la atención sobre el hecho de que, aunque la aplicación de ayuda de Civio les consideraba beneficiarios, su solicitud a la compañía, que solo hacía de intermediadora, era denegada. La pregunta era obligada: ¿qué pasaba con Bosco? ¿Era un algoritmo defectuoso que dejaba fuera a la mayoría de los potenciales beneficiarios? Civio solicita transparencia a la administración y reclama aplicar el derecho a saber de la ciudadanía sobre cómo se toman las decisiones que les afectan.

Los siguientes episodios de este caso ilustran claramente la indefensión de la ciudadanía cuando se aplican algoritmos (y este es uno nada complejo) para decidir la elegibilidad de las personas receptoras de ayudas, becas, solicitantes de empleo y un largo etcétera, que seguirá incrementándose exponencialmente, de procedimientos de las administraciones públicas y ámbitos privados en los que la IA estará presente. Civio quiere revisar Bosco y solicita las especificaciones técnicas, las pruebas de funcionamiento y el código fuente. Las dos primeras cuestiones son facilitadas, pero no la última. El argumento: los derechos de propiedad intelectual y la protección de datos. Sin embargo, la revisión, obtenidas las especificaciones técnicas del algoritmo, ya permite a Civio identificar graves errores. Por ejemplo, a las viudas, si marcaban la casilla de pensionistas, se les denegaba la ayuda, al no tener una pensión mínima; y aquellas familias numerosas que no aportaban la documentación sobre sus ingresos (siendo beneficiarias todas según los criterios) también veían su solicitud denegada. Este análisis permite corregir estos y otros aspectos defectuosos en el formulario y árbol de decisión del algoritmo que, a propuesta de Civio, son realizados por la administración del Estado, pero, a pesar de que han continuado solicitando a través de los tribunales el acceso al código fuente, esto es denegado en varias instancias y procedimientos. Pasados ya unos años de litigio, las respuestas siguen siendo las mismas: seguridad nacional, protección intelectual y protección de datos. Lo importante del caso, más allá de valorar la relevancia y la necesidad de contar con profesionales como los de Civio, que revisan los procedimientos de las administraciones públicas para informar adecuadamente a la ciudadanía, y para litigar cuando es necesario por el derecho a saber y exigir la transparencia de los procesos de tomas de decisión en los asuntos que afectan a nuestras vidas, es que hace emerger una serie de preguntas: si se nos regula mediante algoritmos o códigos fuente secretos y protegidos por la

propiedad intelectual desde las administraciones públicas, ¿seguimos habitando en un Estado social, democrático y de derecho? O, por el contrario, ¿observamos cómo transitamos progresivamente y de forma muy rápida desde el lenguaje natural de las leyes al lenguaje formal, no público, de los algoritmos cada vez más sofisticados y opacos? Además, ¿qué capacidad tendremos en el futuro de saber cómo se realizan los procesos de tomas de decisión que afectan a nuestros derechos? ¿Qué capacidad tendremos de reclamar si ni siquiera tendremos interlocutores humanos que nos expliquen cómo se han tomado esas decisiones?

Habitamos ya en este nuevo régimen de los datos digitales, pero no todas las personas lo experimentan de la misma forma. Los datos y los algoritmos actúan para reforzar los estereotipos presentes en la sociedad y la marginalidad de los grupos más vulnerables, sobre todo cuando se utilizan para señalarles como sospechosos (por ejemplo, el sistema COMPAS[2]), no elegibles o sufren un escrutinio adicional, provocando un ciclo de exclusión y vulnerabilidad, un ciclo de retroalimentación de injusticias (Eubanks, 2018). Otros ejemplos de sesgo algorítmico han llevado a muchos analistas, tecnólogos y académicos e investigadoras a pedir una auditoría más sistemática de los sistemas de aprendizaje automático y un diseño más concienzudo y ético de los procesos de aprendizaje utilizados para entrenar a los sistemas de IA (O'Neill, 2016). La investigadora que documentó la discriminación sistemática en los sistemas de reconocimiento facial, Joy Buolamwini, lanzó la Liga de Justicia Algorítmica[3] para denunciar el impacto de la IA en las capas de población más vulnerables y racializadas: "Los algoritmos, como los virus, pueden propagar el sesgo a gran escala, a un ritmo rápido" (Buolamwini, 2018).

Nuevas estrategias avanzan y comienzan a configurar un campo de propuestas plurales bajo el enfoque de una IA feminista (Toupin, 2023) que pone el foco tanto en los modelos y en el proceso de selección y curación de datos para localizar brechas y corregir sesgos, como en el propio diseño de las herramientas de IA siendo conscientes de cómo impactan negativamente en ciertos grupos. Además, reclaman el desarrollo de una IA responsable, replanteando también los discursos e imaginarios asociados a la misma y continuar con la tarea del análisis crítico de los significados de una tecnología claramente asociada a valores androcéntricos.

---

2. COMPAS (*Correctional Offender Management Profiling for Alternative Sanctions)* está basado en un algoritmo opaco de predicción de riesgos. Utilizado durante un tiempo en los juzgados estadounidenses, asignaba a los nuevos reos una tasa de reincidencia y tal factor era tomado en cuenta a la hora de dictar sentencias. Se comprobó que esas tasas eran más altas en la población racializada y en situaciones de mayor vulnerabilidad y desigualdad social. Recuperado de: https://www.technologyreview.es//s/13800/caso-practico-probamos-por-que-un-algoritmo-judicial-justo-es-imposible.

3. La Liga de la Justicia Algorítmica (https://www.ajl.org) comenzó con el proyecto de la película *Coded Bias* (2020). El documental comienza con la historia de la propia Joy Buolamwini, quien narra su experiencia con los sistemas de reconocimiento facial siendo estudiante en el MIT. Siendo una mujer de color experimentó cómo no funcionaba con ella, no reconocía su cara, ya que el sistema había sido entrenado con personas no racializadas. Tras este punto de partida se destaca las historias de personas que se han visto afectadas por la nueva tecnología y muestra a mujeres pioneras haciendo sonar la alarma sobre las amenazas que la inteligencia artificial representa para los derechos civiles.

10.52712/issn.1850-0013-555

> "*This awareness will encourage a generation of feminist activists, programmers, designers, and engineers to have even more incentive to engage with various areas of AI, because they will realize that the stakes are enormous: designing AI systems is simultaneously designing human systems*" (Hayles, 2023, p. 16).

La IA está transformando radicalmente todos los sistemas humanos, impacta claramente en los hábitos y pautas de vida y en nuestras relaciones con los demás y con nuestro entorno, en los procesos de tomas de decisión en nuestras sociedades y en sus formas de regulación de la vida social, pero también lo hace en los procesos de generación del conocimiento. Es decir: en el núcleo mismo de las habilidades cognitivas humanas al servicio de la construcción de la ciencia.

2. Procesos de generación del conocimiento e IA. ¿Desestabilización epistemológica?

¿El saber de la máquina es simplemente una techné, en el sentido de una destreza o pericia para efectuar una tarea, o puede generar también una episteme? Eso se pregunta el filósofo Dardo Scavino en Máquinas Filosóficas (Scavino, 2022). Porque son los trabajos inmateriales, cognitivos, simbólicos, culturales, la actividad comunicativa de la sociedad en general y la capacidad deliberativa de la ciudadanía la que está ahora atravesada (o sustituida) por las máquinas. Ya no es necesario instruirlas, aprenden solas y son capaces de encontrar reglas, correlaciones entre las x y las y, ni siquiera previstas por los seres humanos (lo que supone avances significativos en disciplinas como la medicina, si de lo que se trata es de diagnosticar con precisión una enfermedad a partir de múltiples y variados síntomas), pero la máquina no sabe por qué es así (Scavino, 2022). Dar cuenta de los porqués implica hacer explícitos los marcos de referencia, los supuestos previos, los valores y los objetivos trazados, una tarea eminentemente humana que, asistida por estos sistemas de IA, puede incrementar el éxito y la velocidad de respuestas para proceder a la toma de decisiones, pero no sustituirla.

En un reciente estudio publicado en *Nature* (Van Noorden y Perkel, 2023) sobre los resultados de una encuesta realizada a 1600 investigadores de todo el mundo, se relata cómo la percepción generalizada es que las nuevas herramientas de IA serán muy importantes, e incluso esenciales, en los trabajos de investigación y generación del conocimiento, pero también expresan preocupación por las formas en las que se verán transformadas estas actividades. Es obvio que las técnicas estadísticas de *machine learning*, incluyendo los modelos generativos de lenguaje (LLM, por la expresión en inglés *large language model*), facilitan muchas de las tareas asociadas a los protocolos de generación y presentación de la información relevante (desde nuevas formas más rápidas y eficaces de procesar datos hasta la generación de gráficas o la elaboración de informes, artículos, escribir código, etc.). Incluso, como se señala en el informe, pueden ayudar a producir nuevas estructuras de proteínas o sugerir diagnósticos médicos, entre otros múltiples resultados que, sin embargo, requieren de la asistencia o revisión humana, porque los errores pueden ser fatales. Una de las preocupaciones mostradas por las personas encuestadas es que, guiados por la aparente eficacia,

rapidez y pulcritud de estos procesos, sesgos y discriminación en los datos acaben siendo desatendidos; además, temen que se tienda a dar más relevancia a los patrones reconocidos por la IA sin que haya implicada una comprensión real de tales patrones, y un número importante de investigadores considera que un uso inadecuado puede conducir a resultados de investigación no reproducibles y, directamente, al fraude. Destacados investigadores afirman que el principal problema es que la IA está cambiando los actuales estándares de prueba y de verdad. Las consecuencias son previsibles y la preocupación muy alta: proliferación de la desinformación, el plagio es más fácil y difícil de detectar, puede producir resultados de investigación falsos y más difíciles de detectar, los sesgos circulan con mayor facilidad y acaban produciendo resultados de investigación sesgados plasmados como conocimiento bien establecido en los textos. Por ello, y a pesar de las innumerables ventajas como asistentes de investigación, "los investigadores han advertido en repetidas ocasiones que el uso ingenuo de las herramientas de IA en la ciencia puede conducir a errores, falsos positivos y hallazgos irreproducibles, lo que podría suponer una pérdida de tiempo y esfuerzo" (Van Noorden y Perkel, 2023, pp. 673-674).

De hecho, nuevas orientaciones para actuar éticamente en los procesos de construcción, difusión y transferencia de conocimientos empiezan a trazarse y proponerse a la comunidad investigadora. Así, unas directrices vivas para el uso responsable de la IA generativa en la investigación han sido publicadas en *Nature* en octubre de 2023.[4] Destacan, en primer lugar, que debido a que no se puede garantizar la veracidad de los resultados generados por la IA generativa, y que las fuentes no se pueden rastrear y acreditar de manera totalmente confiable, es necesario que los actores humanos asuman la responsabilidad final de los resultados científicos. Ello significa que es necesaria la verificación humana en todos los pasos del proceso investigador, desde el diseño de los objetivos de investigación y la elaboración de hipótesis, el proceso de recogida e interpretación de datos, la elaboración de informes y artículos, su revisión por pares, la identificación de sesgos y evaluación de los resultados en el proceso editorial y de difusión de resultados. Además, se debe reconocer y especificar por parte de las personas investigadoras para qué tareas han utilizado la IA generativa, y qué herramientas concretas, en las publicaciones o en las presentaciones de la investigación científica. Además, en el proceso de revisión, las revistas científicas deben declarar si utilizan las IA generativas para la selección de pares y en qué tareas de la revisión son utilizadas por parte de los revisores. Estos son solo unos aspectos relevantes de la multitud de transformaciones que las IA generativas provocarán en el propio proceso de construcción del conocimiento y en su difusión y transferencia, y la necesidad de dotarnos de normas para garantizar los niveles adecuados de control, verificación y adecuación empírica de los resultados de investigación. Normas y códigos éticos institucionales en la investigación y en la innovación para generar una ciencia e innovación responsables son cada vez más comunes en los grandes centros de investigación, conscientes de la necesidad de desarrollar una tecnociencia confiable, y la regulación de la IA en el marco europeo avanza con paso firme instando al desarrollo de la una IA transparente y limitando sus usos más arriesgados. Son

4. Se incluyen en el informe de Bockting *et al.* (2023).

pasos hacia un verdadero control y una verdadera gobernanza democrática de los sistemas tecnocientíficos, y hacia el desarrollo de una tecnoética.

Conclusiones

Los sesgos inscritos en las tecnologías en general y en los sistemas de IA no son un fallo menor corregible con un ajuste del sistema o nuevas capas de código; quizá solo algunos de los más gruesos detectados puedan corregirse parcialmente, pero eso no resolverá la cuestión principal, y es que estas tecnologías reflejan las desigualdades persistentes y las relaciones de poder inscritas en las propias representaciones, significados y datos de nuestra cultura. La viabilidad de un desarrollo de una IA feminista ha sido explorada por grupos de especialistas; algunas de esas iniciativas están recogidas en Toupin (2024) y destaca la relevancia de conocer propuestas de trabajo, narrativas y perspectivas que son infravaloradas (Crawford, 2021) o no tienen el altavoz suficiente como la línea feminista de trabajo en *human computer interaction* (HCI). Así, la categoría IA feminista pretende reflejar las formas en las que feminismo e IA toman una multiplicidad de significados y abren nuevas líneas de investigación, dando cuenta de la pluralidad de enfoques y la capacidad de poner el foco de atención en diferentes aspectos. Toupin (2024, p. 2) señala las siguientes: IA feminista interseccional, IA feminista poshumana, IA feminista interseccional decolonial o IA transfeminista, entre otras. Un conjunto de narrativas alternativas que enriquecen y profundizan en la crítica a los impactos de la IA en nuestras prácticas y conocimientos. También permiten desvelar y atender con mayor rigor a las claves del proceso mismo de diseño de estas tecnologías: identificar los valores implicados en los modelos, en la selección de los datos, en los objetivos y en la trama misma del diseño de herramientas concretas puestas en circulación en nuestras sociedades. Nuevas tareas para la crítica y la práctica feministas cuyas líneas maestras fueron dibujadas hace décadas por las epistemologías feministas (Perdomo, 2024), que han sido tremendamente fructíferas y que nos obligan a seguir repensando la relación tecnociencia-género. Las perspectivas críticas feministas de la ciencia y la tecnología han insistido, además, en la necesidad de una reapropiación crítica de las tecnologías, que permita la participación de las mujeres (a las que se les reconozca autoridad epistémica y práctica) y que permita igualmente la generación de nuevos discursos y narrativas, una nueva cultura superadora de la desigualdad y de la injusticia epistémica (Fricker, 2007).

El camino hacia una IA responsable y feminista implica también evaluar detenidamente los efectos reales en las vidas de la ciudadanía, ya que estos sistemas son cada vez más comunes en los procesos de tomas de decisión de las administraciones públicas y de las empresas privadas. Los casos presentados son solo un pequeño ejemplo de ello y contribuyen a incrementar la exclusión y la desigualdad de los colectivos más vulnerables. Además, hay que subrayar que el desarrollo de estas máquinas inteligentes también depende de una vasta fuerza de trabajo humano invisible que etiqueta los datos que alimentan los algoritmos, limpian código y entrenan herramientas de *machine learning*. Y que esa fuerza de trabajo está compuesta especialmente de legiones de mujeres al frente de pantallas en países del sur global, infravaloradas y mal remuneradas (Gray y Suri, 2019).

Finalmente, no sería una tecnociencia socialmente responsable si no advirtiera de los riesgos de aplicarla al propio proceso de generación de los conocimientos. Un nuevo conjunto de herramientas de IA generativa facilitará la indagación científica del mundo. La gestión de los datos, las correlaciones inesperadas y los patrones antes no considerados abrirán nuevas vías de conocimiento y aplicaciones que permitirán soluciones más eficaces a multitud de problemas. Pero no serán soluciones valiosas si lo son a costa del incremento de vulnerabilidad de los seres vivos y de nuestro propio mundo.

Financiamiento

Proyecto de investigación: "Vulnerabilidad, precariedad y brechas sociales. ¿Hacia una redefinición de los derechos fundamentales?" - PID2020-114718RB-I00, financiado por el Ministerio de Ciencia e Innovación del Gobierno de España (2021-2025). IP: Vicente Navarro y María José Guerra.

Responsabilidad en la investigación

Para la redacción de este texto no se ha hecho uso de modelos generativos de lenguaje tipo ChatGPT u otros.

Bibliografía

Bockting *et al.* (2023). Living guidelines for generative AI -why scientists must oversee its use. Nature, 622, pp. 693-696.

Coeckelbergh, M. (2021). Ética de la Inteligencia Artificial. Madrid: Cátedra.

Crawford, K. (2021). Atlas of AI: Power, Politics, and the Planetary Costs of Artificial Intelligence. New Haven: Yale University Press.

Criado Pérez, C. (2019). Invisible Women: Exposing Data Bias in a World Designed for Men. Londres: Chatto & Windus.

Dastin, J. (2022). Amazon scraps secret AI recruiting tool that showed bias against women. En K. Martin (Comp.), Ethics of Data and Analytics. Concepts and cases (296-299). CRC Press Taylor & Francis Group.

Eubanks, V. (2018). Automating Inequality: How High-Tech Tools Profile, Police and Punish the Poor. Nueva York: St. Martin's Press

Fricker, M. (2007). Epistemic Injustice. Power & the Ethics of Knowing. Oxford: Oxford University Press.

Gray, M. & Suri, S. (2019). Ghost Work: How to Stop Silicon Valley from Building a New Global Underclass. Boston: Harcourt.

Han, B. C. (2022). Infocracia. Madrid: Taurus.

Hayles, N. K. (2023). Technosymbiosis: Figuring (Out) Our Relations to AI. En Feminism and AI. En J. Browne, S. Cave, E. Drage & K. McInerney (Eds.), Feminist AI: Critical Perspectives on Algorithms, Data, and Intelligent Machines (1-18). Oxford: Oxford University Press.

Jasanoff, S. (2004). States of Knowledge. The Co-production of Science and Social Order. Londres: Routledge.

Jasanoff, S. (2016). The Ethic of Invention. Technology and The Human Future. Nueva York: W.W. Norton & Company Ltd.

Larson, E. J. (2022). El mito de la Inteligencia Artificial. Por qué las máquinas no pueden pensar como nosotros lo hacemos. Barcelona: Shackleton Books.

O'Neil, C. (2016). Weapons of Math Destruction: How Big Data Increases Inequality and Threatens Democracy. Londres: Allen Lane.

Perdomo, I. (2024). Tecnociencia feminista. Una propuesta de demarcación. Revista Iberoamericana de Ciencia, Tecnología y Sociedad -CTS, 19(55), 127-143. DOI: https://doi.org/10.52712/issn.1850-0013-424.

Scavino, D. (2022). Máquinas filosóficas. Problemas de cibernética y desempleo. Barcelona: Anagrama.

Toupin, S. (2024). Shaping feminist artificial intelligence. New Media & Society, 26(1), 580-595. Sage Journals. DOI: https://doi.org/10.1177/14614448221150776.

Van Noorden, R. & Perkel, J. (2023). AI and Science: what 1600 researchers think. Nature, vol. 621, 28 de septiembre de 2023, 672-675. DOI: https://doi.org/10.1038/d41586-023-02980-0.

Wajcman, J. (2010). Feminist Theories of Technology. Cambridge Journal of Economics, 34(1), 144.

Wajcman, J. & Young, E. (2023). Feminism Confronts AI. En Feminism and AI. En J. Browne, S. Cave, E. Drage & K. McInerney (Eds.), Feminist AI: Critical Perspectives on Algorithms, Data, and Intelligent Machines (47-64). Oxford: Oxford University Press.

Young, E., Wajcman, J. & Sprejer, L. (2023). Mind the gender gap: inequalities in the emergent professions of artificial intelligence (AI) and data science. New Technology, Work and Employment, 1-24. DOI: https://doi.org/10.1111/ntwe.12278.