EDICIONES COMPLUTENSE

Investigaciones Feministas
ISSN-e: 2171-6080

MISCELÁNEA

Hacia una Inteligencia Artificial Interseccional para el tratamiento de información

Elisa Simó Soler
Universitat de València (España)

https://dx.doi.org/10.5209/infe.81954 Recibido: Diciembre 2022 • Revisado: Mayo 2023 • Aceptado: Junio 2023

Resumen: Introducción. El auge de sistemas de Inteligencia Artificial para el tratamiento de información y la relevancia empírica del enfoque interseccional incita la búsqueda de un punto de confluencia entre ambos métodos de análisis. **Objetivos.** El potencial analítico que ofrecen los sistemas de aprendizaje automático, dada su elevada capacidad de procesamiento de datos masivos, sumado a la mejor disposición del enfoque interseccional para abordar problemáticas sociales al incidir sobre los diferentes ejes de identidad/opresión que vertebran la posición de las personas, invita a plantear una simbiosis entre ellos hasta alcanzar una Inteligencia Artificial Interseccional. **Metodología.** Para ello, se realiza una aproximación conceptual a ambos métodos y se estudian los tres momentos en los que podría evaluarse la interseccionalidad de los sistemas de Inteligencia Artificial: en la configuración de las bases de datos de entrenamiento, en el descubrimiento de correlaciones entre variables durante el desarrollo de los modelos y, finalmente, en la fase de auditoría como categoría de fiabilidad del sistema. **Resultados.** Tras la revisión doctrinal y de supuestos empíricos ya desarrollados, se observa cómo es posible poner al servicio de la sociedad una Inteligencia Artificial que, lejos de generar sesgos, contribuya a la visibilización de realidades olvidadas y colectivos discriminados desde una perspectiva interseccional. **Conclusión.** En una sociedad democrática, una Inteligencia Artificial Interseccional no solo es posible sino deseable como herramienta para potenciar la diversidad y la inclusión.
Palabras clave: Inteligencia Artificial; Interseccionalidad; visibilización; inclusión; diversidad; derechos humanos.

ENG Towards Intersectional Artificial Intelligence for Information Processing

Abstract: Introduction. The rise of Artificial Intelligence systems for information processing and the empirical relevance of the intersectional approach encourages the search for a confluence point between both analysis methods. **Objectives.** The analytical potential offered by machine learning systems, given their high capacity for processing massive data added to the better disposition of intersectional approaches to address social problems by focusing on the different axes of identity/oppression that vertebrate the position of people, invites us to propose a symbiosis between them until reaching an Intersectional Artificial Intelligence. **Methodology.** To this end, a conceptual approach to both methods is made, and the three moments in which the intersectionality of Artificial Intelligence systems could be tested are studied: in the configuration of the training databases, in the discovery of correlations between variables during the development of the models and, finally, in the audit phase as a category of system reliability. **Results.** After reviewing the doctrinal and empirical assumptions already developed, it is observed how it is possible to place at the service of society an Artificial Intelligence that, far from causing biases, contributes to the visibility of forgotten realities and discriminated groups from an Intersectional perspective. **Conclusion.** In a democratic society, an Intersectional Artificial Intelligence is not only possible but desirable as a tool to promote diversity and inclusion.
Keywords: Artificial Intelligence; Intersectionality; visibility; inclusion; diversity; human rights.

Cómo citar: Simó Soler, E. (2024). Hacia una Inteligencia Artificial Interseccional para el tratamiento de información. *Investigaciones Feministas* 15(1), 137-144. https://dx.doi.org/81954

Rainbows include the whole spectrum of different colours,
but how many colours we distinguish depends on
our specific social and linguistic milieu.
Nira Yuval-Davis

1. Introducción

Manifestaba, sin ningún tipo de pudor, el escritor Bob Pop (2021) en la semana por los Derechos Humanos de la Universitat de València que *se legisla desde el privilegio*. Estas palabras ponen en cuestionamiento la vigencia de uno de los valores fundacionales del Ordenamiento Jurídico español en tanto que niegan el contenido del artículo 14 de la Constitución Española según el cual "[l]os españoles son iguales ante la ley, sin que pueda prevalecer discriminación alguna por razón de nacimiento, raza, sexo, religión, opinión o cualquier otra condición o circunstancia personal o social". La mencionada desigualdad frente a la ley fue denunciada por el feminismo jurídico advirtiendo que las notas de universalidad, objetividad y neutralidad otorgadas al Derecho eran caracteres ilusorios (Mestre i Mestre, 2008, 21-22). La elaboración de las leyes y su posterior interpretación y aplicación tomando como sujeto de referencia al hombre cis, blanco, heterosexual, adulto, sin discapacidad, con recursos económicos y urbanita ha supuesto la marginalización del resto de personas necesitadas igualmente de protección.

Tomar como punto de partida la advertencia de que ni siquiera la institución que por excelencia ha sido el reflejo de imparcialidad, independencia, rectitud y asimilacionismo, en el sentido de someter por igual a todas las personas al imperio de la ley, consigue finalmente serlo, lleva a tener que interrogarse acerca de si los nuevos paradigmas, entre los que se encuentra la algoritmización de la justicia (en el sentido digital), pueden generar situaciones de invisibilización y discriminación. Ha quedado patente la preocupación sobre los sesgos de los sistemas de Inteligencia Artificial (IA).

Son conocidas las investigaciones sobre predictibilidad de futuros/as criminales mediante la evaluación de COMPAS que perjudicaba a las personas negras y los supuestos de discriminación hacia las mujeres en múltiples dispositivos como los sistemas de reconocimiento de voz o faciales, la minería de textos o modelos de aprendizaje automático (o *machine learning* en inglés) empleados en el ámbito de la salud, la contratación y la justicia (ProPublica, 2016; European Comisssion, 2019). Sin embargo, el estudio sobre la aplicación de los sistemas de IA típicamente ha atendido a un único eje, raza o género en este caso, mientras que la identidad de una persona y su posición en la sociedad no viene determinada de forma unidimensional, sino por la concurrencia de múltiples categorías atravesadas por diferentes sistemas de opresión o de privilegio, según quien sea el sujeto sobre el que se haga la lectura (Nash, 2008, 9-10).

El análisis interseccional ha ocupado una posición central en los estudios feministas y antirracistas esforzándose por no articular una visión monolítica de *la mujer* (Cho, Crenshaw y McCall, 2013, 787) y su incorporación como metodología garantista de los derechos humanos se ha producido en otros campos de estudios como la sociología, el derecho y las ciencias políticas (Yuval-Davis, 2006, 206). Tratándose de una apuesta teórico-práctica encaminada a desvelar las diversas capas que conforman al individuo y asumiendo la injerencia de los modelos de aprendizaje automático en el espacio personal y profesional de las personas, resulta pertinente interrogarse acerca de la posible confluencia de ambos métodos de análisis.

¿Es posible aplicar un enfoque interseccional en los sistemas de IAI? ¿Tiene sentido entrenar algoritmos con bases de datos interseccionales? ¿Los algoritmos de aprendizaje automático pueden ofrecer correlaciones ("intersecciones") no esperadas entre variables? ¿Debe ponderarse el proceso de supervisión de los sistemas de IA atendiendo a la interseccionalidad como factor de análisis? A partir de la formulación de estas cuestiones se pretende profundizar en la posible relación simbiótica entre interseccionalidad e IA.

2. La interseccionalidad como herramienta metodológica

La idea de interseccionalidad, acuñada por Crenshaw (1989), constituye el enclave epistemológico para teorizar sobre las múltiples realidades que atraviesan a las personas y descartar el marco analítico por el cual se trata la discriminación de los grupos en situación de vulnerabilidad desde un único prisma. A partir de una crítica a los análisis unidimensionales, hegemónicos en la teoría feminista y antirracista y presentes también en el ámbito jurídico, la autora realiza una propuesta de marco teórico multidimensional a partir del cual examinar la violencia contra las mujeres negras (Sales Gelabert, 2017, 231). Para Crensahw, el estudio de la posición que ocupan las mujeres negras no puede abordarse desde parcelas cognitivas impermeabilizadas sino desde el entrecruzamiento de espacios y experiencias. El enfoque interseccional advierte que su metodología no se basa en un solapamiento de variables sino en la interacción sinérgica, en palabras de MacKinnon (2013, 1024). De esta forma, no puede pretenderse un planteamiento sumativo sino relacional entre las diferentes categorías que confluyen en un punto de convergencia. Para profundizar en el esquema conformador de las identidades y los ejes de discriminación, Crenshaw desarrolla una triple clasificación basada en lo que denomina interseccionalidad estructural, política y representativa.

La interseccionalidad estructural pretende poner de manifiesto la forma en que la experiencia de las mujeres negras en la intersección de la raza y el género (en el caso de Crenshaw, pero extensible a otras categorías identitarias) hace que la vivencia de la violencia sea cualitativamente diferente a la de otras mujeres blancas (Crenshaw, 1991, 1124). Por su parte, la interseccionalidad política incluye el enfoque multieje en la

esfera de las políticas públicas y revela la marginación de determinados colectivos que se hallan en la intersección, pese a los postulados progresistas y emancipadores de las iniciativas feministas y antirracistas. Las mujeres negras se sitúan dentro de al menos dos grupos subordinados cuyos sujetos de referencia definen y confinan los intereses de todo el colectivo y cuyas agendas políticas con frecuencia son contradictorias. Para la autora, dado que es el hombre negro quien determina los parámetros de las estrategias antirracistas y la mujer blanca la que fundamenta los intereses de las mujeres, cualquier propuesta política que provenga de dicho antirracismo y feminismo estará limitada en sus propios términos, debido a que las mujeres negras experimentan el racismo y el sexismo de formas diferentes a las experimentadas por los hombres negros y las mujeres blancas (Crenshaw, 1991, 1251-1252).

Por último, la interseccionalidad representativa atiende a la construcción cultural de las mujeres negras. Su representación en el imaginario colectivo condiciona la relevancia y la respuesta otorgada a los actos de violencia contra las mujeres. Aún más, la desatención recíproca entre los discursos antirracista y feminista genera un desempoderamiento interseccional (Crenshaw, 1991, 1282). Forzando al extremo esta ceguera mutua, desde el feminismo es posible contribuir al fortalecimiento de enfoques punitivistas y criminalizadores de los hombres migrantes al tiempo que se invisibiliza la realidad de las mujeres migrantes que sufren violencia, como ocurrió con las jornaleras de Huelva que recibieron menor cobertura mediática y apoyo social que la víctima de la Manada.

La interseccionalidad se constituye como una teoría de la identidad y la opresión, ya que descarta la configuración monista del individuo apelando a una identidad poliédrica y enfatiza la existencia de múltiples ejes de discriminación que diversifican la experiencia de los diferentes sujetos que conforman un colectivo (Nash, 2008, 3). De este modo, tomando como ejemplo la aplicación operada en el seno de los estudios de género, puede advertirse cómo el enfoque interseccional fractura el ideal universalista de *la mujer* para incorporar otras realidades no hegemónicas, pero igualmente conformadoras del colectivo de *las mujeres* que descubren intereses y condiciones diversas en la esfera interna.

Del mismo modo que la teoría crítica feminista evidencia y anula la construcción del universal en torno al hombre, debe producirse una revisión inclusiva similar *ad intra* del feminismo que permita visibilizar el desarrollo de la individualidad con distintos atributos diferentes a los de las mujeres blancas, cis, heterosexuales, de clase media, sin discapacidad y urbanistas y sus experiencias vitales (Cubillos Almendra, 2015, 121-122; McCall, 2005, 1771; Viveros Vigoya, 2016, 8; Crenshaw, 1989, 154). En caso contrario, el enfoque resultará limitado al no considerar otros factores diferenciales y estructurales (clase social, raza, orientación sexual, edad, etnia, capacidad, ruralidad como subsistemas de subordinación) que determinan la posibilidad de sufrir una discriminación en un sistema dado (capitalista, heteronormativo, patriarcal, blanco, capacitista y urbano) condicionando su vivencia y la posibilidad de permanecer en los márgenes de las instituciones formales de gobierno así como de la construcción del conocimiento (Guzmán Ordaz y Jiménez Rodrigo, 2015, 600; La Barbera, 2017, 195; Collins, 2017, 23). Así, es a partir de la incorporación de la interseccionalidad como se entiende la construcción integral, justa y polifónica de los relatos y categorías identitarias.

Entre los aspectos críticos que incorpora el enfoque interseccional se halla el carácter dinámico de las identidades y la naturaleza codependiente de la opresión y el privilegio. Plantea Viveros Vigoya (2016, 8) que:

> "[e]n algunas ocasiones, el género crea la clase, como cuando las diferencias de género producen estratificaciones sociales en el ámbito laboral. En otras, las relaciones de género son utilizadas para reforzar las relaciones sociales de raza, como cuando se feminiza a los hombres indígenas o se hipermasculiniza a los hombres negros; inversamente, las relaciones raciales sirven para dinamizar las relaciones de género, como cuando se crean jerarquías entre feminidades y masculinidades a partir de criterios raciales".

Bajo esta ejemplificación es posible atribuir un carácter fluido y cambiante a lo largo del tiempo y en los diferentes contextos sociopolíticos a las subjetividades, el poder y el privilegio. Los sistemas de opresión-privilegio son mutuamente constitutivos y fluctuantes, intercambiando al sujeto en la posición de subordinación (Nash, 2008, 11-12). Un ejemplo gráfico es el lugar privilegiado de una empresaria como Ana Botín, de una presidenta como Angela Merkel o de una modelo como Georgina Rodríguez frente a un trabajador de su compañía, un ciudadano del país germano o un asistente de la celebridad.

La constitución del eje opresión-privilegio deja de ser inmutable (Shields, 2008: 304). Sobre esta cuestión Collins introduce el concepto de "matriz de dominación" en un nivel de análisis macro para referirse al modo en el que se organiza la dominación política a través de sistemas de opresión entrelazados. Para la autora, el heteropatriarcado, el neocolonialismo, el capitalismo, el racismo y el imperialismo constituyen formas de dominación que caracterizan la geopolítica mundial, adoptan siluetas heterogéneas en los Estados-nación e influyen en todos los aspectos de la vida social. El aporte de la interseccionalidad consiste en sugerir diferentes manifestaciones de dominación, adoptando cada una su propia red de poder al contexto-sujeto y generando una "matriz" distintiva de dinámicas de poder que se entrecruzan, sostienen y multiplican (Collins, 2017, 22).

A modo de recapitulación, son útiles los principios propuestos por Collins (2017, 22) como síntesis definitoria de la interseccionalidad:

> "(1) racismo, sexismo, explotación de clase y sistemas de opresión similares están interconectados y se construyen mutuamente; (2) la configuración de las desigualdades sociales toma forma dentro de las opresiones interseccionales; (3) las percepciones de los problemas sociales también reflejan cómo se sitúan los actores sociales en las relaciones de poder de determinados contextos históricos y sociales; y (4) dado que los individuos y los grupos y se sitúan de forma diferente dentro de las opresiones que se entrecruzan, tienen puntos de vista distintos sobre los fenómenos sociales".

3. Una propuesta de confluencia: Inteligencia Artificial Interseccional

Pese al auge, tanto académico como práctico y mediático, relativamente reciente de la IA y la interseccionalidad, ambas han sido utilizadas con anterioridad. Cabe recordar que la preocupación por la conexión entre las categorías de género, raza y clase se manifestó con anterioridad a su conceptualización como interseccionalidad. Ya en 1791, Olympia de Gouges, quien redefiniera la declaración de derechos *rousseauniana* en favor de las mujeres, comparaba la dominación colonial y la situación de los esclavos con la realidad patriarcal de las mujeres.

En 1851, Sojourner Truth, interpeló a su audiencia en la convención por los derechos de las mujeres en Ohio con la pregunta "¿Acaso no soy una mujer?" para denunciar la doble opresión que sufría por ser una mujer negra, confrontando su situación con las mujeres blancas burguesas. Desde Latinoamérica, el feminismo postcolonial se manifestaba a través de diferentes expresiones artísticas de denuncia de la opresión de las mujeres indígenas y negras. En el siglo XX, antes del aporte conceptual de Crenshaw, la Colectiva del Río Combahee contribuyó debatiendo acerca de la hegemonía del feminismo blanco al ensamblaje de la teoría de la interseccionalidad atendiendo a las categorías de raza, género y clase en la experiencia de las mujeres negras (Viveros Vigoya, 2016, 3-4).

Por otra parte, es posible remontarse a la era de la Grecia clásica y del antiguo Egipto para comprobar cómo la resolución de diferentes problemas de forma automática simulando el comportamiento humano ha sido uno de los objetivos perseguidos a lo largo de la historia (Rainer Granados y Rodríguez Baena, 2017, 19-20) Ya en 1842, Ada Lovelace, primera programadora de la historia y matemática, escribió el primer algoritmo para ser procesado por una máquina. En 1850, Alan Turing, formuló una prueba, renombrada y conocida como "Test de Turing" para averiguar si una máquina exhibía un comportamiento inteligente. Seis años más tarde, John McCarthy organizó la conferencia de Darmouth, considerada el momento fundacional de la IA al acuñar este término y reivindicar la replicación de la inteligencia humana en máquinas digitales.

Estos son solo algunos de los hitos en la revolución de la IA que sirven como precedente para justificar el empleo de estos sistemas antes del imponente desarrollo en computación e informática que ha marcado un punto de inflexión. Concretamente, los avances de *hardware* especializado, el aumento de datos digitalizados y servicios de etiquetado a bajo costo, ha situado a la IA como uno de los campos que mayor interés suscita por su versatilidad para aplicarse en ámbitos muy diversos (Abeliuk y Gutiérrez, 2021, 17).

Como se ha mencionado, dada su preeminencia en diversas disciplinas de estudio, conviene investigar la posible confluencia entre IA e interseccionalidad con el objetivo de ofrecer una fiel representación de la realidad y potenciar la eficacia de ambos instrumentos de trabajo. Podría anticiparse que el encuentro entre IA e interseccionalidad constituye una asociación simbiótica, al beneficiarse los sistemas de IA de la riqueza de los análisis interseccionales, por un lado, y el enfoque intrínsecamente interseccional del potencial analítico que ofrecen los modelos de aprendizaje automático, por otro. En las páginas que siguen se pretende profundizar en los interrogantes planteados al inicio del texto para dar respuesta al primero de ellos: ¿es posible aplicar un marco interseccional en los sistemas de IA?

3.1. ¿Tiene sentido entrenar algoritmos con bases de datos interseccionales?

Una de las principales críticas a los sistemas de IA ha sido la generación de resultados discriminatorios. No solo se denuncia la perpetuación de la subordinación de los colectivos históricamente oprimidos sino la amplificación de dicha situación dado el potencial que albergan dichos modelos. Debido a la elevada capacidad de procesamiento de datos masivos y su uso prácticamente generalizado, tanto en el sector privado como en el público, sin la adopción de medidas correctivas, la vulneración del derecho a la igualdad se convierte en una problemática de primer orden.

Fruto de las investigaciones sobre esta cuestión se ha puesto de relieve la importancia de la selección de los datos de entrenamiento. Una de las principales reivindicaciones en las investigaciones estadísticas, pronunciadas incluso en el marco de Naciones Unidas (Yuval-Davis, 2006, 204), es la obtención de datos desagregados, esto es, la subdivisión de una muestra en sus componentes individuales de información. Tanto las bases de datos como los modelos de aprendizaje automático tienden a ser entrenados con un etiquetado binario o con categorías limitadas que, si bien cumple con la función de búsqueda de diferencias o patrones particulares, no lo hace con el nivel de complejidad que la realidad requiere (Costanza-Chock, 2018, 6). Sin una descomposición interseccional no es posible alcanzar una representación fiel del entorno, debiendo considerar aún más que la suma de datos puede favorecer la interseccionalidad. La representación de los grupos minoritarios y de las relaciones entre atributos más marginales –no por ello menos importantes–, podría beneficiarse del aumento de la información recolectada. El tamaño de la base de datos no solo incrementa la robustez estadística del entrenamiento del modelo, sino que permite mejorar el tratamiento de clases desbalanceadas.

Los modelos de clasificación habitualmente se enfrentan a bases de datos desbalanceadas, aquellas que presentan un desequilibrio significativo entre las diferentes categorías o clases, también conocido como Problema de Desequilibrio de Clases. Este escenario es especialmente frecuente cuando se intentan incluir categorías no hegemónicas o, en otras palabras, cuando se intenta configurar una base de datos interseccional. El modo de salvar esta infrarrepresentación podría lograrse a partir de la manipulación de la muestra, sobrerrepresentando a las categorías menos frecuentes, como una suerte de política de cuotas virtual. No obstante, la articulación de un enfoque interseccional conlleva una serie de riesgos. La aplicación de la interseccional desde un enfoque *top-down* podría derivar en una epistemología colonizadora, asumiendo de antemano las carencias sin entablar un proceso de diálogo con los sujetos situados en la periferia (Cubillos

Almendra, 2015, 131). Esto se traduce en el diseño de bases de datos que incluyan categorías y parámetros que definan la realidad que se pretende analizar de manera fehaciente. Por otro lado, la heterogeneidad de los datos podría traducirse en una pérdida de precisión en el entrenamiento de los sistemas (Buolamwini y Gebru, 2018, 12; Guo y Caliskan, 2020): los datos marginales podrían ser malinterpretados como ruido disminuyendo el rendimiento de los modelos para estas categorías.

En todo caso, la incorporación de la interseccionalidad en el ámbito de los algoritmos de aprendizaje automático tiene una incidencia directa en la diversidad y la inclusión social si se atiende al potencial amplificador de la IA. Los colectivos subalternos recibirían un trato más justo y equitativo, pudiendo disminuir los efectos discriminatorios de este tipo de sistemas. Además, desde una vertiente orgánica, parece lógico apostar por la creación de grupos diversos en el diseño de bases de datos tanto por la propia función integradora como por la mayor probabilidad de que sus miembros conozcan en profundidad todas las aristas identitarias y características sociales implicadas en el análisis (Ireni-Saban y Sherman, 2021, 48).

3.2. ¿Los algoritmos de aprendizaje automático pueden ofrecer correlaciones ("intersecciones") no esperadas entre variables?

El uso de la estadística inferencial tradicional en ciencias sociales está siendo reemplazado por las técnicas de aprendizaje automático. En el paradigma de la estadística inferencial se suele emplear un marco deductivo para proponer una relación matemática general entre una o más variables. El resultado se infiere a partir de unas premisas teóricas, unas hipótesis preestablecidas, que se presuponen en la muestra. Por su parte, en los sistemas de aprendizaje automático prima la lógica inductiva. Se recopila un conjunto de datos, se elige un algoritmo para descubrir patrones y, posteriormente, se propone una teoría más general basada en las correlaciones encontradas que pueden implicar o no causalidad (Nelson, 2021, 2-3).

Tanto los modelos de aprendizaje automático supervisado como no supervisado pueden ofrecer representaciones de datos mucho más complejas y predecir con mayor precisión un resultado o clase (supervisado) o identificar patrones emergentes (no supervisado) a partir de un conjunto de variables de entrada. Debido a la confluencia entre una lógica fundamentalmente inductiva y las representaciones sofisticadas de datos, uno de los enfoques de investigación que podría beneficiarse del potencial del aprendizaje automático es la interseccionalidad (Nelson, 2021, 2-3). El valor añadido del aprendizaje automático en comparación con los modelos tradicionales reside, como se puede observar, en el tratamiento de datos multidimensionales y en la estimación de correlaciones, incluidas también las no aparentes (no evidentes a simple vista), según los marcos conceptuales o culturales definidos.

Un ejemplo paradigmático de cómo la IA puede superar los límites del conocimiento humano es el desarrollo en 2018 de un algoritmo de aprendizaje profundo (o *deep learning* en inglés) entrenado para predecir factores de riesgo cardiovascular en fotografías del fondo de ojo que, inesperadamente, aprendió a predecir el sexo de una persona, simplemente analizando la retina cuando, científicamente, se desconoce cuáles son las diferencias entre mujeres y hombres del tejido interno del globo ocular. Este hallazgo pone de manifiesto no solo que la IA logra captar estructuras desapercibidas, sino que, incluso, desvela relaciones que somos incapaces de explicar racionalmente (Poplin *et al.*, 2019; Korot *et al.*, 2021).

Además, la cosmovisión humana constriñe la capacidad de experimentación impidiendo el ensayo de relaciones no presumibles (más si cabe en supuestos de alta complejidad), mientras que los modelos de IA, desprovistos de ese componente cultural, solo conocen parámetros y exploran todo tipo de conexiones entre ellos. A partir de este razonamiento cabría plantear la posibilidad de que los sistemas de IA pudieran otorgar análisis y predicciones con un mayor grado de interseccionalidad que el propio ser humano.

De hecho, se tiene constancia de varios proyectos de creación de sistemas algorítmicos interseccionales, desde clasificadores que proporcionan una mayor protección a grupos minoritarios sin una pérdida significativa de rendimiento (Foulds et al., 2019, 1921) hasta el uso de *word embeddings*[1] para desarrollar análisis interseccionales sobre categorías e instituciones sociales (Nelson, 2021, 2 y 12). También surgen iniciativas de auditorías de bases de datos, realizando una ponderación sencilla con conteos para elaborar rankings (Bryant y Howard, 2019), y de modelos para valorar la equidad utilizando técnicas complejas de agrupamiento de subgrupos (Cabrera *et al.*, 2019).

3.3. ¿Debe ponderarse el proceso de supervisión de los sistemas de IA atendiendo a la interseccionalidad como categoría de análisis?

Las auditorías, entendidas como mecanismos para investigar la funcionalidad de los algoritmos, están demostrando cumplir una función esencial respecto a la detección de sesgos y otros comportamientos no deseados de los sistemas de IA sin necesidad de conocer los detalles específicos de su diseño. Dado que consisten en una verificación sistemática del cumplimiento de criterios objetivos, las auditorías se focalizan en los efectos problemáticos sobre los resultados de los modelos automatizados de toma de decisiones cumpliendo con una función de certificación de la confiabilidad de los algoritmos (Mohseni, Zarei y Ragan, 2021, 3). La consecución de una IA confiable constituye una aspiración fundacional, en palabras del "Grupo Independiente de Expertos de Alto Nivel sobre Inteligencia Artificial" creado por la Comisión Europea en junio de 2018.

---

1 Técnica de aprendizaje automático en el campo del procesamiento de lenguaje natural donde se representan vectorialmente palabras o frases con números reales.

La condición para que las personas confíen en el desarrollo tecnológico y sus aplicaciones reside en la adopción de un marco claro y detallado que garantice su fiabilidad (Ausín, 2021, 13)[2]. Para ello, se establecen cuatro principios éticos basados en los derechos fundamentales –respeto de la autonomía humana, prevención del daño, equidad y explicabilidad– que tienen su traducción posterior en siete requisitos para una IA confiable –1) acción y supervisión humanas, 2) solidez técnica y seguridad, 3) gestión de la privacidad y de los datos, 4) transparencia, 5) diversidad, no discriminación y equidad, 6) bienestar social y ambiental, 7) rendición de cuentas– que deben ser evaluados constantemente a partir de métodos técnicos y no técnicos a lo largo de la vida de los sistemas de IA (Grupo de Expertos de Alto Nivel sobre Inteligencia Artificial, 2019, 12).

El fundamento sobre el que se alzan las exigencias de confiabilidad es un enfoque centrado en las personas para asegurar el respeto a los valores sociales y derechos fundamentales, así como la instrumentalización de los sistemas de IA para la mejora de la vida humana, no humana y de la naturaleza (Cortina Orts, 2019, 9-10). Sin embargo, tal y como apuntan autoras como Constanza-Chock (2018, 6), pese a que se ha avanzado en el objetivo de exigir justicia, responsabilidad y transparencia a los modelos de aprendizaje automático, la ausencia de un enfoque interseccional en el proceso auditor trae como resultado una evaluación incompleta. No es suficiente con situar a las personas en el centro de los avances tecnológicos, sino que su posición debe ser leída desde los múltiples ejes que las definen y potencialmente oprimen.

Son varios los estudios que revelan la importancia de considerar la interseccionalidad como uno de los criterios analíticos de los modelos de IA. Al examinar el rendimiento de los algoritmos asociados a diferentes atributos en la intersección de raza y género, Buolamwini y Gebru (2018, 6) remarcaron la importancia de realizar auditorías interseccionales al comprobar que los clasificadores funcionaban mejor para personas de piel más clara y hombres en general, mientras que se obtenían los peores resultados con mujeres negras. Un equipo del Departamento de Sistemas de Información de la Universidad de Maryland (Foulds et al., 2019) propuso la incorporación de la interseccionalidad en el diseño de un algoritmo de aprendizaje a partir de una métrica de equidad (ϵ-diferencial o DF) que tiene como objetivo mantener la equidad interseccional para la IA, mejorando los resultados respecto a los subgrupos minoritarios, circunstancia que quedó contrastada tras realizar experimentos con el conjunto de datos de reincidencia criminal de COMPAS. Siguiendo esta línea de investigación, un trabajo elaborado por Quantum Black (Morina *et al.*, 2020) profundiza sobre la equidad en relación con múltiples atributos sensibles e identifica diferentes momentos en los que evaluar la interseccionalidad del algoritmo. La equidad interseccional se puede medir en las bases de datos, en los problemas de clasificación, en los resultados y en la robustez de las métricas elegidas, consiguiendo una evaluación integral. Las posibilidades que presenta la conexión entre IA e interseccionalidad junto a la capacidad para identificar sesgos, acceder al fallo y corregirlo durante esta fase de auditoría, invita a plantear comparativamente el grado de facilidad y factibilidad que supone la corrección técnica respecto a la deconstrucción humana de esos mismos sesgos (Simó Soler y Rosso, 2022, 5).

Por último, si se conceptualiza la IA como una tecnología disruptiva, transformadora de los sistemas sociales, económicos, políticos, judiciales o naturales, con una incidencia evidente en las personas –desde una asistente virtual o la publicidad personalizada hasta el diagnóstico de una enfermedad o la evaluación de riesgos de impacto cósmico– (Ausín, 2021, 4), la interseccionalidad debería formar parte de su esquema de funcionamiento y de su modelo de evaluación. Asimismo, asumiendo que la justicia distributiva se sitúa como uno de los fines ideados para la IA, el empleo de los datos evitando supuestos de discriminación y la aplicación de los resultados con un fin inclusivo reclaman la consideración de la interseccionalidad como uno de los prismas principales en la configuración y desarrollo de los sistemas algorítmicos.

4. Conclusiones

Resulta difícilmente cuestionable la adopción de un enfoque interseccional tanto en el ámbito de las ciencias sociales como en los avances científico-técnicos. Las múltiples dimensiones que conforman la identidad de una persona obligan a realizar un análisis interseccional para obtener una imagen más fiel de la posición de cada individuo en los sistemas de privilegio-opresión que vertebran los modelos sociopolíticos, económicos, judiciales, culturales y medioambientales. El uso creciente de los sistemas de IA incitan a proponer una conexión entre IA e interseccionalidad. Tanto los datos de entrenamiento, como los resultados obtenidos, y las auditorías implementadas deberían adoptar la interseccionalidad como garantía de una mejor representación de la realidad, mostrando la pluralidad intrínseca que caracteriza a la sociedad. Una multiplicidad de atributos que alejan los análisis de patrones monolíticos y dicotómicos tendentes a invisibilizar las múltiples aristas de la identidad.

Frente a la crítica sobre la atomización neoliberal de los discursos que conlleva la interseccionalidad, se comparte la necesidad de mantener una perspectiva estructural de la discriminación (Salem, 2018, 404; Bilge, 2013, 404, 411 y 414), al tiempo que se postula la ventaja que puede suponer recibir un trato individualizado, no asignado en función de la pertenencia a un grupo, así como la contribución que la interseccionalidad, potenciada por sistemas de IA, puede ofrecer en términos democráticos (Collins, 2017, 35-36).

Una Inteligencia Artificial Interseccional no solo es posible sino deseable como herramienta para potenciar la diversidad y la inclusión. La confluencia entre ambas puede ofrecer resultados muy positivos, dando respuesta a nuevas preguntas irresolubles de acuerdo con los modelos tradicionales, provocando un cambio sistémico al incorporar como sujetos políticos a colectivos históricamente excluidos (Hancock, 2007, 249) y mejorando las garantías y vías de acceso al reconocimiento de derechos humanos y su

2 Habrá que comprobar si el Reglamento de IA del Parlamento Europeo y del Consejo es capaz de cumplir con ese fin.

visibilización. Dado que no existe una única forma de desarrollar sistemas de IA (Joyce et al., 2021, 4), cabe apostar por aquella que incluya las virtudes de la interseccionalidad para brindar herramientas empíricas destinadas al bien común.

Financiación

El artículo ha sido realizado en el marco del Proyecto "Claves para una justicia digital y algorítmica con perspectiva de género", PID2021-123170OB-I00 financiado por MCIN/AEI/10.13039/501100011033.

Referencias bibliográficas

Almendra, Javiera Cubillos (2015). La importancia de la interseccionalidad para la investigación feminista. *Oxímora revista internacional de ética y política*, *7*, 119-137.

Bilge, Sirma. (2013). Intersectionality undone: Saving intersectionality from feminist intersectionality studies. *Du Bois review: Social science research on race*, *10*(2), 405-424.

Buolamwini, Joy & Gebru, Timnit (2018, January). Gender shades: Intersectional accuracy disparities in commercial gender classification. In *Conference on fairness, accountability and transparency* (pp. 77-91). PMLR.

Bryant, De'Aira & Howard, Ayanna (2019, January). A comparative analysis of emotion-detecting AI systems with respect to algorithm performance and dataset diversity. In *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society* (pp. 377-382).

Cabrera, Ángel Alexander, et al. (2019, October). FairVis: Visual analytics for discovering intersectional bias in machine learning. In *2019 IEEE Conference on Visual Analytics Science and Technology (VAST)* (pp. 46-56). IEEE.

Casella, Marco (2015). Historia y evolución de la Inteligencia Artificial. *Inteligencia Artificial*, 14-21.

Cho, Sumi, Crenshaw, Kimberlé Williams & McCall, Leslie. (2013). Toward a field of intersectionality studies: Theory, applications, and praxis. *Signs: Journal of women in culture and society*, 38(4), 785-810.

Collins, Patricia Hill (2019). The difference that power makes: Intersectionality and participatory democracy. In *The Palgrave handbook of intersectionality in public policy* (pp. 167-192). Palgrave Macmillan, Cham.

Cortina Orts, Adela (2019). Ética de la inteligencia artificial. In *Anales de la Real Academia de Ciencias Morales y Políticas* (pp. 379-394). Ministerio de Justicia.

Costanza-Chock, Sasha. (2018). Design Justice, A.I., and Escape from the Matrix of Domination. *Journal of Design and Science*, 1-12.

Crenshaw, Kimberlé (1989). Demarginalizing the intersection of race and sex: A black feminist critique of antidiscrimination doctrine, feminist theory and antiracist politics. *u. Chi. Legal f.*, 1989(1), 139-167.

Crenshaw, Kimberlé (1990). Mapping the margins: Intersectionality, identity politics, and violence against women of color. *Stan. L. Rev.*, 43(6), 1241-1299.

European Commission (2019). *Women in Artificial Intelligence: mitigating gender bias.* Disponible en:
https://ec.europa.eu/jrc/communities/en/community/humaint/news/women-artificial-intelligence-mitigating-gender-bias

Foulds, James R. *et al.* (2020, April). An intersectional definition of fairness. In *2020 IEEE 36th International Conference on Data Engineering (ICDE)* (pp. 1918-1921). IEEE.

Gelabert, Tomeu Sales. (2017). Repensando la interseccionalidad desde la teoría feminista. *Ágora: papeles de Filosofía*, 36(2), 229-256.

Granados, José Javier Rainer y Baena, Luis Rodríguez (2019). Perspectiva histórica y evolución de la inteligencia artificial. En *La inteligencia artificial, aplicada a la defensa* (pp. 17-38). Instituto Español de Estudios Estratégicos.

Grupo de Expertos de Alto Nivel sobre Inteligencia Artificial. (2019). *Directrices éticas para una IA fiable*. Comisión Europea, 1-53.

Guo, Wei & Caliskan, Aylin (2021, July). Detecting emergent intersectional biases: Contextualized word embeddings contain a distribution of human-like biases. In *Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society* (pp. 122-133).

Guzmán Ordaz, Raquel y Jiménez Rodrigo, María (2015). La Interseccionalidad como Instrumento Analítico de Interpelación en la Violencia de Género. *Oñati Socio-Legal Series*, 5(2), 596-612.

Hancock, Ange-Marie (2007). Intersectionality as a normative and empirical paradigm. *Politics & Gender*, 3(2), 248-254.

Ireni-Saban, Liza & Sherman, Maya (2020). Incorporating Intersectionality into Ai Ethics. In *Democracy and Fake News* (pp. 40-52). Routledge.

Joyce, Kelly, et al. (2021). Toward a sociology of artificial intelligence: A call for research on inequalities and structural change. Socius, *7*, 2-11.

Korot, Edward, *et al.* (2021). Predicting sex from retinal fundus photographs using automated deep learning. *Scientific reports*, *11*(1), 1-8.

La Barbera, María Caterina. (2017). Interseccionalidad. *EUNOMÍA. Revista En Cultura De La Legalidad*, (12), 191-198.

MccalL, Leslie. (2005). The complexity of intersectionality. Signs: Journal of women in culture and society, *30*(3), 1771-1800.

Mackinnon, Catherine. (2013). Intersectionality as method: A note. *Signs: Journal of Women in Culture and Society*, *38*(4), 1019-1030.

Mestre i Mestre, Ruth. (2008). "Mujeres, Derechos y Ciudadanías". En *Mujeres, Derechos y Ciudadanías* (pp. 17-44). Tirant lo Blanch.
Mohseni, Sina, Zarei, Niloofar & Ragan, Eric D. (2021). A multidisciplinary survey and framework for design and evaluation of explainable AI systems. *ACM Transactions on Interactive Intelligent Systems (TiiS)*, *11*(3-4), 1-45.
Nash, Jennifer C. (2008). Re-thinking intersectionality. *Feminist review*, *89*(1), 1-15.
Nelson, Laura K. (2021). Leveraging the alignment between machine learning and intersectionality: Using word embeddings to measure intersectional experiences of the nineteenth century US South. *Poetics*, 88(101539), 1-14.
Poplin, Ryan, et al. (2018). Prediction of cardiovascular risk factors from retinal fundus photographs via deep learning. *Nature Biomedical Engineering*, 2(3), 158-164.
ProPublica. (2016). Machine Bias. There's software used across the country to predict future criminals. And it's biased against blacks.
https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing
Salem, Sara (2018). Intersectionality and its discontents: Intersectionality as traveling theory. *European Journal of Women's Studies*, *25*(4), 403-418.
Shields, Stephanie A. (2008). Gender: An intersectionality perspective. *Sex roles*, *59*(5), 301-311.
Simó Soler, Elisa & Rosso, Paolo (2022). Inteligencia artificial y derecho: Entre el mito y la realidad. «La destrucción algorítmica de la humanidad». *Diario La Ley*, 9982, 1-9.
Viveros Vigoya, Mara (2016). La interseccionalidad: una aproximación situada a la dominación. *Debate feminista*, 52, 1-17.
Yuval-Davis, Nira. (2006). Intersectionality and feminist politics. *European journal of women 's studies*, *13*(3), 193-209.