

Inteligencia artificial, ética y salud pública

FUNDACIÓ
VÍCTOR
GRÍFOLS
i LUCAS

Con la colaboración de:



Cuadernos de la Fundació Víctor Grífols i Lucas
Inteligencia artificial, ética y salud pública Nº 69 (2024)
Edita: Fundació Víctor Grífols i Lucas. c/ Jesús i Maria, 6 - 08022 Barcelona
fundacio.grifols@grifols.com www.fundaciogrifols.org
ISBN 978-84-09-62351-8 Depósito Legal: B 13286-2024

Inteligencia artificial, ética y salud pública

Cuadernos
de la Fundació **69**
Víctor Grífols i Lucas

SUMARIO

Pág.

Introducción

Andreu Segura y Àngel Puyol 7

Los retos éticos de la IA en la salud pública

Inteligencia artificial y salud pública

Francisco Javier García León 12

Desafíos éticos de la IA

David Casacuberta 27

Decisiones automatizadas en sanidad

Inteligencia artificial y toma de decisiones

Enrique Bernal-Delgado y Francisco Estupiñán-Romero 40

Decisiones automatizadas en sanidad: IA generativa

Jordi Vallverdú 51

Injusticia algorítmica en salud pública

Aspectos éticos asociados al uso de algoritmos

Alicia de Manuel 66

La injusticia social relacionada con el uso de la inteligencia artificial

Miguel Ángel Royo 79

Aportaciones de expertos

«Iceberg» de la inteligencia artificial y las desigualdades en la salud

Brenda Biaani y Pere Torán 98

Inteligencia artificial y relación clínica: punto de encuentro

Valle Coronado 102

Inteligencia, artificial y medicina

Bernabé Robles 105

Relación de autores 110

Títulos publicados 112

INTRODUCCIÓN

Este es el decimosegundo encuentro anual –decimotercero según se mire^a del grupo de trabajo sobre ética de SESPAS (Sociedad Española de Salud Pública y Administración Sanitaria). De todos estos encuentros, la Fundació Víctor Grífols i Lucas ha editado su correspondiente Cuaderno; es decir, una quinta parte de su actual colección, lo que es remarcable puesto que el ámbito de actuación de la Fundació es la bioética, y la salud pública no ha sido un campo particularmente atractivo para la filosofía moral hasta ahora, ni tampoco para la sanidad, a no ser que padezcamos alguna calamidad colectiva, como la reciente pandemia.

Por tanto, desde la salud pública, y específicamente desde SESPAS, es de obligado reconocimiento y gratitud la acogida, colaboración y promoción –también material– que la Fundació nos está prestando desde hará casi quince años; una colaboración que ha convertido a la Fundació en un referente español e internacional de la ética aplicada a la salud pública, dada la cantidad y relevancia de las aportaciones compartidas en los sucesivos encuentros anuales, que abarcan muchos de los temas y dilemas de interés en el área de la salud comunitaria y de las poblaciones, y que ha proporcionado a SESPAS y, en general, a los profesionales de la salud pública, un amplio inventario de materiales útiles para la aplicación de la perspectiva de la ética en la adopción de las decisiones y el desarrollo de las intervenciones pertinentes para la prevención y el control de los problemas colectivos de salud. Y todo ello sin olvidar que, además, ha contribuido a despertar el interés de la filosofía moral sobre la salud pública, como ilustra la reciente creación de ESPACyOS (Ética Salubrista para la Acción, Cuidados y Observación Social), una red interdis-

^a El encuentro sobre «Maleficencia en los programas de prevención» fue organizado conjuntamente por la Fundació y el Institut d'Estudis de la Salut del Departament de Sanitat del gobierno catalán, en el que estaba ubicado el programa de coordinación de las unidades docentes de Medicina Preventiva y Salud Pública, cuyo responsable fue posteriormente presidente de SESPAS, entidad que suscribió el acuerdo con la Fundació Víctor Grífols i Lucas para el desarrollo de estas actividades.

ciplinar, aunque eminentemente filosófica, que nació al calor de la pandemia por coronavirus en 2020 y de la que Àngel Puyol, y también Maite Cruz y David Casacuberta –que han participado en este último encuentro–, además de Txetxu Ausin –participante en otros encuentros– o incluso Javier Padilla –flamante secretario de Estado de Sanidad–, son destacados componentes.

En el presente Cuaderno recogemos las aportaciones del último encuentro dedicado a las consideraciones éticas que suscita la Inteligencia Artificial (IA) en el ámbito de la salud pública.

Sin duda, la IA está de moda. Aunque, en realidad, hace mucho tiempo que está entre nosotros. Un móvil o una simple base de datos informatizada ya usa IA. Por otra parte, todos recordamos las máquinas que ganan partidas de ajedrez a los mejores expertos del mundo, o la aparición, no sin polémica y sin problemas, de los trenes sin conductor y, sobre todo, de los coches autónomos. Sin embargo, en los últimos dos años ha habido una enorme popularización de la IA con la aparición de tecnologías como el ChatGPT, los asistentes virtuales de todo tipo o las cámaras inteligentes de reconocimiento facial.

Sin duda, la IA es muy beneficiosa, pero también provoca miedo, sobre todo por la posibilidad de que las máquinas usen su «inteligencia» de un modo que se escape a nuestro control, o incluso para rebelarse contra los humanos. La mayoría de nosotros mantiene en la memoria al ordenador Hall de la película *2001: Una odisea del espacio* que, en un momento dado, se hace con el mando de la nave espacial generando una inquietud que a nadie pasa desapercibida. Por otra parte, si las máquinas son inteligentes, enseguida surge el temor de que tal vez acaben siendo conscientes (ya están apareciendo pruebas científicas que apuntan a ello), con voluntad propia y quién sabe si con agencia moral. Estos avances tecnológicos de la IA y las oportunidades y los miedos que producen están muy presentes en el debate académico y social, y también en el político. Así, a finales de 2023 se ha aprobado en el Parlamento europeo la primera ley sobre IA.

La revolución tecnológica de la IA no solo no es ajena a la salud, sino que muchas de sus aplicaciones más llamativas tienen que ver precisamente con la salud, la sanidad y la salud pública. La IA está revolucionando la ciencia de los datos sanitarios, proponiendo diagnósticos y pronósticos más precisos, e

incluso interviniendo en la asistencia directa a los enfermos (pensemos, por ejemplo, en la robótica asistencial). Este nuevo escenario no está pasando inadvertido para las autoridades sanitarias, comenzando por la propia OMS, que hace dos años publicó una extensa guía sobre IA y salud, señalando los principales desafíos éticos y de gobernanza que esta debe afrontar; siempre, claro está, en torno al respeto a la integridad y la dignidad de las personas, las libertades individuales y la justicia social.

Los textos del seminario que recoge este nuevo Cuaderno pretenden contribuir a un mayor conocimiento entre nosotros de esos desafíos éticos de la IA en la salud pública. La digitalización masiva ha permitido incrementar sustancialmente la cantidad de datos relacionados con la salud y la enfermedad, tanto de la ciudadanía como de las sociedades humanas, incluido su entorno geográfico y los eventuales factores de protección, promoción y riesgo para la salud individual y colectiva. Mediante la IA, es factible elaborar información significativa acerca de los determinantes potencialmente relevantes de las actividades sanitarias protectoras, promotoras y restauradoras de la salud, tanto personal como comunitaria. Esperemos, pues, que las aportaciones del Cuaderno consigan aclararnos las cosas un poco más y que, a partir de ellas, logremos entender mejor los desafíos éticos de la IA en la salud pública, así como las inquietudes y preocupaciones éticas del amplio y diverso calado que suscita, desde las relativas a una probable mercantilización abusiva hasta las que pueden resultar de una orientación injusta a la hora de establecer prioridades en los ámbitos de las políticas públicas, de la salud y de la sanidad.

Àngel Puyol

Catedrático de Ética de la Universitat Autònoma de Barcelona

Andreu Segura

Coordinador del Grupo de Trabajo de Ética y Salud Pública de SESPAS

**Los retos éticos de la IA
en la salud pública**

Inteligencia artificial y salud pública

Francisco Javier García León

*Epidemiólogo. Asistente de Investigación ad honorem.
Departamento de Metafísica y Corrientes Actuales de la
Filosofía, Ética y Filosofía Política. Universidad de Sevilla*

El paradigma digital

En las últimas cuatro décadas hemos pasado de un mundo casi exclusivamente analógico a un mundo tremendamente digitalizado. Esta es la perspectiva que nos puede ayudar a entender la repercusión de la Inteligencia Artificial (IA) en la salud pública.

El paso a un mundo digitalizado ha ido acompañado de otros cambios de orden social y político, con gran impacto en la forma en la que se está desarrollando y aplicando la IA, así como el propio relato de este fenómeno. Destacan entre estos cambios la consolidación del neoliberalismo como teoría política y económica hegemónica que tiende a reducir al mínimo la intervención del Estado, la definición de una nueva era geológica, el Antropoceno, reflejo de nuestro impacto como especie sobre el planeta, los cambios demográficos y sociales profundos en un mundo globalizado y un desplazamiento en el plano geopolítico del poder hacia Asia.

Mientras tanto, la salud pública, al igual que el resto de la sociedad, ha experimentado un proceso de digitalización, entendido como proceso sociotécnico por el cual se han incorporado las tecnologías digitales a las operaciones y tareas previamente existentes, con el objetivo de mejorar la eficiencia y añadir valor a los usuarios, pero sin modificar sustancialmente los procesos productivos. Esta digitalización debe dar paso a un nivel superior todavía no alcanzado: el de la transformación digital. La transformación digital es un proceso disruptivo que implica una estrategia y un esfuerzo organizado para integrar de forma significativa las tecnologías, introduciendo nuevos procesos. Para llevarla a cabo se requiere la colaboración interdisciplinaria, con nuevos acto-

res, cambios en la cultura y el modelo de prestación de servicios y superando las barreras de los silos de información, tanto dentro de la propia salud pública como entre esta y los servicios sanitarios o extrasanitarios como medioambiente, servicios sociales, educación, etc.

Esta digitalización de la salud pública ha ido paralela a una pérdida de interés en sus enfoques comunitarios, y a una tendencia, tanto en la investigación como en la intervención, hacia los factores de riesgo desde la perspectiva individual; es posible que se esté produciendo una crisis de identidad en la disciplina.

En los servicios de salud llevan décadas planteándose estrategias que tienden a su transformación digital, pero no así en el ámbito de la salud pública; en ella, este tipo de iniciativas es más reciente. La primera es la Public Health England's digital strategy de 2017. Ha habido igualmente un debate acerca de si la salud pública digital es –o debería ser– algo distinto a la salud pública sin apellidos. Con la aparición del concepto de *medicina personalizada* en el ámbito clínico, se llegó a hablar de su aplicación a gran escala a nivel poblacional –especialmente en los aspectos preventivos personales–, lo que llevaría a la «salud pública personalizada», término que sufrió un amplio rechazo y ha dado paso a la llamada «salud pública de precisión».

La COVID-19 ha supuesto un avance notable en la digitalización de la salud, con la mayor disponibilidad de datos administrativos y clínicos, la incorporación de datos no sanitarios, el uso de datos de telefonía móvil y redes sociales, el rastreo de contactos y control de los vacunados mediante desarrollos tecnológicos, y las mejoras en la comunicación e integración de la información en los informes de seguimiento –sin que eso haya supuesto realmente una integración de los sistemas de información.

Durante la pandemia se desarrollaron sistemas de IA relacionados con la atención a los pacientes (cribados, diagnóstico por imagen, establecimiento de pronósticos, evaluación de tratamientos, etc.). Igualmente, ha habido avances en la salud pública para la identificación de brotes (con datos de rastreo de contactos, historias clínicas electrónicas [HCE] o tráfico aéreo, entre otros), el cumplimiento de normas de uso de mascarillas mediante reconocimiento facial, medidas contra la desinformación en redes sociales,

predicciones en la evolución de las incidencias, o farmacovigilancia, entre otros. La IA no ha tenido un papel relevante en el control de la pandemia, pero está claro que en el futuro su papel será cada vez mayor.

Las inversiones del Plan de Resiliencia son una oportunidad de avance, si bien carecemos de una estrategia digital que les dé coherencia, y posiblemente el foco esté excesivamente centrado en la vigilancia, descuidando otros ámbitos de la salud pública cuyo desarrollo ha de ser armónico.

¿Inteligencia y Artificial?

En los últimos años, la IA ha tomado un gran protagonismo. Para ello han tenido que confluír factores como la disponibilidad de grandes volúmenes de datos digitalizados (*big data*) y la mayor capacidad de cálculo computacional; esto ha posibilitado probar algoritmos ya existentes pero con limitaciones tecnológicas para su cálculo y ha estimulado el desarrollo de otros nuevos.

A pesar del uso masivo del término IA, no hay una única definición de la misma. Sin necesidad de remontarnos a la máquina pensante de Ramon Llull en el siglo XIII, se considera que el término fue utilizado por primera vez en 1955, en la solicitud de financiación del Dartmouth Summer Research Project on Artificial Intelligence celebrado en verano de 1956. En ella se indicaba que:

«...cada aspecto del aprendizaje o cualquier otra característica de la inteligencia puede, en principio, describirse con tanta precisión que se puede lograr que una máquina lo simule. Se intentará averiguar cómo hacer que las máquinas utilicen el lenguaje, formen abstracciones y conceptos, resuelvan tipos de problemas ahora reservados a los humanos y se mejoren a sí mismas», y

«...el desafío de la inteligencia artificial es hacer que una máquina se comporte de manera que se consideraría inteligente si un ser humano se comportara así».¹

Esta otra definición más concreta podría ser representativa de las definiciones de los años más recientes:

«...la capacidad de un sistema para interpretar correctamente datos externos, aprender de esos datos y utilizar esos aprendizajes para lograr objetivos y tareas específicas mediante una adaptación flexible».²

Por último, esta es la definición normativa incluida en la propuesta de Ley de IA de la UE:³

«...software que se desarrolla empleando una o varias técnicas y estrategias (...) y que puede, para un conjunto determinado de objetivos definidos por seres humanos, generar información de salida como contenidos, predicciones, recomendaciones o decisiones que influyan en los entornos con los que interactúa».

Entre las técnicas y estrategias que incluye se encuentran: estrategias de aprendizaje automático, estrategias basadas en la lógica y el conocimiento y estrategias estadísticas.

A pesar del gran número de definiciones sobre IA y la diversidad de enfoques,⁴ se observan elementos comunes entre todas ellas, que sintetizan las principales capacidades de la IA:

- Percepción del entorno, teniendo en cuenta la complejidad del mundo real.
- Procesamiento de información, recopilando e interpretando los datos de entrada.
- Toma de decisiones (incluyendo razonamiento y aprendizaje), ejecución de acciones y realización de tareas (incluyendo la adaptación y reacción a los cambios que se producen en su entorno), con cierto nivel de autonomía.
- Cumplimiento de objetivos específicos, que es la razón última de los sistemas de inteligencia artificial.

Se especula entre lo que la IA hace, lo que podrá hacer a corto y medio plazo y lo que podría lograr en el futuro. Para delimitar su alcance, evitar confusiones y centrar el debate, se ha diferenciado entre la IA débil o estrecha y la IA fuerte o general. La IA existente en la actualidad es la IA débil, que opera dentro de un rango limitado de tareas previamente definidas; en este tipo de IA no hay ninguna inteligencia genuina, ni conciencia, ni vida.

La IA fuerte es, en teoría, un tipo de sistema automático que iguala o excede a la inteligencia humana promedio; la inteligencia de una máquina que puede realizar con éxito cualquier tarea intelectual propia de un ser humano. Es un

objetivo importante para la investigación sobre la IA y un tema interesante para la ciencia ficción y el debate filosófico.

Debido a la diversidad conceptual y a la multitud de dimensiones de la IA, se han desarrollado varias taxonomías para fines específicos, construidas con criterios tales como las familias de algoritmos utilizadas, las capacidades cognitivas desarrolladas, las funciones que realiza o sus usos en los distintos sectores económicos y sociales.

Dada la especificidad de las taxonomías anteriores, se ha propuesto una clasificación amplia que sea de utilidad desde distintas perspectivas: política, investigadora e industrial. Esta taxonomía establece unos dominios nucleares (razonamiento, planificación, aprendizaje, comunicación y percepción) y otros transversales (integración e interacción, servicios y ética y filosofía).

Dominios nucleares

El dominio del razonamiento aborda la forma en que las máquinas transforman los datos en conocimiento, o infieren hechos a partir de los datos. Un ejemplo claro sería la búsqueda de patrones en un lago de datos, mostrándonos relaciones *a priori* desconocidas. Entre sus subdominios encontraríamos la representación del conocimiento, el razonamiento automatizado y el razonamiento de sentido común.

El dominio de la planificación permite el diseño y la ejecución de estrategias para llevar a cabo una actividad. Sus subdominios incluyen planificación y programación, búsqueda de alternativas y optimización de tareas.

El dominio del aprendizaje se refiere a la capacidad de los sistemas para aprender, decidir, predecir, adaptarse y reaccionar automáticamente a los cambios, mejorando a partir de la experiencia, sin ser programados explícitamente. Incluye el aprendizaje automático en sus distintas modalidades.

El dominio de la comunicación se refiere al procesamiento del lenguaje natural, es decir, la capacidad de la máquina para identificar, procesar, comprender y generar información en las comunicaciones humanas escritas y habla-

das. Engloba aplicaciones como la generación de textos, la minería de textos, la clasificación y la traducción automática.

El dominio de la percepción se refiere a la capacidad de los sistemas para darse cuenta de su entorno a través de los sentidos: visión, oído, manipulación táctil, etc., siendo la visión y el oído las áreas más desarrolladas en la IA. Incluye también los subdominios de visión artificial y tratamiento de audio.

Dominios transversales

La integración e interacción aborda la combinación de los dominios de percepción, razonamiento, acción, aprendizaje e interacción con el entorno, así como el desarrollo de características como la orientación, coordinación, cooperación, autonomía, interacción e integración. Incluye asimismo sistemas multiagente, robótica y automatización y vehículos conectados o autónomos.

El dominio de servicios de la IA se refiere a cualquier infraestructura, *software* y plataformas, generalmente ubicados en la nube, que proporcionan servicios o aplicaciones que se ejecutan bajo demanda, reduciendo la gestión de infraestructuras complejas.

El desarrollo y aplicación de la IA suscita debates éticos; el significado de la IA actualmente y en el futuro es motivo de interés filosófico, especialmente en lo referente a la IA general.

Para desmitificar⁵ relatos más próximos al mundo de la ciencia ficción y centrar el debate en los elementos que nos permitan comprender estos sistemas, es importante tener en cuenta que la IA no busca explicar cómo funciona el cerebro, no intenta operar de manera similar al mismo, ni induce al pensamiento en los ordenadores. La IA no tiene una finalidad diferente a la estadística o al aprendizaje automático, con los cuales comparte su objetivo de análisis de datos; lo que diferencia a unos y otra es la complejidad de los datos utilizados, la complejidad de la pregunta de investigación y el control que tenemos sobre el proceso. A medida que aumenta la complejidad en estas

dimensiones, disminuye nuestra capacidad para interpretar la respuesta del sistema.

La IA no es una tecnología en sí misma, sino una metodología basada en algoritmos de aprendizaje, que, para su ejecución, necesita elementos tecnológicos como los ordenadores; sin embargo, el componente intelectual de la IA no es el ordenador en sí, sino el *software* que aplica las reglas de aprendizaje. La propuesta de Ley de la IA en la UE combina estos elementos y se refiere a «sistemas de inteligencia artificial» mientras que, en informática, la investigación sobre la IA se define como el estudio de los «agentes inteligentes» (dispositivos que perciben su entorno y emprenden acciones para alcanzar sus objetivos).

Se ha especulado sobre cuándo la IA alcanzará su objetivo final de igualar o superar a los humanos. Hay tareas específicas, como el juego de ajedrez o el reconocimiento de imágenes, en las que la IA ya obtiene mejores resultados que los humanos; sin embargo, la inteligencia general, que abarca una amplia gama de habilidades cognitivas, es un desafío más complejo y no hay consenso acerca de si se podrá alcanzar.

En la década de 1960, hubo grandes expectativas sobre la IA y se realizaron predicciones que sugerían que en dos décadas las máquinas serían capaces de hacer cualquier trabajo realizado por humanos, o que en una generación el problema de crear inteligencia artificial estaría sustancialmente resuelto. A pesar del entusiasmo inicial, las expectativas no se cumplieron y a principios de la década de 1970 decayó el interés por la IA, hasta que ha tomado nuevo impulso en la última década, y especialmente en los últimos cinco años, favorecido por el hecho de disponer de grandes volúmenes de datos y de una capacidad computacional sin precedentes. Tras la irrupción de ChatGPT, la IA generativa se encuentra en su momento de máxima expectativa, estando previsto que, como cualquier tecnología emergente, estas expectativas decaigan hasta unos niveles más adecuados a la realidad, y se espera que en un plazo de dos a cinco años esté incorporada en la rutina de las organizaciones.⁶

Por otra parte, parece que a medida que vamos avanzando en el campo de la IA van surgiendo más dudas acerca de la idoneidad de su denominación.

Inteligencia artificial y salud pública

La aplicación más destacada de los sistemas de IA en salud pública se observa en el campo de la vigilancia epidemiológica y los chatbots de promoción de la salud.

En el campo de la vigilancia epidemiológica, los desarrollos más avanzados se encuentran en la Public Health Agency of Canada, los Centers for Disease Control and Prevention (CDC) de EE. UU. y en China. Generalmente utilizan datos de HCE, incluidos los datos de laboratorio, datos específicos de registros de enfermedades y redes sociales. También es destacable la iniciativa privada, como BlueDot, que proporciona informes epidemiológicos a sus clientes mediante suscripción.

Respecto a la promoción de la salud, los mayores avances se están realizando desde el campo tecnológico, desarrollando herramientas que buscan modificar o reforzar conductas individuales de promoción de la salud o la eliminación de riesgos individuales. Los ámbitos en los que han mostrado mayor interés son los del ejercicio físico, alimentación, consumo de tabaco y salud mental. La evaluación de la efectividad de estos dispositivos en términos generales no es satisfactoria, debido a la comunicación insuficiente de la metodología utilizada y a las dificultades para la reproducibilidad de los resultados.

No es sorprendente que la promoción de la salud esté polarizada en este tipo de productos, dado que la propia investigación en salud pública lleva décadas centrada en la identificación de factores de riesgo individuales, agravada con la llamada medicina y salud pública personalizadas o de precisión.

En cuanto a la medicina comunitaria, se observa que las aplicaciones más habituales de la IA son herramientas de ayuda al diagnóstico clínico, la detección precoz de enfermedades y la vigilancia epidemiológica.

Hay ámbitos en los que la IA podría ser de gran ayuda en salud pública –como el análisis del estado de salud de la población, la exploración de estrategias de intervención y la evaluación de las propias intervenciones en salud pública–, en los que, de momento, no se observa interés. Disponemos de sistemas de IA que podrían ser muy útiles para avanzar en los determinantes sociales de la

salud a partir de la información estructurada y no estructurada de la HCE; sin embargo, estos determinantes no suelen ser considerados en la investigación sobre la IA a partir de los datos de la HCE.

En cuanto a sanidad ambiental, hay experiencias en el uso de la IA en el estudio de abastecimientos de agua e identificación de tóxicos y otros contaminantes, así como modelos de simulación de transmisión de legionela y previsiones en la aparición de picos de contaminación ambiental. Se echan en falta iniciativas sobre intervenciones comunitarias para lograr entornos más saludables, con la inclusión de la perspectiva social. En lo que respecta a la protección de la salud, hay experiencias de intervenciones en lugares de restauración en relación con las toxiinfecciones alimentarias.

Se están haciendo valoraciones acerca del impacto que la IA pueda tener en la consecución de los Objetivos del Desarrollo Sostenible. En términos generales, podría contribuir a alcanzar 134 objetivos y dificultar la consecución de otros 59. Sin embargo, existe preocupación por su posible impacto en el aumento de las desigualdades entre países y personas, la concentración de poder en unas pocas empresas tecnológicas, con un progresivo descenso del papel de los centros académicos y gubernamentales, y el impacto medioambiental debido a su alto consumo de energía. Para evitar estos riesgos, será necesario un marco de gobernanza justo, así como el desarrollo normativo adecuado.

La iniciativa One Health podría beneficiarse con la incorporación de la IA, para intentar resolver problemas complejos de interacción entre microorganismos, animales, medioambiente y humanos. Se destaca la importancia de la genómica en este campo. Sin embargo, estos posibles avances encuentran grandes dificultades debido a la dispersión, diversidad y limitaciones en la accesibilidad a los datos.

En cuanto a la regulación de los sistemas de IA, la referencia de mayor trascendencia en la actualidad es la propuesta del Reglamento de la Ley de IA en Europa. Tiene por finalidad garantizar que los sistemas de IA comercializados y utilizados en el mercado de la UE sean seguros y respeten la legislación vigente sobre derechos fundamentales y valores de la UE, garantizar la seguridad jurídica para facilitar la inversión y la innovación, mejorar la gobernan-

za y la aplicación efectiva de la legislación, y facilitar el desarrollo de un mercado único para la IA.

Esta propuesta de reglamento indica los sistemas de IA que tienen prohibido su uso y aquellos que son considerados de alto riesgo en función de su finalidad prevista, en los cuales se precisa una evaluación de la conformidad con anterioridad a su uso. El resto de los sistemas no tiene requerimientos previos y se somete a la autorregulación de los fabricantes.

En relación con la salud, se consideran de alto riesgo los sistemas de IA destinados a enviar o establecer prioridad en el envío de servicios de primera respuesta de emergencia, incluidos los bomberos y la asistencia médica, así como el *software* que se utiliza para diagnosticar pacientes, tomar decisiones terapéuticas y monitorear procesos fisiológicos o usados con fines anticonceptivos (son los dispositivos médicos clasificados como IA o superiores en el Reglamento (UE) 2017/745 sobre los productos sanitarios).

En consecuencia, los fabricantes de sistemas de IA tendrán que seguir un código de buenas prácticas, pero no estarán obligados a realizar una evaluación previa en aquellos sistemas cuya finalidad abarque los distintos ámbitos de la salud pública, la planificación sanitaria o la distribución de recursos sanitarios. Debido a su impacto poblacional, los daños causados por estos sistemas pueden ser cuantiosos.

Sería recomendable que la propuesta incluyera, entre los sistemas de alto riesgo, aquellos cuyo resultado pueda tener efectos negativos en la salud de las personas o de las poblaciones, así como en la equidad sanitaria, o que se utilicen en la gestión de riesgos poblacionales, la planificación, gestión del personal, la ruta y clasificación de la atención sanitaria, así como el descubrimiento de medicamentos, ensayos clínicos y promoción farmacéutica, entre otros.⁷

La IA tiene el potencial de mejorar la capacidad de la salud pública para el ejercicio de sus funciones, pero para que este potencial se desarrolle es preciso que los servicios de salud pública tengan las estrategias adecuadas. Buena parte de estas estrategias no son específicas de la IA, pero son requisitos previos no resueltos, relacionados con la insuficiente estrategia digital en salud pública y el debilitamiento de estos servicios.

Se han identificado seis líneas prioritarias de actuación para que las organizaciones de salud pública puedan incorporar con éxito los sistemas de IA:⁸

1. Gobernanza de datos actualizada

Desarrollar mecanismos propios de gobernanza de datos con apoyo de expertos, teniendo en cuenta los marcos de referencia de la Administración pública y de los servicios de salud. En este apartado se incluyen los aspectos relacionados con normas, procedimientos, roles y responsabilidades, procesos, calidad, seguridad, interoperabilidad, metadatos o privacidad, entre otros.

Deberá tenerse en cuenta el marco de referencia establecido por la OMS sobre ética y gobernanza de la IA: 1) proteger la autonomía, 2) promover el bienestar humano, la seguridad de las personas y el interés público, 3) garantizar la transparencia, la explicabilidad y la inteligibilidad, 4) fomentar la responsabilidad y la rendición de cuentas, 5) garantizar la inclusión y la equidad, y 6) promover una inteligencia artificial responsable y sostenible.

2. Inversión en infraestructura, datos y sus procedimientos analíticos

La IA requiere grandes volúmenes de datos para cuya gestión y análisis muchas organizaciones de salud pública no cuentan con la infraestructura computacional necesaria. Será preciso tomar decisiones acerca de contar con centros de datos locales o tender hacia centros compartidos o almacenamiento en la nube.

En cuanto a los datos, será necesario realizar esfuerzos en su calidad y oportunidad, así como conseguir acuerdos de uso en el acceso a los gestionados por otros organismos públicos, empresas y entidades no gubernamentales. Será preciso desarrollar plataformas de datos únicas que eviten los silos de datos actuales, tan dificultosos a la hora de compartir los datos tanto dentro de los propios servicios de salud pública como entre estos, los servicios de salud y otros organismos gubernamentales, especialmente los competentes en servicios sociales, educación, medioambiente o agricultura.

Respecto a la capacidad analítica, será preciso dar paso a las herramientas de cálculo más apropiadas para grandes volúmenes de datos, buscando en cada caso la solución más adecuada, ya sea de desarrollo propio o como solución de mercado.

3. Abordar la brecha de competencias profesionales

Para utilizar adecuadamente grandes volúmenes de datos, métodos analíticos avanzados y sistemas de IA son necesarios profesionales con competencias adecuadas en gestión de datos, estadística, informática, privacidad y ética, entre otras. Por tanto, es preciso identificar las competencias sobre las que la organización debe mejorar. Se requiere tanto el desarrollo de actividades propias como la participación en otras iniciativas gubernamentales, facilitando la rotación por otras unidades para adquirir competencias. En términos generales, los profesionales de salud pública no han recibido la formación necesaria para desarrollar, evaluar o aplicar tecnologías basadas en la IA.

Para los servicios de salud pública, al igual que para los organismos públicos en general, resulta difícil incorporar profesionales de la IA; estos se encuentran mayoritariamente en el sector privado, por lo que es preciso tener iniciativas para atraer y conservar estos perfiles profesionales. El interés que los datos de salud tienen para desarrollar nuevos proyectos, junto con becas, ayudas a la investigación y facilitación de doctorados, puede ayudar a atraer estos nuevos profesionales.

4. Desarrollo de asociaciones estratégicas de colaboración

Es necesaria la colaboración con otras instituciones públicas para adquirir conocimiento especializado, obtener o compartir el acceso a datos e infraestructuras y ampliar las perspectivas; igualmente, es necesaria la colaboración con científicos e institutos de investigación en materia de ciencia de datos e inteligencia artificial. Por último, es preciso establecer unas reglas justas y transparentes en la colaboración pública-privada, dada la asimetría de conocimiento entre los organismos gubernamentales y las grandes corporaciones tecnológicas.

Los profesionales de salud pública poseen competencias en el análisis de los sesgos, la generalización de resultados y la equidad. Por tanto, pueden colaborar eficazmente con los expertos en informática e IA para desarrollar estrategias de mitigación de estos problemas.

5. Uso de buenas prácticas para garantizar la transparencia y la reproducibilidad

La transparencia es uno de los principios éticos más comunes en el ámbito de la IA, con el objetivo de fomentar la confianza y evitar posibles daños. Este principio implica esfuerzos para mejorar la explicabilidad e interpretabilidad de la IA, lo que a su vez requiere una divulgación detallada sobre cómo se desarrolla un modelo o tecnología de IA, cómo funciona, qué datos usa, cómo se despliega y utiliza, y un debate sobre sus limitaciones. Además, la transparencia fomenta la libertad y autonomía.

Aunque existen normas para la comunicación adecuada de los resultados de la investigación con IA, nos enfrentamos comúnmente a problemas de reproducibilidad de los resultados. A pesar de las posibles limitaciones por los derechos de propiedad intelectual, es exigible un determinado nivel de transparencia. Los principios de transparencia y ciencia abierta deben equilibrarse cuidadosamente con la privacidad y la confidencialidad.

6. Incorporación explícita de consideraciones de equidad y justicia

En todas las fases del desarrollo y la aplicación de la IA es crucial tener en cuenta consideraciones éticas. Esto abarca desde la selección del tema de estudio, la recopilación de datos, el desarrollo e implantación del modelo, e incluso las etapas posteriores. La preocupación por evitar sesgos debe incluir tanto el origen y calidad de los datos como su validación en subgrupos de población según su condición social, etnia, sexo o ámbito geográfico.

Para que el uso de la IA sea ético, es fundamental que la población esté informada sobre cómo se utilizan sus datos y cómo los sistemas de IA pueden influir en su vida, y que tenga un espacio para expresar sus preferencias y preocupaciones. La colaboración en paneles deliberativos o experiencias de ciencia ciudadana pueden ser de ayuda. La diversidad y multidisciplinariedad de los equipos han mostrado ser de gran utilidad para prevenir sesgos y discriminación.

Conclusión

La IA puede contribuir de manera significativa al desempeño de las funciones esenciales de la salud pública. Para lograrlo, es imperativo que estas organizaciones adopten lo antes posible una estrategia adecuada. Entre las acciones prioritarias se destaca el desarrollo de una estrategia de transformación digital en salud pública, hasta ahora no abordada, y la revisión del rol mismo de la salud pública, en la que hay una tendencia a poner el foco en la perspectiva individual respecto a la poblacional.

Notas

1. McCarthy J, Minsky M, Rochester N, Shannon C. A Proposal for the Dartmouth Summer Research Project on Artificial Intelligence; 1955.
2. Kaplan A, Haenlein M. Siri, Siri, in my hand: Who's the fairest in the land? On the interpretations, illustrations, and implications of artificial intelligence. *Bus Horiz.* 2019;62(1):15-25.
3. Bruselas. Propuesta de Reglamento del Parlamento Europeo y del Consejo por el que se establecen normas armonizadas en materia de inteligencia artificial (Ley de Inteligencia Artificial) y se modifican determinados actos legislativos de la Unión. COM(2021) 206 final. 21 de abril de 2021.
4. Samoili S, López Cobo M, Delipetrev B, Martínez-Plumed F, Gómez E, De Prato G. AI Watch. Defining Artificial Intelligence 2.0. Towards an operational definition and taxonomy for the AI landscape, EUR 30873 EN, Publications Office of the European Union, Luxemburgo; 2021.
5. Emmert-Streib F, Yli-Harja O, Dehmer M. Artificial Intelligence: A Clarification of Misconceptions, Myths and Desired Status. *Front Artif Intell.* 2020; 3: 524339. Disponible en: doi: 10.3389/frai.2020.524339.
6. Gartner Inc, Jaffri A. Hype Cycle for Artificial Intelligence, 2023. ID G00791179. Publicado: 19 de julio de 2023. Stamford, CT 06902, EUA.

7. van Oirschot J, Ooms G. Interpreting the EU Artificial Intelligence Act for the Health Sector. Health Action International. Febrero de 2022. Amsterdam, Países Bajos.
8. Fisher S, Rosella LC. Priorities for successful use of artificial intelligence by public health organizations: a literature review. BMC Public Health. 2022; 22(1):1-14.

Desafíos éticos de la IA

David Casacuberta

Departamento de Filosofía de la Universitat Autònoma de Barcelona

I. Introducción. Entender la IA

La Inteligencia Artificial (IA) se está convirtiendo en una herramienta esencial en el ámbito de la salud pública, ofreciendo posibilidades sin precedentes para el diagnóstico, tratamiento y prevención de enfermedades. En particular, la así llamada IA generativa está remodelando todo lo que es posible en el campo de la salud. Estos sistemas pueden procesar y analizar grandes cantidades de datos a una velocidad y precisión que superan a las capacidades humanas. Sin embargo, el rápido avance de esta tecnología también plantea importantes desafíos éticos que deben ser abordados para garantizar su uso responsable y justo.

A la hora de clarificar estos retos es importante entender con cierto detalle qué es en última instancia la IA y cómo funcionan los algoritmos sobre los que se desarrollan estas tecnologías. En particular los medios de comunicación han generado una serie de mitos que están enraizando en el público general, mezclando de forma caótica máquinas benevolentes casi divinas con monstruos fuera de control dispuestos a esclavizarnos, o directamente a exterminarnos. Por ello dedicaremos esta introducción a despejar algunos de esos mitos y exageraciones.

En primer lugar, cabe clarificar que estamos muy lejos de la así llamada inteligencia artificial general (AGI son sus siglas en inglés). Una inteligencia artificial general sería un tipo de programa que tendría las mismas capacidades intelectuales que un humano, más la velocidad de procesamiento y las capacidades de almacenamiento de una computadora, de manera que rápidamente superaría a los humanos en todos los ámbitos intelectuales. Así,

una AGI podría procesar cualquier tipo de problemas, entendiendo el contexto en el que se presentan, y podría exponer las mejores soluciones a esos problemas. Una AGI, pues, podría leer un problema avanzado de física cuántica y resolver las ecuaciones de manera casi instantánea, para después pasar a inventar una receta de fusión que combine la tradición china con la cocina de Ferran Adrià, recomendar novelas feministas de corte histórico con un toque de misterio que tengan lugar en el siglo XIX y, si no encuentra ninguna, escribirla ella misma. Y, evidentemente, si le describimos una serie de síntomas, la AGI podría decirnos con exactitud cuál es nuestra enfermedad e incluso informarnos de cuáles son los mejores medicamentos para tratarla.

Buena parte de los conversos a la revolución de la IA creen que los proyectos de inteligencia artificial generativa como ChatGPT están consiguiendo precisamente eso: una inteligencia artificial general que puede tratar todo tipo de temas con la soltura de un humano. Desgraciadamente –o afortunadamente– estamos todavía muy lejos de disponer de semejantes tecnologías. Si usted ha jugado un poco con el ChatGPT, Bard o equivalentes, habrá notado que estos programas no se caracterizan tanto por su capacidad de crear textos verídicos, sino por la de ofrecer textos coherentes que *suenan* verídicos.

Ello es un resultado directo de la forma en que funcionan los algoritmos de la IA generativa. En resumen, muchos de estos sistemas tienen enormes mapas conceptuales que establecen las probabilidades con las que una palabra puede seguir a otra, en función del contexto anterior. Imagine que estamos en medio de un proceso de generación de textos mediante IA generativa. Empezamos una nueva frase, y de momento solo tenemos el artículo «La». El programa tiene en su enorme base de datos el género marcado de cada palabra del castellano que forma parte de la base de datos, de manera que es muy improbable que una palabra de género masculino o neutro aparezca detrás de ese «la». Imaginemos ahora que el contexto sobre el que está generando frases la IA generativa es sobre un viaje en tren. Ello hará que palabras como «estación» o «ciudad» sean mucho más probables que «páncreas» o «esotérico». Si finalmente añadimos «estación» y tenemos la frase «la estación», entonces una expresión como «de tren» tendrá muchas más probabili-

dades de aparecer que «de radio», ya que el contexto es un viaje en tren. Así y todo, «de radio» o «primaveral» tienen más probabilidades de aparecer que «elástica», pues en la base de datos de la AGI hay muchos más ejemplos de la frase «la estación de radio» que de «la estación primaveral» –aunque alguno de esta última habrá–, pero es muy improbable que haya ni un solo ejemplo de «la estación elástica».

Ello significa, simplemente, que cada palabra surge como la más probable dado un contexto, pero que no hay ninguna búsqueda de veracidad o exactitud al respecto. Así, por ejemplo, acabo de preguntarle a Bard –la IA generativa de Google– por una enfermedad ficticia, la *supralisianosis atípica*, que me acabo de inventar. Al preguntarle por cuáles son sus síntomas me ha contestado lo siguiente:^a

La supralisianosis atípica (MSA) es un trastorno neurodegenerativo progresivo que afecta el movimiento, la cognición y la función autonómica. Los síntomas principales de la MSA son:

Síntomas motores:

- Acinesia: disminución de la velocidad y la amplitud de los movimientos.
- Rigidez: aumento del tono muscular que dificulta el movimiento.
- Temblor: movimientos involuntarios rítmicos de las manos, los brazos, las piernas o la cabeza.
- Postura inestable: dificultad para mantener el equilibrio y la coordinación.
- Caídas: frecuentes caídas debido a la inestabilidad postural y la dificultad para caminar.

^a Este intercambio tenía lugar el 30 de enero de 2024. Seguramente este tipo de resultados ya no serán posibles en un futuro cercano. De hecho, para ser justos, comentar que ChatGPT no cometió el mismo error y me informó de que esta enfermedad no está en ninguna base de datos médica y probablemente sea una enfermedad inventada. Ello no significa que ChatGPT sea más «inteligente» que Bard. Simplemente es señal de que probablemente para evitar demandas futuras, programadores humanos han añadido controles específicos sobre temas problemáticos para los humanos, como puede ser las enfermedades.

El texto sigue, de hecho, y hasta se inventa dos tipos de supralisianosis:

La MSA se divide en dos tipos principales:

- MSA-P: predominio de parkinsonismo. Los síntomas motores son más prominentes que los síntomas cognitivos.
- MSA-C: predominio cerebeloso. Los síntomas cerebelosos, como la ataxia y la disartria, son más prominentes que los síntomas parkinsonianos.

Si le vuelvo a preguntar, aparecerá otra descripción diferente de la enfermedad, con otros síntomas. Fíjense, no obstante, en cómo sigue el esquema de lo que sería un diagnóstico médico. Hasta llega a inventarse unas siglas para referirse a la enfermedad.

En la jerga informática, este tipo de comportamiento se conoce como «alucinar». A mí me divierte más llamarlo «efecto *cuñao*», porque, en el fondo, la IA generativa se comporta como los cuñados de los chistes y series de humor: dicen que saben –y opinan– de cualquier tema, aunque en realidad no sepan prácticamente nada de la mayoría.

Pero las IA no son cuñados. No son personas. No tienen conciencia, ni personalidad. No aprenden como lo hacemos nosotros. Ni son creativas a la manera en que lo somos los humanos. Este tipo de metáforas tan extendidas por los medios de comunicación pueden ser útiles hasta cierto punto, pero si las tomamos demasiado en serio generan confusiones importantes.

Lo cierto es que, en realidad, el tipo de IAs con las que se están desarrollando actualmente la mayoría de aplicaciones se basan en el concepto de aprendizaje automático. Es decir, sistemas que reciben una gran cantidad de datos etiquetados por humanos, que son convertidos en números y algoritmos de «aprendizaje», y que luego buscan similitudes estadísticas entre esos datos, para así poder hacer predicciones y clasificaciones.

Así, no deberíamos decir que el algoritmo que nos permite desbloquear nuestro *smartphone* al acercar nuestro rostro a la cámara ha aprendido a reconocer nuestra cara. Lo que hace es inferir que, entre diferentes imágenes mías, con ángulos e iluminación diferentes, hay una serie de píxeles que son suficientemente estables entre foto y foto como para establecer la unicidad de esa imagen. Pero el algoritmo no reconoce nuestro color de ojos, la forma de nuestra nariz o esa colección de pecas distribuida sobre nuestras mejillas.

Simplemente establece que hay una concordancia estadística suficientemente amplia entre los píxeles de la imagen captada por la cámara y aquella fotografía que nos tomamos al principio para poder desbloquear el móvil. Pero fíjense que hay todo un abismo entre reconocer realmente una cara e inferir simplemente que dos colecciones de números se parecen lo suficiente.

Nuestra inteligencia natural –la única que merece realmente el calificativo de «inteligente» ahora mismo– entiende las cosas y sabe qué datos son relevantes para una tarea y cuáles no. Un algoritmo de aprendizaje automático no sabe nada de eso y lo más probable es que el tipo de regularidad que encuentre no tenga nada que ver con lo que un humano consideraría relevante... a no ser que haya, precisamente, una revisión humana de todo el proceso.

Consideremos el siguiente ejemplo de una aplicación real: se enseñó a un algoritmo de aprendizaje automático a distinguir los lobos de los huskies. Para hacerlo se le dieron una serie de fotos a modo de ejemplo, etiquetadas como «lobo» o como «husky» para que aprendiera. Una vez comprobado que el algoritmo funcionaba de forma adecuada, se contrastó la capacidad real de distinguir lobos de huskies probando con fotos nuevas. El programa lo hacía bastante bien, pero cometía algunos errores importantes (figura 1).

Figura 1. Explain the Prediction



Fuente: Besse, Philippe & Castets-Renard, Céline & Garivier, Aurélien & Loubes, Jean-Michel (2018). Can Everyday AI be Ethical? Machine Learning Algorithm Fairness (english version). 10.13140/RG.2.2.22973.31207.

Como puede verse, en la segunda fila hay dos errores interesantes: en el segundo fotograma aparece un lobo etiquetado como «husky» y en la cuarta foto, un husky ha sido erróneamente etiquetado como «lobo».

Después de una serie de test técnicos, los investigadores se dieron cuenta de lo que pasaba: la inmensa mayoría de las fotos de lobos se tomaron en parajes montañosos en invierno, de manera que prácticamente todos los lobos aparecían rodeados de nieve. En cambio, las fotos de huskies se habían tomado en un entorno urbano y sin nieve. Si se fijan, verán que el lobo etiquetado como husky no tiene ni una mancha de nieve en la imagen, mientras que el husky erróneamente calificado de lobo sí tiene un paisaje nevado de fondo.

Lo que hizo el algoritmo de aprendizaje automático, en definitiva, fue asociar con «lobo» una buena cantidad de píxeles blancos en la foto, y con «husky» la ausencia de tales píxeles.

Resumiendo, el mundo de la IA está lleno de términos que no se adecuan a la realidad. La IA no es realmente inteligente en el sentido en que lo usamos los humanos. Y el adjetivo «automático» con el que caracterizamos los algoritmos responsables de buena parte de la IA –los algoritmos de aprendizaje automático– se debería revisar también, pues da esa falsa imagen de que el algoritmo «aprende solo» cuando la realidad es que ha de haber un complejo proceso previo de etiquetado desarrollado por humanos para que el proceso acabe generando una aplicación realmente útil.

Esta dificultad de establecer qué es lo que un algoritmo de IA debe considerar saliente a la hora de establecer una predicción o llevar a cabo una clasificación está, en mi opinión, en la base de buena parte de retos éticos que la IA presenta para la medicina. En la siguiente sección desarrollo esta propuesta.

II. Desafíos éticos de la IA en medicina

Los desafíos éticos directos de la IA giran en torno a la coherencia y la verdad en la información generada por la IA, así como la necesidad de explicabilidad. Estos aspectos son fundamentales para asegurar que las decisiones tomadas con el apoyo de la IA sean correctas, justas y transparentes.

Hablo aquí de riesgos directos para distinguirlos de otros tipos de riesgos éticos, que surgen de manera indirecta en el proceso de creación o difusión de los algoritmos de IA. Así, la erosión a nuestra privacidad que deriva de una mala gestión de las bases de datos con las que se ha entrenado un algoritmo de IA no es un resultado directo de aplicar el algoritmo en un contexto médico, sino que surge de manera indirecta al no tomar las precauciones adecuadas para garantizar la protección de datos personales.

A continuación exploraremos los dos problemas más centrales que resultan directamente del uso de la IA en un entorno médico: el error en la predicción y los sesgos por falta de representación.

Error en la predicción

El problema ético más claro a la hora de utilizar la IA en un entorno médico es el del error en la predicción. ¿Cómo podemos estar seguros de que el algoritmo está haciendo una predicción o una clasificación correcta cuando analiza los datos médicos de un paciente? Observen que este problema no es equivalente a un error médico. El clásico «los médicos también se equivocan» de los *cuñaos* pro-inteligencia artificial no es suficiente para resolver este problema ético. Ciertamente, el personal médico puede equivocarse en un diagnóstico, pero ese error está confinado a una comprensión del cuerpo humano y de las normas de la biomedicina, y no al albur de elementos externos, como el de un algoritmo real de aprendizaje automático entrenado para detectar melanomas a partir de fotografías de manchas benignas y malignas y que decidió que el indicador más estable de la existencia de un melanoma era la presencia de una cinta métrica en la fotografía.

Cuando el personal médico sospecha que una mancha en la piel es candidata a ser un melanoma, muchas veces toma una fotografía con una cinta métrica o una regla al lado para medirla en ese momento preciso y así, pasado un tiempo, comprobar si ha crecido o sigue con el mismo tamaño. Ello significa que una parte estadísticamente relevante de las fotografías de melanomas tendrán una cinta métrica al lado. Un algoritmo de aprendizaje automático que busca regularidades estadísticas no puede evitar incluir ese elemento en su proceso de predicción.

Estos errores de interpretación pueden tener efectos muy perniciosos debido a la forma en que se diseñan las interfaces de comunicación entre los resultados del algoritmo y el humano que va a usar esas predicciones o clasificaciones. La forma típica en la que esas interfaces presentan sus predicciones y clasificaciones es junto a un número que establece la probabilidad de que la predicción sea correcta. Así, un detector artificial de melanomas acostumbra a producir un *output* del estilo de: «Probabilidad de que la mancha sea un melanoma: 17%».

La persona que va a usar esa predicción se encuentra así en una situación complicada: ¿hasta qué punto es fiable esta predicción? Y la mayoría de veces caemos en la falacia de la objetividad numérica: un número tan bien perfilado como «17%» suena la mar de objetivo, mucho más fiable que un humano, que puede estar cansado, confuso o con una opinión sesgada. Las máquinas no se cansan ni tienen prejuicios. Y, así, puede suceder que el profesional médico desoiga su propia intuición sobre el tema y acabe aceptando el número «objetivo» generado por la computadora.

Nos encontramos entonces con otro mito más, generado por los medios de comunicación. Aunque cansarse no se van a cansar, los algoritmos sí pueden cometer errores, como ya hemos visto, y si la base de datos con la que se ha entrenado el algoritmo está sesgada, inevitablemente el algoritmo repetirá esos sesgos. Y a diferencia de un humano, que aprende de sus errores o deja de estar cansado y vuelve a «funcionar» con normalidad, esos errores en el algoritmo permanecen. Una vez terminada la fase de entrenamiento, cuando el algoritmo se ha convertido en una aplicación cerrada para usarse en un hospital, esos errores se ejecutarán una y otra vez, sin el menor atisbo de entender cómo de mal va el proceso.

Sesgos por falta de representación

Una vez hemos revisado un algoritmo de aprendizaje automático y nos hemos asegurado de que sus predicciones están adecuadas a la realidad y no lo llevan a confundir «ser un lobo» con «estar rodeado de nieve» ni «ser un melanoma» con «mancha con una cinta métrica al lado» podemos confiar en

que las predicciones sean coherentes con los datos introducidos, aunque ello no nos asegura que esas predicciones no estén sesgadas.

Consideremos otro programa de clasificación de melanomas, también real. Después de una serie de pruebas, el programa demostraba un fino conocimiento de qué elementos eran significativos para establecer que una mancha era benigna o maligna. El programa recibió un caluroso reconocimiento por parte de los medios de comunicación y también de la clase médica. ¡Se hablaba así de un algoritmo que incluso superaba las capacidades de las personas expertas en dermatología, y además no se cansaba ni cometía errores!

Sin embargo, según se iba revisando el algoritmo, se descubrió que en realidad el programa sí lo hacía muy bien, mejor que una dermatóloga... pero solo con pieles caucásicas. Con pieles más oscuras el algoritmo cometía graves errores. La causa de esos errores estaba en la base de datos utilizada para entrenar al algoritmo. Se trataba de una imponente base de datos de imágenes de manchas de piel, pero en la que las pieles blancas, caucásicas, estaban sobrerrepresentadas, mientras que pieles de otras coloraciones, como las de las personas afrodescendientes, aparecían mucho menos. Al no disponer de suficientes datos sobre etnias no caucásicas, en esos casos el algoritmo cometía muchos más errores.

No hay que acusar a los desarrolladores del algoritmo de racismo o xenofobia. La base de datos simplemente reflejaba la discriminación estructural del sistema sanitario de Estados Unidos. Por una combinación de determinantes sociales como la falta de información, el menor nivel económico o la ausencia de instalaciones sanitarias, el número de personas de origen caucásico advertidas sobre la presencia de melanomas es mucho mayor que el de las personas afrodescendientes.

Recordemos la idea de salida de cómo funcionan los algoritmos de aprendizaje automático: buscan regularidades estadísticas entre los datos que les hemos entregado. Si esos datos están sesgados de salida, inevitablemente, esos sesgos se van a incluir en el diagnóstico y clasificación. De hecho, es lo que buscan: regularidades; y no les importa si están asociadas de manera causal con la predicción que queremos hacer. Lo vimos con el caso del reco-

necedor de lobos y huskies, en el que el algoritmo consideraba que la nieve es un elemento mucho más relevante que el color del pelaje o el aspecto del morro. Y lo volvemos a ver aquí, donde de forma totalmente acientífica el algoritmo deduce que tener la piel oscura es un maravilloso protector contra los melanomas, ya que la mayoría de melanomas –en su base de datos– van asociados a pieles más blancas.

III. Enfrentarse a los desafíos: la búsqueda de la explicabilidad

Para solventar estos desafíos, necesitamos disponer de explicaciones sobre cómo funcionan los algoritmos. El término «explicabilidad» es ahora un *hot topic* en el mundo de la IA, y cada cual lo usa a su manera. En este texto defiendo una explicabilidad dirigida al usuario final. Es decir, no es suficiente con dar el diagnóstico y poner al lado un número que represente la probabilidad de que el diagnóstico sea acertado. El algoritmo ha de ser capaz también de generar algún tipo de narración en el que se expliquen las causas o razones de la predicción propuesta. Así, en lugar de establecer que la probabilidad de que haya un melanoma es del 17%, el algoritmo ha de producir una explicación que el personal sanitario pueda entender, algo del estilo de: considerando que la mancha es simétrica, monocolor, los bordes están equilibrados y es relativamente pequeña (no llega a 5 milímetros) es muy improbable que nos encontremos ante un melanoma.

Cuando esta explicación vaya asociada a las regularidades estadísticas que el algoritmo usa *de facto* para establecer sus clasificaciones y predicciones, los desarrolladores –con la ayuda de asesores médicos para corroborar los diagnósticos– podrán detectar errores y sesgos antes de que el programa sea introducido en entornos sanitarios.

Volviendo a nuestros ejemplos, si en lugar de «la probabilidad de que el animal que aparece en la fotografía sea un husky es del 89%» lo que tenemos es «el animal que aparece en la fotografía es casi seguro un husky porque no hay un gran manto blanco en la foto», sabremos que hay que revisar la base de datos y poner lobos sin nieve y huskies con nieve y volver a empezar. Igualmente, si

nuestro algoritmo detector de melanomas indica que «la mancha de la fotografía no es con seguridad un melanoma, pues la piel de la fotografía es oscura y casi no hay ejemplos de pieles oscuras con melanomas en mi base de datos», sabremos que el algoritmo está replicando la discriminación existente en la actualidad y que hay que revisar la base de datos incluyendo más personas con piel oscura. Igual piensan que esto es una obviedad y que no haría falta plantearlo, pero no es así. Pocos años después de mostrarse que el algoritmo entrenado con la base de datos de fotografías de manchas de piel estaba sesgado, otro equipo de investigadores desarrolló un algoritmo fantástico que también sobrepasaba en habilidad a los humanos. Cuando la comunidad científica fue a revisar el algoritmo, se encontraron con que había sido entrenado con la misma base de datos que infrarrepresentaba poblaciones no caucásicas. Y, más recientemente, Google anunció otro algoritmo maravilloso para detectar melanomas que también se basaba en esa misma base de datos.

IV. Conclusión

Debemos sustituir el paradigma actual, que ve los algoritmos de aprendizaje automático como un oráculo que da una respuesta numérica que el usuario solo puede acatar o rechazar, por un paradigma de interacción entre humano y máquina. En este nuevo paradigma, el diagnóstico se desarrolla de forma conjunta, combinando las capacidades de procesar grandes datos de información por parte de los algoritmos, con la intuición y el conocimiento causal del personal sanitario, que entiende las causas y correlaciones que emergen de la salud y la enfermedad. Simplemente sustituyendo esa mentalidad oracular, la mayoría de los problemas éticos directamente ligados a la IA se desvanecerían.

La integración de la IA en la salud pública presenta desafíos éticos significativos que requieren una atención cuidadosa. A través de la implementación de soluciones enfocadas en la explicabilidad, la equidad y la inclusión, podemos avanzar hacia un futuro en el que la IA contribuya positivamente a la salud pública sin comprometer principios éticos fundamentales.

**Decisiones automatizadas
en sanidad**

Inteligencia artificial y toma de decisiones

Enrique Bernal-Delgado y Francisco Estupiñán-Romero

Ciencia de Datos para la Investigación en Servicios y Políticas Sanitarias. Instituto Aragonés de Ciencias de la Salud, Zaragoza

Fue una conversación con Jeanette Kusel,¹ directora de asesoramiento científico en el National Institute for Health and Care Excellence (NICE),² en el marco de un seminario organizado por la Asociación de Economía de la Salud (AES), donde se empezaron a cimentar algunas de las reflexiones introducidas en este trabajo. La conversación tenía que ver con una creciente preocupación por la incorporación de la Inteligencia Artificial (IA) en los Sistemas de Salud. Se le pedía a Jeanette que reflexionara sobre cómo NICE, en su papel de agencia evaluadora, estaba afrontando la valoración de las IAs y, en su caso, la recomendación de adopción y financiación pública de las mismas como tecnología sanitaria. Como era de esperar, la conversación dejó más preguntas que respuestas y un cierto poso de incertidumbre que quedaba aliviado por la impresión de que las agencias reguladoras ya estaban pensando seriamente en las consecuencias de una incorporación acrítica de la IA. Reconozco que me interesó mucho el pragmatismo de la aproximación de NICE porque contrastaba con la fascinación mediática por la inteligencia artificial, particularmente la IA generativa, una suerte de arcano inaccesible.

De esta conversación, que sucedió en plena pandemia, se destilan algunas reflexiones e interrogantes posteriores que pueden tener impacto en futuras decisiones de salud pública.³ Pero antes de entrar en ellas, quizá sea conveniente sugerir una precisión semántica: cuando nos refiramos a la IA nos referiremos también al, en inglés, *machine learning* (ML) o aprendizaje estadístico. ML es la rama de la IA que utiliza los datos como sustrato de aprendizaje y cuyo principal instrumento es la algoritmia matemática y estadística, que permite resolver cuatro problemas fundamentales: la simulación de

eventos, la identificación de patrones, la clasificación de observaciones y la predicción de sucesos. Para los propósitos de este trabajo nos referiremos indistintamente a IA o a ML.

¿Qué añade la inteligencia artificial a la toma de decisiones en salud pública?

Si, paradigmáticamente, la salud pública confía en poder atribuir causalidad a una exposición poblacional (por ejemplo, nivel de privación, calor, programa de cribado poblacional o copago) sobre un determinado efecto (por ejemplo, obesidad, muerte, supervivencia libre de enfermedad u hospitalizaciones), y así enfocar las políticas y las acciones (las decisiones) a mitigar dicha exposición, es natural preguntarse qué añade (o puede añadir) el uso de la IA a este paradigma tradicional.

Aunque algunos defienden con vehemencia un cambio en el paradigma de conocimiento, no parece que la IA al menos epistemológicamente, esté en condiciones de refutar el falsacionismo de Popper, matizado por Lakatos, ni su correlato para la salud pública, los criterios de causalidad de Austin Bradford Hill.⁴ Dicho de otro modo, la causa debe seguir precediendo al efecto, la asociación observada debe ser consistente con el conocimiento previo, varios estudios con la misma hipótesis deben producir resultados similares, una asociación fuerte es más indicativa de causalidad que una débil, el incremento de exposición debería intensificar el efecto, eliminar la causa debería disminuir el riesgo de tener el efecto, el diseño de investigación que soporta la asociación entre causa y efecto debería disponer de validez interna... o cuantos argumentos acompañen a la determinación de causalidad.

Si no es en el fundamento epistemológico, cabe preguntarse, entonces, si es en lo analítico donde la IA, panoplia de instrumentos construidos sobre las bases de la matemática bayesiana, puede mejorar la toma de decisiones en la salud pública. O, dicho de otro modo, preguntarse si las capacidades del ML para la simulación de eventos, identificación de patrones, clasificación de observaciones y predicción de sucesos mejora las de otros instrumentos

de uso cotidiano –otros los llamarían tradicionales– para resolver problemas de interés de la salud pública.

Pero pasemos de lo general a lo concreto. En la tabla 1 se reflejan tipologías de preguntas sobre salud pública que, en suma, reflejan las necesidades de conocimiento de un agente de la salud pública en su proceso de toma de decisiones. Podemos distinguir dos tiempos en dicho proceso: un momento preadoptivo y un momento postadoptivo. En el primero, el agente de salud pública está planteándose la base de conocimiento, la teoría, que sustentaría el diseño y la implementación de una determinada intervención.⁵ El momento postadoptivo, por su parte, es aquel en el que dicha intervención se ha implementado en la vida real –contexto y personas reales– y es necesario conocer cómo se está desarrollando el proceso de implementación, cómo mejorar la adopción y la efectividad de las medidas o cuál es el impacto parcial o global de la intervención, su efectividad y su eficiencia.

La cuestión es dilucidar si para estas preguntas la IA está mejor equipada que los métodos tradicionales. En las preguntas en el período preadoptivo subyace la necesidad de descubrimiento y actualización de la base de conocimiento. Los métodos tradicionales aquí, revisiones de literatura y síntesis de conocimiento, son limitados en el alcance, utilizan lo publicado en repositorios reconocibles y exigen altos costes de actualización. Sin entrar en cuestiones sobre la calidad de la información que se recupera, los modelos largos del lenguaje (LLM, en sus siglas en inglés), muy eficientes en el descubrimiento y síntesis de la información, mejoran con mucho las capacidades de actualización de la misma.

En las preguntas postadopción, en las que subyacen problemas de clasificación y predicción, el ML resolverá mejor los problemas de gran dimensionalidad, en los que el número de atributos (las columnas en una tabla de datos) es apreciablemente mayor que el número de observaciones (las filas en la tabla de datos). Sin embargo, cuando las tablas tienen muchas filas y pocas columnas, los métodos tradicionales funcionarán perfectamente y conservarán la ventaja de la transparencia analítica y la facilidad de explicación de los hallazgos, cosa que no siempre resulta obvia en el uso de técnicas de *machine learning*, o particularmente de *deep learning*.

Tabla 1. Algunas preguntas arquetípicas en la toma de decisiones en salud pública

| Período preadoptivo | Período postadoptivo |
|--|---|
| ¿Qué intervenciones innovadoras disponen de evaluaciones formales de efectividad y eficiencia? | ¿Cómo se está adoptando la intervención en salud pública? |
| ¿Hay lagunas de conocimiento sobre la efectividad y eficiencia de las intervenciones en salud pública? | ¿Se observan desigualdades en la adopción asociadas a características de la población? |
| ¿Qué intervenciones han demostrado ser más eficaces en entornos como el nuestro? | ¿Se observa un uso inadecuado de la intervención, por su aplicación fuera de protocolo – poblaciones distintas a las previstas? |
| ¿Cuáles serían los criterios de valoración pertinentes para responder a una pregunta PICO-t? | Ante la necesidad de priorización, ¿cuáles son los métodos que maximizan el impacto de mis decisiones? |
| ¿Cuáles serían los comparadores pertinentes para evaluar una intervención? | ¿Cuál ha sido la experiencia de las personas sobre las que se ha intervenido? |
| ¿Cuáles podrían ser los conjuntos de datos pertinentes para utilizar en la evaluación de una intervención interregional? | ¿Cuál ha sido el impacto organizativo y presupuestario de la implementación de la intervención? |
| ¿Cuál es la base de conocimientos pertinente para fundamentar nuestras hipótesis? | ¿Se han reducido los riesgos para la salud de las personas destinatarias de la intervención? |
| ¿Cuáles son las áreas grises de reglamentación que sustentan una intervención? | ¿Se han reducido las desigualdades en la población tras la implementación de la intervención? |
| (...) | (...) |

En cualquier caso, es pertinente subrayar que la IA no resolverá las maldiciones del uso de datos en la toma de decisiones, particularmente cuando se movilizan datos masivos o los llamados datos «de la vida real», observacionales por naturaleza y coleccionados tras el contacto de las personas con los sistemas de salud. Entre esas maldiciones: a) la aparición de correlaciones espurias, difíciles de distinguir de asociaciones plausibles; b) la existencia de extra-heterogeneidad, derivada del uso de múltiples fuentes de datos; c) la incompletitud de información o la presencia de celdas vacías, con frecuencia no distribuidas al azar o, d) la dificultad para controlar fenómenos como la confusión o los sesgos de selección, particularmente cuando existen fenómenos latentes no observables.

En resumen, la IA/ML no supone un cambio de aproximación epistemológica para la toma de decisiones en la salud pública, aunque es más eficiente en el descubrimiento y actualización de la base de conocimiento y puede obtener resultados más robustos cuando se necesita resolver problemas de clasificación y predicción en condiciones de alta dimensionalidad. Sin embargo, estas ventajas no eximen de la necesidad de diseños de investigación con fuerte validez interna ni de la obligación de mitigar el error sistemático asociado a la calidad de los datos y a los sesgos, particularmente cuando la toma de decisiones requiere información sobre atribuciones causales.

Incorporar la IA a la toma de decisiones en salud pública

La discusión sobre la incorporación de la IA a la toma de decisiones está ciertamente sesgada hacia su utilización en el ámbito clínico y sanitario, y es desde esta perspectiva desde donde planteo las siguientes reflexiones. No obstante, considero que estas reflexiones son, en general, de aplicación a las decisiones en salud pública.

Vuelvo aquí a la conversación con Jeanette Kusel y a la aproximación de NICE para la incorporación de la IA en el sistema de salud británico. A efectos de la evaluación de la IA, NICE establece dos condiciones *a priori*. La primera asumiría que la IA, es decir el código o el *software* que la soporta, debe ser

considerado de la misma naturaleza que un dispositivo tecnológico, y, por tanto, ser evaluable mediante los criterios habituales de evaluación de tecnologías; y la segunda asumiría que se trata de una tecnología que no debe evaluarse de forma aislada, sino dentro del proceso asistencial y en el contexto organizativo en el que se vaya a implementar.

Esta aproximación pragmática plantea algunas cuestiones metodológicas que retan al planteamiento. En primer lugar, cabe preguntarse qué procedimiento se debe seguir para valorar el proceso precomercial de desarrollo de la inteligencia artificial (y sus modificaciones); en segundo lugar, no es obvio definir un estándar de cuidados contra el que comparar la IA objeto de evaluación (es decir, cómo determinar su valor añadido con respecto a la práctica habitual); en tercer lugar, mientras que las tecnologías tradicionales son inmutables, la IA se adapta a las condiciones de uso (aprende de su heurístico y del uso que se hace de ellas) o requiere modificaciones más frecuentes, no resultando obvia ni su evaluación ni su elaboración de recomendaciones de adopción universal; y, en cuarto lugar, no es obvio, diferenciar entre señal y ruido en aquellas IA no autoexplicativas y, por tanto, es preciso elaborar recomendaciones de uso apropiado.

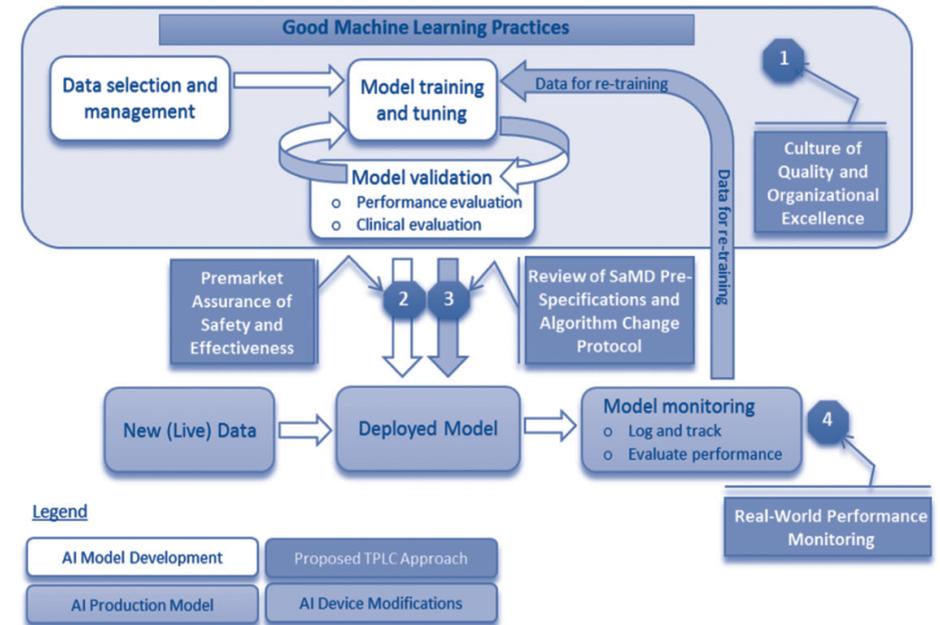
Sobre los principios rectores del desarrollo de la IA, la Food and Drugs Administration (US), Health Canada y la Medicines and Healthcare Products Regulatory Agency of the United Kingdom acordaron en 2021 algunos de los parámetros básicos para el desarrollo e incorporación de la IA.⁶ Los principios rectores que aparecen en esta guía señalan que: 1) debería existir una comprensión multidisciplinar y exhaustiva del valor –clínico, sanitario, de salud pública– de la incorporación de la IA en el momento en el que se está desarrollando; 2) el diseño e implementación de los algoritmos de la IA deberían seguir los principios de buena práctica del desarrollo de *software*, calidad de datos y ciberseguridad; 3) los participantes en el desarrollo de la IA y los conjuntos de datos utilizados deberían representar adecuadamente la población sobre la que se aplicaría el producto; 4) los conjuntos de entrenamiento deberían ser independientes de los conjuntos de datos sobre los que la IA se valida; 5) el desarrollo de la IA debería realizarse sobre la base de los métodos y algoritmos más apropiados a cada caso concreto; 6) la validez de la IA, una vez superada la fase de pruebas,

debería confirmarse en condiciones reales de uso; 7) el despliegue de los modelos en la vida real debería seguirse de una monitorización de los efectos positivos o negativos de su utilización, tanto a nivel del paciente como de la organización; y, 8) los usuarios de la IA, ya sean clínicos, pacientes o poblaciones, deberían disponer de la información adecuada sobre los beneficios y riesgos derivados del uso del producto.

En definitiva, lo que estos principios proponen, coincidiendo por cierto con los postulados de la aproximación de NICE, es un enfoque regulador que abarque el ciclo de vida del algoritmo (en inglés, Software as a Medical Device: SaMD), tanto en las fases preadoptivas (desarrollo previo a la comercialización) como postadoptivas, (la interacción entre la IA y el contexto organizativo en el que se implementa, el impacto real sobre las poblaciones diana). En la figura 1, tomada del documento de discusión para la elaboración de un marco regulatorio de la IA,⁷ se reflejan los pasos críticos de la regulación:

1. Establecer cuáles son las expectativas de desarrollo de la IA, fundamentadas en las buenas prácticas de desarrollo de la IA/ML.
2. Llevar a cabo la revisión previa a la comercialización para demostrar una garantía razonable de seguridad y eficacia y establecer las bases para el seguimiento de los potenciales riesgos para pacientes y poblaciones después de la adopción.
3. Instruir a los desarrolladores para que incorporen un enfoque de gestión de riesgos, así como qué procedimiento se debe seguir en el caso de un cambio de *software* en un dispositivo existente.
4. Estudiar los efectos positivos y negativos del uso de la IA/ML en el mundo real, con objeto de mantener una garantía continua de efectividad y seguridad.

Figura 1



Este enfoque regulador, todavía en desarrollo, tiene matices, de modo que tanto la intensidad regulatoria como la aproximación metodológica pueden ser distintas dependiendo de: 1) si la IA objeto de regulación es un algoritmo cerrado o una IA generativa; 2) la relevancia de la información proporcionada por la IA en la toma de decisiones; o, 3) la criticidad de la condición clínica o de salud pública que potencialmente puede beneficiarse de la IA. El IMDRE, International Medical Devices Regulators Forum, propuso en 2014 una gradación de riesgos que orientase la intensidad regulatoria.⁸ Este marco ha sido adoptado por la FDA y clasificaría las IA en cuatro categorías (tabla 2):

Tabla 2. Categorización de riesgos para la regulación de la IA

| Situación que afecta al sistema de salud, la población o la condición clínica | Relevancia de la información aportada por la IA en la toma de decisiones | | |
|---|--|---|--|
| | Diagnosticar y tratar –o clasificar– grupo poblacional e intervenir | Orientar decisiones de gestión clínica o de la intervención | Informar sobre la gestión clínica o de la intervención |
| Crítica | IV | III | II |
| Seria | III | II | I |
| No seria | II | I | I |

n.b. Esta tabla está modificada por el IMDRF para aportar la visión poblacional.

Las cuatro categorías (I, II, III, IV) se basan en los niveles de impacto que sobre los pacientes, la salud pública y el sistema, tiene la información proporcionada por la IA. La IA destinada a producir información para diagnosticar y tratar, o para decidir quién se beneficia de un programa de salud pública e intervenir, estaría clasificada en la categoría IV de alto impacto potencial en situaciones de afección crítica. En el otro extremo, la IA pensada para proporcionar información sobre la gestión clínica o la gestión de una intervención, en una situación de poco nivel de criticidad, estaría clasificada en la categoría I, lo que implica que la desarrollada para cubrir ese espectro tendría un menor impacto relativo. Un ejemplo de inteligencia artificial clasificada en la categoría IV podría ser el de una que combina datos de inmunoensayos para detectar de forma temprana patógenos mutables o brotes pandémicos que pueden ser altamente transmisibles por contacto directo u otros medios y con alta tasa de letalidad. En el otro extremo, un ejemplo de inteligencia artificial clasificada en la categoría I podría ser el de una destinada a pacientes de edad avanzada con múltiples afecciones crónicas que recibe datos de sensores de salud vestibles, transmite datos al servidor de monitorización e identifica información de nivel superior, como taquicardias y signos de infecciones respiratorias, basándose en conocimientos médicos establecidos, y comunica esta información a los cuidadores.

Independientemente de las cuestiones regulatorias, la utilización de la IA en las decisiones de salud pública requerirá que los actores e instituciones de salud pública se familiaricen con los instrumentos de la IA/ML y valoren si esta podría añadir valor a sus actividades habituales; por ejemplo, monitorización y vigilancia de los determinantes de salud, evaluación de la adopción de las intervenciones de salud pública y evaluación de la efectividad y eficiencia de las mismas. Adicionalmente, será preciso que quienes toman las decisiones en salud pública añadan cautelas adicionales cuando quieran adoptar una IA clasificada como categoría III o IV. Finalmente, los actores de salud pública necesitarán entrenamiento en el uso e interpretación de los resultados obtenidos por los algoritmos ML.

Corolario

La salud pública se beneficiará de la adopción de la IA en múltiples aspectos relacionados con el descubrimiento y actualización de las bases de conocimiento necesarias para la toma de decisiones y en los aspectos más evaluativos de la acción en salud pública.

La adopción de la IA requiere de una evaluación previa, minuciosa, adaptada al impacto potencial de la IA sobre la salud de las personas y las poblaciones, sobre los sistemas de salud y sanitarios, y que se realice a lo largo de todo el ciclo de vida de la IA. Es tan relevante la evaluación precomercial como la evaluación continuada en situaciones de vida cotidiana.

Las agencias reguladoras se encargarán de proporcionar inteligencias artificiales confiables. Pero es responsabilidad de los actores e instituciones de salud pública el buen uso de la IA. Es necesario capacitar a la fuerza de trabajo en estas herramientas y su interpretación.

Notas

1. <https://www.nice.org.uk/news/article/jeanette-kusel-is-appointed-as-director-for-nice-scientific-advice>

2. Para aquellos que no conozcan NICE, siglas de esta reconocida institución, cabe decir que se trata de un organismo independiente, vinculado al departamento de Salud y Atención Social del Reino Unido, cuya misión es ayudar a los profesionales y gestores a proveer los mejores cuidados, a tiempo, mientras asegura que los contribuyentes obtengan a cambio un sistema sanitario de valor.
3. Estas incluyen las decisiones a nivel macro, meso y micro que pretenden impactar sobre los determinantes de salud, incluyendo el sistema sanitario.
4. Bradford Hill A, The Environment and Disease: Association or Causation?, Proceedings of the Royal Society of Medicine, 58 (1965), 295-300.
5. Intervención es sinónimo de práctica, acción política, incluso de tecnología.
6. Good Machine Learning Practice for Medical Device development: Guiding principles. Disponible en: <https://www.fda.gov/media/153486/download>
7. Proposed Regulatory Framework for Modifications to Artificial Intelligence/Machine Learning (AI/ML)-Based Software as a Medical Device (SaMD) - Discussion Paper and Request for Feedback (2019). Disponible en: <https://www.fda.gov/media/122535/download?attachment>
8. Software as a Medical Device: Possible Framework for Risk Categorization and Corresponding Considerations. Authoring Group: IMDRF Software as a Medical Device (SaMD) Working Group, 2014. Disponible en: <https://www.imdrf.org/sites/default/files/docs/imdrf/final/technical/imdrf-tech-140918-samd-framework-risk-categorization-141013.pdf>

Decisiones automatizadas en sanidad: IA generativa

Jordi Vallverdú

ICREA Academia. Universitat Autònoma de Barcelona

Es importante señalar que el análisis y las conclusiones de esta ponencia están parcialmente basados en el artículo «Challenges and Controversies of Generative AI in Medical Diagnosis» de Vallverdú, J. (2023), publicado en *Euphyia. Revista de Filosofía*, 17(32), páginas 88-121. Este trabajo proporciona una visión integral de los desafíos éticos, filosóficos y prácticos asociados a la inteligencia artificial generativa (IAG) en el ámbito médico, y explora temas como la autoría y responsabilidad, autonomía del paciente, impacto social y económico, consideraciones computacionales y médicas, así como una perspectiva holística centrada en el paciente.

La paciente Sarah, de 38 años, presenta manifestaciones clínicas caracterizadas por erupciones cutáneas y episodios febriles. Inicialmente, la paciente se somete a una interacción con un chatbot médico; sin embargo, la Inteligencia Artificial (IA) empleada en dicho sistema no logra discernir la etiología subyacente de los síntomas. Consciente de las limitaciones de la IA en el ámbito de la telemedicina, la paciente, identificada como Emily, busca la asesoría de su médico de atención primaria. El profesional médico, con el objetivo de realizar un análisis exhaustivo de los datos clínicos de la paciente, incorpora el uso de inteligencia artificial generativa, en particular, empleando Redes Generativas Adversarias (GAN) y Auto-Encoders Variacionales (VAE). La IA, al ser desplegada en este contexto, logra identificar patrones médicos complejos, simulando diversos escenarios diagnósticos potenciales. El resultado de esta simulación se traduce en una variedad de diagnósticos, que son presentados al médico en función de su probabilidad de ocurrencia. Esta integración de la IA no solo enriquece la consideración diagnóstica, sino que también expande el espectro de posibles condiciones patológicas, algunas de las cuales no habían sido inicialmente contempladas. La sinergia entre la inteligencia artificial y la

experiencia clínica del médico contribuye de manera significativa al proceso diagnóstico. Como conclusión, se obtiene un diagnóstico más preciso y se reconoce la contribución crucial de la inteligencia artificial en el esclarecimiento de los desafíos diagnósticos en el ámbito médico.

Este fragmento describe un caso hipotético que refleja la reflexión sobre el posible impacto de la inteligencia artificial generativa en el diagnóstico médico desde una perspectiva multidimensional. Aquí se exponen varios elementos clave, que procedemos a remarcar y enumerar:

1. Presentación del caso:

- La paciente, Sarah, de 38 años, presenta síntomas como erupciones cutáneas y episodios febriles.
- Inicialmente, la propia paciente recurre a un chatbot médico, pero la IA no logra discernir la causa subyacente de los síntomas.

2. Limitaciones de la IA en telemedicina:

- Se destaca la conciencia de las limitaciones de la IA en el ámbito de la telemedicina.

3. Consulta con el médico de atención primaria:

- La paciente, identificada como Emily, busca la asesoría de su médico de atención primaria ante la incapacidad del chatbot.
- El médico decide realizar un análisis más exhaustivo y recurre a la inteligencia artificial generativa.

4. Uso de redes generativas adversarias (GAN) y autocodificadores variacionales (VAE):

- Se menciona el uso específico de tecnologías avanzadas como GAN y VAE para analizar datos clínicos.

5. Simulación de escenarios diagnósticos potenciales:

- La inteligencia artificial simula varios escenarios diagnósticos potenciales, identificando patrones médicos complejos.

6. Resultados de la simulación:

- La simulación arroja una variedad de diagnósticos con sus respectivas probabilidades de ocurrencia.

7. Ampliación del espectro diagnóstico:

- Se destaca cómo la IA expande el espectro de posibles condiciones patológicas, algunas de las cuales no habían sido consideradas inicialmente.

8. Sinergia entre la IA y la experiencia clínica:

- Se resalta la colaboración entre la inteligencia artificial y la experiencia clínica del médico, subrayando la contribución significativa al proceso diagnóstico.

9. Diagnóstico más preciso:

- Como conclusión, se obtiene un diagnóstico más preciso gracias a la sinergia entre la inteligencia artificial y la experiencia del médico.

10. Contribución crucial de la IA en desafíos diagnósticos:

- Se reconoce la contribución crucial de la inteligencia artificial en el esclarecimiento de desafíos diagnósticos en el ámbito médico.

El caso plantea una situación compleja en la que la integración de la inteligencia artificial generativa y la experiencia clínica resulta en un proceso diagnóstico más enriquecedor y preciso. La narrativa busca resaltar cómo la IA puede ser una herramienta valiosa para los profesionales de la salud, complementando y ampliando sus capacidades.

El panorama de la atención médica ha experimentado una transformación significativa en los últimos años con la incorporación de tecnologías de vanguardia. Uno de los avances más destacados es la introducción de la IAG en el ámbito médico (Vallverdú, 2023). Este enfoque innovador utiliza algoritmos avanzados, específicamente las GAN y los VAE, para simular y generar datos, revolucionando la forma en que se abordan los diagnósticos y tratamientos médicos. La IA generativa desempeña un papel crucial en diversas aplicaciones médicas, destacándose particularmente en el ámbito del diagnóstico. Los modelos generativos, como las GAN, han resultado efectivos en la generación de imágenes médicas realistas. Esto facilita el entrenamiento de algoritmos de diagnóstico por imágenes, al utilizar datos más diversos y extensos. Ciertos estudios respaldan el potencial de la IA generativa para mejorar la precisión en el diagnóstico por imágenes, al emplear la capacidad

de generar imágenes sintéticas en el entrenamiento de algoritmos de aprendizaje automático. La investigación se ha ampliado para incluir el uso de modelos generativos en datos clínicos, como historias clínicas electrónicas. Esto ofrece la posibilidad de generar datos sintéticos para el entrenamiento de algoritmos de diagnóstico sin comprometer la privacidad de los pacientes. La IA generativa, a través de técnicas como los VAE, ha tenido éxito en la realización de diagnósticos diferenciales al simular diversos escenarios y proporcionar opciones de diagnóstico basadas en datos clínicos. No obstante, la integración de la IA generativa en el ámbito médico plantea desafíos éticos, especialmente en la interpretación de decisiones tomadas por algoritmos. Se destaca la importancia de una validación rigurosa para garantizar la seguridad y eficacia de estos modelos, abordando preocupaciones éticas y optimizando su aplicación en entornos clínicos. Es crucial mantenerse al tanto de los desarrollos más recientes, ya que la investigación en este campo sigue evolucionando. Inmersos en el ámbito de los diagnósticos y tratamientos médicos, visualicemos un escenario en el que los profesionales médicos, empoderados con la inteligencia artificial, puedan analizar patrones complejos dentro de vastos conjuntos de datos, conduciendo a diagnósticos más precisos y oportunos. Esta es la promesa que la inteligencia artificial generativa tiene para el futuro de la medicina. Imitando los procesos de aprendizaje del cerebro humano, estos sistemas inteligentes comprenden relaciones intrincadas dentro de los datos médicos, ofreciendo un espectro de posibles diagnósticos que podrían pasar desapercibidos.

En este contexto, uno de los avances significativos en la aplicación de la inteligencia artificial generativa a la atención médica se manifiesta en la autoevaluación del paciente. Imaginemos un mundo en el que, independientemente de la experiencia médica, podamos interactuar con chatbots impulsados por la IA para evaluar síntomas, recibir posibles diagnósticos y obtener información sobre nuestra salud. Este enfoque no solo capacita a las personas para desempeñar un papel proactivo en su bienestar, sino que también abre nuevas vías para la accesibilidad y eficiencia en la prestación de servicios de salud. En este escenario, la intersección de la inteligencia artificial generativa con la autoevaluación del paciente promete transformar la dinámica de la atención médica, ofreciendo una perspectiva innovadora hacia la participación activa

y la eficacia en el cuidado de la salud. A medida que profundizamos en el impacto de la inteligencia artificial generativa en las prácticas médicas, se hace evidente que estos sistemas inteligentes van más allá de las tareas rutinarias, contribuyendo al rápido análisis de extensos conjuntos de datos, mejorando las capacidades de diagnóstico e incluso sirviendo como coautores en artículos de investigación científica. Sin embargo, junto con estas promesas surgen consideraciones éticas, desafíos a la hora de garantizar la privacidad de los datos y la necesidad de esfuerzos colaborativos entre profesionales de la salud y expertos en tecnología.

Esta introducción tiene como objetivo desmitificar el ámbito de la inteligencia artificial generativa y arrojar luz sobre su potencial para revolucionar los diagnósticos médicos, mejorar los resultados para los pacientes y allanar el camino hacia una experiencia de atención médica más personalizada y eficiente. Acompañennos en este viaje hacia el futuro de la medicina, donde la sinergia entre la experiencia humana y la inteligencia artificial promete redefinir los límites de lo posible en la búsqueda de una mejor salud para todos.

Usos de la IA en diagnóstico

En el ámbito de la inteligencia artificial generativa aplicada a la medicina, diversos desafíos y estudios proporcionan una visión integral de sus capacidades y limitaciones. Uno de los principales desafíos apela a la necesidad de una validación rigurosa, la interpretación de los diagnósticos generados por la IA y ciertas consideraciones éticas y de privacidad. La utilización de modelos de lenguaje extensos, como ChatGPT, emerge como una herramienta significativa para diversas tareas, como por ejemplo la interacción con bases de datos, lo que simplifica la búsqueda de información clínica de manera accesible. Asimismo, se resalta el potencial de la inteligencia artificial generativa para simular situaciones clínicas con el propósito de aprender y perfeccionar tratamientos. A pesar de estas ventajas, se reconoce la importancia del control de calidad, especialmente en contextos médicos, como medida crucial para garantizar la confiabilidad de los resultados. Recordemos el caso del niño llamado Alex, residente en una ubicación no revelada, que atravesó una

odisea médica de tres años que involucró visitas a 17 médicos por dolor crónico, y que permaneció sin diagnosticar. A pesar de múltiples consultas médicas, sus síntomas persistieron, lo que llevó a su madre, Courtney (apellido omitido por privacidad), a recurrir a ChatGPT en busca de ayuda. Courtney, expresando su frustración al tiempo que su determinación, creó una cuenta en la plataforma de inteligencia artificial y compartió extensa información sobre los síntomas de Alex y todos los datos disponibles de sus resonancias magnéticas. La plataforma de inteligencia artificial, conocida como ChatGPT, sugirió un posible diagnóstico de síndrome de médula anclada, una condición en la que el movimiento de la médula espinal se ve restringido debido a anexos anormales, a menudo asociados con la espina bífida. Esta revelación impulsó una investigación médica adicional, y Courtney programó una cita con un nuevo neurocirujano. Durante la consulta, compartió sus sospechas con el médico, quien confirmó el diagnóstico después de examinar las imágenes de resonancia magnética de Alex. El síndrome de médula anclada está estrechamente asociado con la espina bífida y ocurre cuando el movimiento de la médula espinal se ve limitado, causando un estiramiento anormal. En muchos niños con espina bífida hay una abertura visible en la espalda del niño. Sin embargo, el tipo que tenía Alex es cerrado y se considera «oculto». De ahí que se le conozca también como espina bífida oculta, según los centros para el control y la prevención de enfermedades de EE. UU. Alex, descrito por su madre como «feliz y despreocupado», además de inteligente, disfrutaba jugando con otros niños y participando en actividades como el béisbol y el hockey. A pesar de tener que enfrentarse a numerosos desafíos debido a su condición médica, encuentra formas de adaptarse y participar en las actividades que ama. Este caso atrajo la atención mundial cuando apareció en las noticias el 11 de septiembre de 2023. La historia del periplo médico de Alex y el papel de ChatGPT a la hora de identificar su condición alcanzó un reconocimiento generalizado, arrojando luz sobre los desafíos que afrontan las familias para obtener diagnósticos médicos completos. Después del diagnóstico, hace unas semanas Alex se sometió a una cirugía para abordar su síndrome de médula anclada y actualmente se está recuperando. Courtney compartió la historia de su familia con la intención de ayudar a otros que se enfrenten a desafíos similares, enfatizando la importancia de ser un defensor proactivo de sus hijos al navegar en situaciones médicas complejas.

En otro estudio (Kuroiwa *et al.*, 2023), se evaluó la eficacia de ChatGPT en el autodiagnóstico de condiciones ortopédicas durante un período de cinco días. Los autores realizaron preguntas sobre el síndrome del túnel carpiano (CTS), la mielopatía cervical (CM), la estenosis espinal lumbar (LSS), la osteoartritis de rodilla (KOA) y la osteoartritis de cadera (HOA). Las respuestas se categorizaron como correctas, parcialmente correctas, incorrectas o diagnósticos diferenciales. Se observó una variabilidad en la precisión de las respuestas para diferentes condiciones, y se destacó la necesidad de mejorar la identificación de síntomas y la claridad en el procesamiento del lenguaje natural. En otro estudio (Mehnen *et al.*, 2023), se evaluó ChatGPT 4 en diagnósticos médicos comunes y se comparó con su versión anterior, ChatGPT 3.5. Los resultados mostraron que ChatGPT 4 fue efectivo en diagnósticos comunes para individuos no expertos, pero que para casos raros se requerían más sugerencias. No obstante, ChatGPT 4 superó a ChatGPT 3.5 en precisión diagnóstica, resaltando su mejora en la capacidad diagnóstica. La utilidad potencial de ChatGPT 4 como fuente estándar para la información médica se exploró en un contexto más amplio. Se comparó con fuentes estáticas de internet en términos de información médica, y se destacó su comparabilidad. Asimismo, se llevó a cabo una evaluación de calidad mediante la herramienta EQIP en condiciones hepato-pancreático-biliares, demostrando que ChatGPT 4 es comparable a otras fuentes de información médica en la web. Mirando hacia el futuro, la incorporación de la IA generativa en diagnósticos médicos representa un cambio de paradigma con oportunidades prometedoras, pero también desafíos significativos. Se enfatiza la necesidad de un enfoque equilibrado y colaboración entre profesionales de la salud, tecnólogos y expertos éticos para abordar de manera efectiva estos desafíos y aprovechar el potencial transformador de la inteligencia artificial en el campo de la medicina.

Controversias

Las controversias en torno al uso de la IAG en el ámbito médico se manifiestan en diversas dimensiones, desde la interpretación de datos hasta las implicaciones en la relación médico-paciente y la autonomía del individuo. La

primera controversia desarrollada por Diálogo socrático sobre causalidad en epidemiología (Cox, 2023) se centra en la discusión de las asociaciones entre PM 2.5 (partículas muy pequeñas en el aire con un diámetro de 2,5 micrómetros) y los riesgos de mortalidad. Aquí, se plantean desafíos relacionados con la interpretación y confiabilidad de las conclusiones generadas por la IA. Problemas epistémicos, interpretativos y de calidad de datos surgen como cuestionamientos fundamentales, ya que la complejidad de las relaciones entre variables puede llevar a conclusiones erróneas o sesgadas. En cuanto a las aplicaciones clínicas de ChatGPT (Eysenbach, 2023), la segunda controversia, destaca el potencial transformador de la IA en la medicina clínica, pero también señala desafíos significativos. Las limitaciones en diagnósticos detallados y aplicaciones críticas generan controversias en la educación médica y las políticas editoriales de revistas. La participación de la IA en la generación de becas, el proceso de revisión por pares y la publicación de resultados científicos plantea preguntas sobre la validez y la confiabilidad de la información generada por estos sistemas. La tercera controversia se centra en el fenómeno de «Dr. Google 2.0» y el autodiagnóstico, que veremos con más detalle en la próxima sección. Aquí, se destaca el impacto en la relación médico-paciente y la autonomía del individuo al buscar información médica en línea. La interpretación incorrecta de la información, el exceso de énfasis en las condiciones raras y el sesgo de confirmación son preocupaciones clave. Existe el riesgo de retraso en la búsqueda de ayuda profesional debido a la confianza excesiva en la información en línea, lo que podría tener impactos psicológicos y conducir a la medicalización de variaciones normales. En respuesta a estas controversias, se enfatiza la necesidad de promover la alfabetización en salud y destacar el papel complementario del asesoramiento médico profesional para garantizar una toma de decisiones informada y equitativa. En resumen, las controversias en el uso de IA generativa en medicina abarcan desde la calidad de los datos y la interpretación de resultados hasta las implicaciones en la relación médico-paciente y la autonomía individual. Estas cuestiones subrayan la importancia de abordar cuidadosamente los desafíos éticos, clínicos y sociales asociados con la implementación de la inteligencia artificial en el campo de la salud.

Dr. Google 2.0

El uso de ChatGPT en la comunicación con los médicos, dentro del marco de la ciencia ciudadana o de la sociedad civil informada y participativa, refleja la tendencia del fenómeno conocido como «Dr. Google». En ambos casos, los individuos buscan activamente comprender y gestionar su salud, utilizando herramientas tecnológicas para acceder a información relevante. Las personas, al enfrentarse a condiciones médicas complejas, a menudo experimentan barreras en la comunicación con profesionales de la salud. La capacidad de utilizar ChatGPT para simplificar y explicar diagnósticos puede mejorar significativamente la comprensión entre médicos y pacientes. En un enfoque de ciencia ciudadana, en el que los pacientes se convierten en participantes activos en la gestión de su salud, ChatGPT se convierte en una herramienta de empoderamiento. Algunas de las maneras en las que ChatGPT puede facilitar la comunicación son la explicación detallada de diagnósticos, la preparación para consultas médicas y la formulación de preguntas específicas para abordar inquietudes. Esto se alinea con los principios de la ciencia ciudadana al involucrar activamente a los ciudadanos en la toma de decisiones relacionadas con su salud y brindarles herramientas que faciliten una comunicación más efectiva con los profesionales de la salud. Cuando los pacientes utilizan ChatGPT en su preparación para citas médicas, la realidad se alinea con el enfoque de ciencia ciudadana, al poner a los individuos en el centro de su propio cuidado. Los usuarios pueden aprovechar ChatGPT para formular preguntas específicas, comprender mejor sus condiciones y expresar preocupaciones de manera efectiva. Esto refleja la idea fundamental de la ciencia ciudadana, que implica la participación activa de los ciudadanos en la investigación y la toma de decisiones científicas.

En un contexto de ciencia ciudadana, los pacientes pueden considerar al ChatGPT como una herramienta colaborativa que les ayuda a estructurar sus interacciones con los médicos. Esto puede resultar en un diálogo más informado y enriquecedor entre el profesional de la salud y el ciudadano, en el que la información generada por la IA complementa la experiencia y conocimientos del médico. La «teoría de las cucharas», que ilustra cómo las personas con enfermedades crónicas tienen recursos limitados de energía, se conecta con el papel de la IA, como ChatGPT, en la investigación de la

salud. La teoría de las cucharas es una metáfora utilizada para describir la gestión de la energía en personas que viven con enfermedades crónicas o condiciones de salud que limitan sus recursos físicos y mentales. Fue acuñada por Christine Miserandino (Miserandino, 2017), una mujer con lupus, para ilustrar cómo las personas con enfermedades crónicas deben hacer elecciones cuidadosas sobre cómo gastar su limitada energía a lo largo del día. La metáfora se desarrolla de la siguiente manera: se le da a una persona un número limitado de cucharas al comienzo de cada día, representando la cantidad de energía que tienen disponible para realizar diversas actividades. Cada actividad, desde levantarse de la cama hasta realizar tareas cotidianas, requiere el uso de una cuchara. Cuando las cucharas se agotan, la persona no tiene más energía para el resto del día. Esta metáfora destaca la necesidad de que las personas con enfermedades crónicas tomen decisiones conscientes sobre cómo utilizan su energía para evitar agotarse demasiado rápido. También ayuda a explicar a otras personas la naturaleza limitada de la energía de alguien con una enfermedad crónica, ya que pueden no comprender completamente las restricciones a las que se enfrenta esa persona. Aquellos que viven con condiciones médicas no diagnosticadas a menudo se vuelven «informados médicamente» por necesidad, y la IA puede facilitar este proceso al proporcionar acceso rápido y eficiente a información relevante. En términos de ciencia ciudadana, ChatGPT se convierte en una herramienta valiosa para todos aquellos que busquen respuestas a preguntas específicas sobre su salud. La capacidad de realizar búsquedas detalladas, plantear preguntas relacionadas con condiciones específicas y almacenar conversaciones previas facilita la investigación continua de quienes participan activamente en la gestión de su salud. La combinación de la teoría de las cucharas y la investigación de la IA destaca cómo las herramientas tecnológicas pueden aliviar la carga de la búsqueda de información para aquellos con recursos limitados.

La Macdonalización de la medicina

La integración de la inteligencia artificial generativa en el ámbito médico, especialmente cuando es desarrollada y mantenida por empresas privadas,

plantea una serie de desafíos significativos. Nos encontramos ante la perspectiva de lo que podría denominarse «McThinking», o una suerte de «McMedicina», en la que los modelos y patrones diseñados por estas entidades privadas pueden influir de manera considerable en la toma de decisiones clínicas y en la formulación de diagnósticos. En este contexto, la tiranía invisible del algoritmo privado se manifiesta a través de la capacidad de estas empresas para diseñar modelos que asignen pesos o condicionen variables de manera que puedan influir en el resultado final. Este enfoque comercial puede introducir sesgos indeseados en la interpretación de los datos médicos, afectando a la objetividad y la equidad en la atención médica.

La dependencia de algoritmos privados en el campo de la inteligencia artificial médica plantea preocupaciones sobre la transparencia y la rendición de cuentas. La opacidad en la formulación de estos algoritmos puede limitar la comprensión de los profesionales de la salud sobre cómo se llega a un determinado diagnóstico, lo que, a su vez, puede socavar la confianza en el uso de la inteligencia artificial en el entorno clínico. Además, la influencia de empresas privadas en el desarrollo de algoritmos médicos puede conducir a una falta de diversidad y representación en los conjuntos de datos utilizados para entrenar dichos modelos. Esto podría resultar en desafíos a la hora de abordar de manera precisa y equitativa las condiciones médicas en diversas poblaciones, exacerbando las disparidades existentes en la atención médica.

En resumen, la introducción de la inteligencia artificial generativa de manos de empresas privadas plantea retos significativos, desde la posibilidad de sesgos algorítmicos hasta la falta de transparencia y diversidad en el desarrollo de modelos médicos. Abordar estos desafíos es crucial para garantizar que la implementación de la inteligencia artificial en la medicina beneficie a la sociedad en su conjunto, sin comprometer la integridad y la equidad de la atención médica.

Conclusiones

En la intersección entre la tecnología y la salud, la toma de decisiones automatizadas en el ámbito sanitario ha experimentado avances notables, desta-

cándose la inteligencia artificial generativa (IAG) como un factor clave en esta transformación. Este fenómeno no solo representa una evolución en la prestación de la atención médica, sino que también plantea cuestiones fundamentales sobre cómo las decisiones automatizadas, impulsadas por algoritmos inteligentes, influyen en el diagnóstico, tratamiento y gestión de la salud.

La IAG, con sus redes neuronales avanzadas y su habilidad para generar datos, está redefiniendo la toma de decisiones críticas en el ámbito de la salud. Desde simular patrones médicos complejos hasta mejorar la precisión diagnóstica, la IAG ha constatado la relevancia de su impacto. Sin embargo, estos avances también han suscitado preocupaciones éticas y filosóficas. La atribución de determinada autoría a sistemas de IAG plantea interrogantes sobre responsabilidad y ética en la investigación médica, especialmente cuando contribuyen a la generación de conocimiento científico.

La capacidad de los pacientes para realizar autodiagnósticos a través de la IAG ha generado preguntas profundas sobre la autonomía del paciente, la interpretación de la información y la toma de decisiones informada. Al mismo tiempo, el impacto social y económico de la IAG en la atención médica plantea desafíos considerables. El acceso desigual a la tecnología y a la IAG podría amplificar disparidades en el acceso a diagnósticos precisos entre diferentes segmentos de la sociedad. Además, la automatización en el ámbito médico, impulsada por la IAG, podría tener implicaciones económicas, sin descartar el posible desplazamiento de ciertos roles profesionales.

Desde una perspectiva computacional, la opacidad en el funcionamiento interno de los modelos de la IAG representa un desafío para la interpretabilidad. Comprender el proceso de toma de decisiones de la inteligencia artificial es esencial para construir confianza entre profesionales de la salud y pacientes. La gestión de grandes cantidades de datos médicos para el entrenamiento de modelos de IAG también demanda atención especial a la seguridad y privacidad, protegiendo la información sensible del paciente.

En el ámbito médico, la validación rigurosa y la mejora continua de la precisión son esenciales para garantizar la confiabilidad de los diagnósticos generados por la IAG. La colaboración efectiva entre la IAG y los profesionales de la salud es crucial, pues maximiza el impacto positivo en el diagnóstico

médico, al combinar la experiencia humana con la capacidad de razonamiento clínico de la inteligencia artificial.

Desde una perspectiva holística, el desarrollo e implementación de la IAG en el diagnóstico médico debe priorizar un enfoque centrado en el paciente, y garantizar que las decisiones y recomendaciones generadas sean comprensibles, transparentes y beneficiosas para el bienestar del paciente se convierte en un imperativo ético. A pesar de los desafíos, la IAG representa un paso audaz hacia el futuro de la medicina, donde la inteligencia artificial y la experiencia humana convergen para brindar una atención médica excepcional.

Bibliografía

- Cox Jr LA. Pushing Back on AI: A Dialogue with ChatGPT on Causal Inference in Epidemiology. In *AI-ML for Decision and Risk Analysis: Challenges and Opportunities for Normative Decision Theory*; 2023: 407-423. Cham: Springer.
- Eysenbach G. The role of ChatGPT, generative language models, and artificial intelligence in medical education: a conversation with ChatGPT and a call for papers; 2023. *JMIR Medical Education* 9(1): e46885.
- Kuroiwa T, Sarcon A, Ibara T, Yamada E, Yamamoto A, Tsukamoto K, Fujita K. The potential of ChatGPT as a self-diagnostic tool in common orthopedic diseases: exploratory study. *Journal of Medical Internet Research*. 2023. 25: e47621.
- Mehnen L, Gruarin S, Vasileva M, Knapp B. ChatGPT as a medical doctor? A diagnostic accuracy study on common and rare diseases. *medRxiv Preprint*; 2023. Disponible en: <https://www.medrxiv.org/content/10.1101/2023.04.20.23288859v2.full>
- Miserandino C. The spoon theory. In *Beginning with disability*. Routledge; 2017: 174-178.
- Vallverdú J. Challenges and Controversies of Generative AI in Medical Diagnosis. *Euphyia*; 2023. 17(32), 88-121.

**Injusticia algorítmica
en salud pública**

Aspectos éticos asociados al uso de algoritmos

Alicia de Manuel

Investigadora del Observatorio de Ética e Inteligencia Artificial de Catalunya, Universitat de Girona

1. Introducción

La atención médica y la salud pública no han permanecido ajenas a la creciente integración de la Inteligencia Artificial (IA) en la mayoría de sectores. La aplicación de algoritmos de IA en la toma de decisiones clínicas y en la gestión de datos son algunos de los usos que se le está dando a esta tecnología y que prometen revolucionar la eficiencia y la precisión en el diagnóstico y el tratamiento de enfermedades. Esta auténtica revolución tecnológica que se vio impulsada de alguna manera debido a crisis mundiales como la de la COVID-19,^a se encuentra, sin embargo, rodeada de desafíos críticos que deben tenerse en cuenta. En particular, hablaremos sobre la problemática de la injusticia algorítmica derivada del uso de la IA en el contexto de la salud pública, explorando de qué manera estos sistemas pueden influir en la equidad y accesibilidad de los servicios médicos. A medida que los algoritmos ayudan a los profesionales de la salud en la toma de decisiones críticas, la asignación de recursos y la formulación de diferentes políticas de salud, es necesario examinar de qué manera estos procesos se pueden ver influenciados por los sesgos inherentes en los datos de entrenamiento, así como los retos éticos y sociales que subyacen a estas tecnologías.

^a Este *paper* está basado en la investigación que llevamos a cabo durante el proyecto «Detección y eliminación de sesgos en algoritmos de triaje y localización para la COVID-19», financiado por el BBVA, en el que investigamos la detección y minimización de los posibles sesgos existentes en los sistemas algorítmicos utilizados en el triaje de pacientes y en los sistemas de detección de contactos utilizados durante la crisis sanitaria de la COVID-19. El proyecto puede consultarse a través de la web <https://airesearch-sesgos-covid19.com/>.

2. Aplicación de la IA en la salud pública

Como bien decíamos, la inclusión de la IA en múltiples sectores de la sociedad ha poblado también, lógicamente, la salud pública. Factores como la reciente crisis de la COVID-19 desencadenaron la creciente demanda y aplicación de soluciones basadas en la IA para la ayuda a la toma de decisiones, identificación temprana de pacientes en riesgo, desarrollo de medidas de control y mitigación de contagio, identificación de focos de infección, etc. Esto no significa que anteriormente no se utilizaran los sistemas de inteligencia artificial en el cuidado de la salud, ya que diferentes estudios han demostrado cómo el uso de la IA ha ayudado a realizar el triaje médico en aquellas situaciones en las que los recursos sanitarios son escasos. Y debido a esto, el uso de esta tecnología en un momento de crisis global hizo que se acelerara un proceso ya iniciado.

En ese sentido, esa aceleración de la implantación tecnológica en el campo de la salud, la democratización del uso de la tecnología y la modernización de los sistemas han favorecido la aplicación de la IA en la salud pública. Sin embargo, frente a los evidentes beneficios de la utilización de esta tecnología, debemos hablar sobre los riesgos y los retos éticos que se incorporan con su uso.

3. ¿Cuáles son los principales aspectos éticos asociados al uso de algoritmos de la IA en medicina?

Buena parte de los algoritmos de la IA usados en el cuidado de la salud tienen como función principal tratar de determinar el riesgo de agravamiento en la salud de un paciente, la probabilidad de que una persona sufra una enfermedad, etc., a partir de información clínica y de otro tipo, como, por ejemplo, variables sociodemográficas. Si bien es cierto que en la mayoría de los casos esta asignación probabilística que ofrece el algoritmo no constituye un diagnóstico médico en sentido estricto, sí que se trata de un juicio automatizado que, desde una perspectiva ética, puede presentar diversos riesgos para las personas. Entre ellos destacamos:

1. **Errores de juicio:** el juicio automatizado puede ser erróneo por diversas razones, como por ejemplo el «sobreajuste» u *overfitting*, relativo a los datos utilizados para entrenar el modelo, o puede ocurrir que obtengamos un falso positivo –el o la paciente no tiene la supuesta enfermedad, pero el algoritmo diagnostica que sí– o un falso negativo –el algoritmo declara que la persona no tiene la enfermedad, pero en realidad sí la tiene–. Estos errores de juicio pueden producir efectos negativos en el bienestar físico y emocional de la persona, al inducir a acciones fútiles que además conllevan riesgos asociados –por ejemplo, una intervención innecesaria– o la omisión de tratamientos necesarios. Por otra parte, estos errores de juicio automatizado podrían afectar a la autonomía decisional de los pacientes, que se ven privados de participar en la toma de decisiones sobre el tratamiento.
2. **Errores de diagnóstico:** este tipo de falsos positivos o falsos negativos, a diferencia de los que pueden suceder por la interpretación automática de una prueba concreta –como una PCR– o de un algoritmo, se deben a la interpretación que hace un profesional de la salud a partir de los resultados de dichas pruebas o algoritmos. Estos falsos resultados repercuten tanto en el bienestar de los pacientes como en el del propio personal sanitario, porque pueden entorpecer la labor diagnóstica y contribuyen a la «fatiga por alertas». Las consecuencias de un error de diagnóstico pueden ser múltiples; desde reacciones adversas banales hasta errores que llevan al fallecimiento del paciente. Por su importancia, es imprescindible que la tecnología siempre se utilice como una ayuda a la toma de decisiones, es decir, que exista un sistema de revisión o de cuestionamiento por parte del personal, que priorice su propio criterio clínico por encima del resultado que arroje el algoritmo.
3. **Privacidad:** la información sobre los pacientes o usuarios queda registrada en una base de datos y, si no se ponen las medidas adecuadas de anonimización y seguridad, podría volverse accesible a terceros o personas no autorizadas. Gracias a las distintas normativas europeas, como el GDPR o la futura regulación en materia de inteligencia artificial denominada AI Act, las empresas y organismos públicos cuentan con una serie de normas y obligaciones sobre el uso de los datos de pacientes y usuarios.
4. **Errores por la descontextualización del paciente:** los datos se descontextualizan de los cuerpos y de las condiciones de vida, que se vuelven irrelevantes para el algoritmo. Esto tiene implicaciones éticas notables sobre cómo se entienden comúnmente la salud y los cuidados, la enfermedad y sus tratamientos, situados en contextos sociales y culturales, y con una visión tan holística como sea posible de la salud. Esto lo veremos más adelante, cuando hablemos sobre los determinantes sociales de la salud.
5. **Sobresaturación de información:** los profesionales sanitarios pueden acabar con una sobresaturación burocrática si hay una introducción excesiva de datos para leer e interpretar, así como constantes alertas emergentes en la pantalla –la «fatiga por alertas» del punto 2–. Ello puede desembocar en una malinterpretación de la identidad profesional y de las competencias clínicas que haga que la responsabilidad hacia el paciente se transforme en responsabilidad hacia los datos.
6. **Introducción de nuevos actores en los procesos de toma de decisiones:** la aplicación de algoritmos de inteligencia artificial puede suponer la intervención de nuevas entidades –por ejemplo, empresas– en procesos tradicionalmente definidos, coordinados y ejecutados por profesionales sanitarios y gestores públicos. Garantizar que los criterios que guían la práctica médica también guíen la entrada de proveedores de servicios de la IA puede ser un reto.
7. **Sobrediagnóstico:** la utilización de sistemas de IA puede conllevar una mayor predicción de posibles enfermedades que, sin embargo, jamás se desarrollarán en la persona, pero que la someterán a un mayor número de pruebas diagnósticas, cribados o tratamientos que pueden empeorar otros cuadros clínicos existentes.
8. **Sesgo de automatización y complacencia:** la disponibilidad de herramientas de automatización en la toma de decisiones puede

llevar a una dependencia excesiva de estas ayudas y sistemas de apoyo en detrimento del propio juicio profesional. En ningún caso el equipo médico debe confiar ciegamente en las propuestas generadas por algoritmos, prescindiendo de su propio criterio. Asumiendo que estos sistemas son imperfectos, el personal sanitario puede cometer todo tipo de errores de juicio y razonamiento, tanto de omisión, como de comisión, no solo técnicos, sino también éticos. Este fenómeno psicológico se ve exacerbado por la forma exagerada en que se promocionan actualmente las capacidades de las aplicaciones de inteligencia artificial.

9. **Justicia y equidad:** este es uno de los problemas éticos más relevantes en la IA y uno de sus principales principios (Jobin *et al.*, 2019). En el campo del cuidado de la salud, nos referimos a aquellas situaciones en las que los datos de entrenamiento del algoritmo que tienen poca o ninguna representación de ciertos grupos sociales o determinadas características de la condición bajo estudio exacerbaban las disparidades ya existentes en salud y llevan a juicios sesgados. Por ejemplo, la baja cantidad de imágenes de pieles oscuras aportadas para entrenar un *software* de detección de melanomas provoca peores resultados en el análisis de personas afrodescendientes (Buolamwini, 2017). Otro ejemplo es la presencia de «correctores raciales» en la priorización de acceso a los trasplantes de órganos. Se trata de factores que se añaden, a menudo erróneamente, para corregir la puntuación obtenida según el color de piel, asumiendo una falsa variabilidad del funcionamiento de órganos internos y provocando desigualdades raciales (Simonite, 2020). Debemos decir que, tal y como pone de manifiesto el proyecto AI FORA, los algoritmos de la IA no son universales, sino que están vinculados a valores culturales y éticos dependientes del contexto. Los problemas de la IA relacionados con la justicia y la equidad han sido tradicionalmente infravalorados en un contexto de salud. Por ello, es necesario el establecimiento de mecanismos para asegurar que las decisiones en las que se basan sean justas y no contengan algún tipo de sesgo que haga que el sistema discrimine a ciertos grupos sociales y, por tanto, no se produzca una injusticia algorítmica.

4. ¿Qué entendemos por injusticia algorítmica y sesgo algorítmico?

La injusticia algorítmica en modelos de aprendizaje se refiere a la aplicación de principios éticos y de equidad en el desarrollo y uso de algoritmos de aprendizaje automático, es decir, en el uso de algoritmos que no produzcan daños morales o materiales a los usuarios, ya sea debido a la violación de privacidad, la falta de transparencia o a la aparición de sesgos dentro del algoritmo.

Podemos entender sesgo como la tendencia a beneficiar a un grupo y/o perjudicar a otro de manera injusta o éticamente inadmisibles. En los sistemas informáticos, se refiere a errores sistemáticos que crean resultados injustos, como, por ejemplo, privilegiar a un grupo de usuarios frente a otro. Los sesgos algorítmicos que se pueden dar a lo largo de todo el ciclo de vida de un sistema de IA, suponen la amplificación de los sesgos sociales de género, raza, etnia, sexualidad, etc. y pueden conducir a una discriminación sistemática e injusta en una amplia variedad de situaciones. Es importante destacar que, aunque hablemos de «sesgo algorítmico», el problema puede no estar en el algoritmo en sí mismo, sino en la forma en que se han recopilado los datos que el algoritmo usa para construir el modelo, en un error en la implementación del algoritmo como un programa concreto, o en darle un uso específico para el que no estaba originalmente diseñado, entre otros.

En definitiva, decimos que un sistema está sesgado y/o que un resultado está sesgado cuando una predicción o recomendación es injusta, independientemente de si el resultado es o no estadísticamente correcto. Diremos que un sistema de IA que recomienda más a un hombre que a una mujer para un puesto de alta dirección está sesgado, aunque los cálculos se basen en la realidad del mundo de la alta dirección, en la que la mayoría son todavía hombres. En el caso de la salud puede suceder que el algoritmo infradiagnostique a las mujeres de infarto de miocardio porque está entrenado para dar más importancia a la sintomatología prevalente en hombres –dolor en el pecho y el brazo izquierdo frente al dolor en el epigastrio en mujeres–, debido a una subrepresentación de las mujeres en el modelo algorítmico. Debemos indicar que los sesgos a los que nos referimos son primordialmente sesgos no intencionados, es decir, que aparecen y se manifiestan en y a través de la actividad

computacional al reflejar y reproducir injusticias y desigualdades existentes en la sociedad.

La sobrerrepresentación o subrepresentación también puede afectar a colectivos socioeconómicos. Así pues, un algoritmo está sesgado si favorece o perjudica a ciertos colectivos cuya salud está condicionada por circunstancias sociales desiguales e injustas. Decimos que un algoritmo de acceso a la UCI está injustamente sesgado cuando tiende a desfavorecer a una persona enferma cuya esperanza de vida es menor debido a condiciones sociales de salud desiguales e injustas. Incluso si en apariencia está tratando a todas las personas candidatas por igual en función de su estado de salud actual. Como no tienen en cuenta que el estado de salud está determinado por condiciones sociales injustas que perjudican sistemáticamente a ciertos colectivos, dichos algoritmos actúan como refuerzo de la inequidad, perpetuando, e incluso exacerbando, la injusticia y las desigualdades sociales en la salud.

Por ejemplo, durante la investigación que llevamos a cabo con la finalidad de explorar los sesgos en algoritmos de triaje para la COVID-19, descubrimos que las disparidades raciales se hacen muy evidentes en la aplicación de algoritmos para la toma de decisiones, debido a las diferencias históricas y el impacto de los determinantes de la salud (Delgado *et al.*, 2022). Especialmente en un contexto como el de EE. UU., donde las comunidades negras y latinas estuvieron muy afectadas durante la COVID-19. En el artículo de Williams *et al.* (2020) acerca de la utilización de algoritmos «daltónicos» –en inglés *colorblind*– en referencia a ciegos de color o a la identificación de razas dentro de los datos, se puede ver cómo los sesgos históricos son un factor determinante en la asignación de recursos críticos en emergencias sanitarias.

La idea de utilizar algoritmos que eliminen factores susceptibles de ser sesgados, como por ejemplo la raza, propone una manera equitativa de proporcionar un criterio de distribución, en este caso para la creación de un protocolo en la asignación de respiradores durante la pandemia. Sin embargo, lo que se pudo observar es cómo los sesgos están arraigados profundamente en los datos, y el hecho de eliminar ese dato de la base de datos no elimina el problema. Williams comenta cómo históricamente los pacientes de raza negra acceden a los servicios sanitarios más tarde que las personas de raza blanca debido a las desigual-

dades económicas, el acceso limitado a los recursos, etc., con lo que llegan con dolencias más severas y, por tanto, con unas probabilidades de fallecer más altas. De esta manera, estas cuestiones aparecen reflejadas en las bases de datos como sesgos históricos, pasando a formar parte de la toma de decisiones.

5. ¿Qué factores debemos tener en cuenta para prevenir la aparición de sesgos algorítmicos cuando se aplica la IA en los sistemas de salud?

Como hemos dicho anteriormente, los sesgos se pueden producir a lo largo de todo el ciclo de vida de un modelo de IA, empezando por la fase de propuesta de valor y adquisición de datos, entrenamiento, monitorización y despliegue. Algunos de los sesgos que debemos tener en cuenta son:

1. El **sesgo codificado en los datos**. Sucede cuando las personas responsables de generar, etiquetar el conjunto de datos o crear el modelo, lo hacen de forma sesgada inconscientemente. De esta manera, aquellos sesgos que tenemos como personas y partes de una sociedad, repercuten inconscientemente en la manera de ordenar y catalogar los datos. Debido a que los algoritmos de aprendizaje automático están diseñados con precisión para ajustarse a los datos que reciben, es inevitable que reproduzcan estos sesgos. Así, por ejemplo, en el desarrollo de un algoritmo diseñado para decidir si un paciente debe ingresar en la UCI o no, si los datos se toman sobre decisiones anteriores en las que se discriminaba a las personas de edad avanzada, el resultado reflejará, muy probablemente, un sesgo hacia este colectivo. Este sesgo ocurre con mucha facilidad en equipos que no son diversos.
2. El **uso de sustitutos o proxies**, es decir, de variables que sustituyen a otras que no pueden utilizarse por razones operativas o legales. En ciertos contextos, como en recursos humanos, el uso de la edad como criterio de selección está oculto para evitar discriminaciones por razón de edad o edadismo. Sin embargo, la edad de una persona se puede inferir en la mayoría de los casos a partir de otros datos, como el año en que se graduó

en la universidad, etc. En el caso de la salud, un sistema que use un indicador de riesgos para asignar tratamientos a pacientes podría utilizar el número de consultas o pruebas realizadas como sustituto de la severidad de la enfermedad. Esto podría provocar que colectivos que ya son discriminados al recibir menos atención fueran doblemente discriminados al subestimarse su riesgo y se priorizara a grupos que ya se benefician de un mayor acceso a prestaciones sanitarias.

3. **Sobreajustar un modelo** a la población mayoritaria para minimizar el error promedio. Una base de datos realmente equitativa es una quimera. Inevitablemente, las diferentes comunidades tendrán diferentes distribuciones en relación con las propiedades que queremos modelizar. Si el sistema se entrena principalmente con datos de un tipo de comunidad y el algoritmo está diseñado para minimizar el error promedio, el sistema cometerá más errores al intentar etiquetar a individuos pertenecientes a minorías. Es el caso del diagnóstico de enfermedades de la piel antes mencionado: el algoritmo lo hace relativamente bien con personas de origen caucásico, pero comete errores importantes con personas de otras etnias porque su base de datos está sobreajustada frente a un tipo de paciente.
4. Debemos tener en cuenta que **no existe un modelo universal**. Las poblaciones son sistemas dinámicos, de modo que es necesaria la continua revisión de los modelos para adaptarse a los cambios.

6. ¿De qué mecanismos disponemos para minimizar los sesgos e injusticias en un algoritmo de IA?

Frente a los retos éticos y aquellos factores que pueden suponer la aparición de sesgos, surgen dudas sobre cómo minimizar los sesgos o injusticias que se produzcan en un algoritmo de IA.

En primer lugar, debemos asegurar que el algoritmo represente el mayor número de grupos sociales relevantes en nuestra base de datos en un número suficiente para evitar la aparición de sesgos y el etiquetado incorrecto de

miembros de minorías por la falta de casos suficientes. Los problemas de tener una sobrerrepresentación de las mayorías o de los grupos socialmente favorecidos aumentan cuanto más compleja es la tarea que queremos asignar al algoritmo. Para prevenir los sesgos, nos podemos guiar de forma general asegurando que el sistema tiene en cuenta los vectores tradicionales relacionados con injusticias y desigualdades estructurales (género, edad, sexo, discapacidad, nivel educativo, color de piel, nacionalidad, religión, clase social, lugar de residencia, orientación sexual, etc.).

En segundo lugar, tenemos que asegurarnos de que el etiquetado de los datos utilizados durante la creación de las bases de datos y el entrenamiento del modelo sea preciso, detallado, fiable y sin discriminaciones a primera vista. Tanto si se trata de datos que van a ser etiquetados desde cero como si están producidos por algoritmos previos y vienen ya etiquetados, debemos asegurarnos de que el etiquetado no está contaminado por algún tipo de prejuicio o dinámica estructural de la sociedad. Aquí debemos considerar no solo el proceso del etiquetado, sino también las propias etiquetas utilizadas. Existen diversos mecanismos técnicos que nos permiten aumentar la calidad del proceso de etiquetado y descubrir futuros problemas, desde la revisión de las listas de etiquetas a incluso utilizar el requisito de que los datos sean etiquetados por varias personas pertenecientes a colectivos diversos. No obstante, puede ser complejo descubrir sesgos en el etiquetado, ya que estos pueden ser poco evidentes o resultar de la combinación de diferentes etiquetas individualmente no discriminatorias. En cualquier caso, se debe promover la transparencia sobre la metodología seguida en cada una de las fases del diseño, para que esta información sea pública y, además, se explique de forma clara y sencilla.

En tercer lugar, y relacionado con el punto anterior, debemos conocer con seguridad que las personas responsables de preparar la base de datos y entrenar al algoritmo entienden bien dos aspectos clave: por un lado, las características biomédicas y las sociales de las diferentes comunidades que el algoritmo quiere abarcar; por otro, la relevancia de los determinantes sociales de la salud (DSS) que entran en juego. Este es el aspecto menos analizado y reconocido en el proceso de establecer la equidad de un algoritmo de IA.

Por último, es recomendable trabajar mediante enfoques colaborativos que involucren de forma participativa, y no meramente consultiva o nominal, a la

sociedad civil y en especial a los colectivos y grupos que pueden verse afectados por el sistema que se quiere desarrollar.

En ese sentido, y con el objetivo de ayudar a aquellos agentes involucrados en el diseño e implementación de sistemas de inteligencia artificial en el ámbito de la salud a detectar, así como eliminar la injusticia algorítmica y contribuir en la toma de conciencia del uso ético de la tecnología, creamos una *Guía de autoevaluación ética de uso de la IA en salud* dentro del marco de desarrollo del proyecto «Detección y eliminación de sesgos en algoritmos de triaje y localización para la COVID-19». Esta guía se encuentra en el formato de un formulario o listado de preguntas que evalúa el sistema en todas las fases de su ciclo vital y contempla áreas como la justicia, la equidad y las dinámicas participativas, los errores de juicio, la adecuación con la práctica clínica o la transparencia y la privacidad, y se encuentra disponible de manera *online* en <https://airesearch-sesgos-covid19.com/>.

7. ¿De qué forma podemos mejorar la equidad de un algoritmo de IA?

Tradicionalmente, un algoritmo de inteligencia artificial se diseñaba teniendo en mente la fiabilidad y exactitud de su predicción. Es decir, se buscaba que el algoritmo se acercara de la manera más precisa al resultado correcto –positivos y negativos verdaderos– y se minimizara la diferencia entre el resultado real y el predicho por la máquina. Sin embargo, podemos incluir otras métricas que nos ayuden a juzgar la bondad del algoritmo, como por ejemplo una que nos ayude a medir la mejora de la equidad a la hora de representar los resultados. Aquí deberíamos tener en cuenta qué son los determinantes sociales de la salud (DSS), ya que juegan un papel fundamental en el estado de salud de las personas, esto es, las condiciones sociales, económicas y físicas del lugar donde una persona nace, vive, estudia, trabaja, etc., que pueden afectar a su salud, su bienestar y su calidad de vida. Por ejemplo, usar un algoritmo que tuviera en cuenta de qué manera afectan los DSS podría resultar útil para tomar decisiones más justas a la hora de priorizar a personas en una UCI con recursos limitados. Eso implicaría que la decisión de priorización estaría formada por criterios éticos y no solo clínicos, a fin de evitar decisiones que perpetúen ciertas injusticias.

Asimismo, existe la posibilidad de combinar datos sintéticos –es decir, datos artificiales generados por el propio algoritmo– con datos reales para proteger la privacidad. Así, el proceso de reidentificación se hace prácticamente imposible. También se puede optar por usar solo datos sintéticos, lo que no implica la posible violación de la privacidad de ningún usuario. Sin embargo, el precio de incluir ruido mediante datos ficticios es la reducción de la exactitud y la pérdida de fiabilidad en los resultados. Por ello, desarrollar algoritmos más éticos implica repensar qué usos vamos a dar al algoritmo y qué aspectos de este valoramos más para llegar a compromisos y acuerdos. Se puede objetar que un campo como el de la salud pública no es el lugar para compensar las injusticias sociales y que el personal sanitario responsable no tiene, entre sus obligaciones profesionales, la de evitar mayores injusticias sociales. Sin embargo, no es posible eludir el componente ético de determinadas decisiones sanitarias que priorizan a personas para usar servicios escasos que afectan a su salud. La idea de que las decisiones de priorización, o de racionamiento de recursos entre personas, son éticamente neutras porque solo tienen en cuenta elementos clínicos o de eficiencia en el uso de los recursos limitados es ilusoria. Por ello, ante la aplicación de la IA en cualquier campo, y más si se trata del cuidado de la salud, se debe tomar conciencia de que las decisiones de priorización que hacen los algoritmos sesgados por raza, género, discapacidad o nivel socioeconómico que el personal utiliza están reproduciendo, o incluso incrementando, DSS y podrían ser injustas.

8. Conclusiones

En conclusión, la integración acelerada de la IA en la salud pública, especialmente impulsada durante la crisis de la COVID-19, promete transformar la eficiencia y precisión del diagnóstico y tratamiento de enfermedades. Sin embargo, este avance tecnológico no está exento de desafíos éticos críticos. La aplicación de algoritmos en la toma de decisiones clínicas y la gestión de datos presenta riesgos –errores de juicio y diagnóstico, problemas de privacidad– y sesgos, lo que puede impactar negativamente en la equidad y accesibilidad de los servicios médicos. La injusticia algorítmica, asociada a la falta de principios éticos y equidad en el desarrollo y uso de algoritmos, junto con

el sesgo algorítmico, que produce resultados injustos, son fenómenos que requieren nuestra atención. No hay un único camino a la hora de prevenir los sesgos en los sistemas de IA, pero se pueden abordar a través de la representación diversa en los datos, el etiquetado preciso o la comprensión de los determinantes sociales de la salud. En definitiva, la conciencia de estos desafíos y la implementación de mecanismos para minimizar sesgos e injusticias son esenciales para lograr una aplicación ética y equitativa de la inteligencia artificial en el campo de la salud pública.

Bibliografía

- AI FORA. Disponible en: <https://www.ai-fora.de/> [Accedido el 25 de enero de 2024].
- Buolamwini J. Gender Shades: Intersectional Phenotypic and Demographic Evaluation of Face Datasets and Gender Classifiers. MIT Master's Thesis; 2017. Disponible en: <https://www.media.mit.edu/publications/full-gender-shades-thesis-17/> [Accedido el 25 de enero de 2024].
- Delgado J, de Manuel A, Parra I, Moyano C, Rueda J, Guersenzvaig A, Puyol A. Bias in algorithms of AI systems developed for COVID-19: A scoping review. *Journal of Bioethical Inquiry*; 2022. 19(3), 407-419.
- Jobin A, Ienca M, Vayena E. The global landscape of AI ethics guidelines. *Nature Machine Intelligence*; 2019. 1. 10.1038/s42256-019-0088-2.
- Moss E, Metcalf J. High tech, high risk: Tech ethics lessons for the COVID-19 pandemic response. *Patterns*; 2020. 1(7): 100102.
- Simonite T. How an Algorithm Blocked Kidney Transplants to Black Patients. *Wired, Business*. (2020, 26 de octubre). Disponible en: <https://www.wired.com/story/how-algorithm-blocked-kidney-transplants-black-patients/> [Accedido el 25 de enero de 2024].
- Williams JC, Anderson N, Mathis M, Sanford III E, Eugene J, Isom J. Color-blind algorithms: Racism in the era of COVID-19. *Journal of the National Medical Association*; 2020. 112(5), 550-552.

La injusticia social relacionada con el uso de la inteligencia artificial

Miguel Ángel Royo

Director del Máster de Salud Pública y Jefe de Estudios de la Escuela Nacional de Sanidad

Introducción

La epidemiología, como disciplina científica de la salud pública, es el estudio de la aparición y distribución de sucesos, estados y procesos relacionados con la salud en poblaciones específicas, incluyendo el estudio de los determinantes que influyen en esos procesos y la aplicación de este conocimiento para controlar los problemas de salud pública.¹ La epidemiología extrae el conocimiento de los datos de la población, mediante la aplicación de análisis estadísticos, para llevar a cabo alguna de las siguientes tareas:²

1. Descripción: proporcionar información sumaria de los fenómenos relacionados con la salud y sus posibles determinantes.
2. Clasificación: clasificar a los individuos en categorías de las variables de exposición y enfermedad (diagnóstico).
3. Predicción: predecir la evolución futura de los fenómenos relacionados con la salud y sus determinantes en los individuos (pronóstico).
4. Inferencia causal: explicar las causas de los fenómenos relacionados con la salud.

La clasificación de las tareas de la epidemiología, y de la ciencia de datos en general, para generar conocimientos científicos, nos puede ayudar a entender mejor de qué hablamos cuando hablamos de inteligencia artificial. Los inicios de la IA, en el siglo pasado, partieron de la idea de crear máquinas inteligentes, capaces de pensar como los humanos y suplantarlos en una charla a ciegas.³ Las primeras aproximaciones se basaron en la IA simbólica, que consiste en incorporar datos y reglas de comportamiento a los programas

informáticos, una estrategia muy exitosa en entornos en los que las reglas están muy claras, como en una partida de ajedrez. De ahí que la supercomputadora Deep Blue terminase por vencer a Garry Kasparov, el campeón del mundo de ajedrez, en una partida lenta.⁴ Es decir, codificamos las reglas en un programa informático y lo alimentamos con datos hasta obtener las respuestas a los problemas planteados. Pero esto tiene poco que ver con la complejidad del mundo real, y con la de la salud pública en particular, donde las reglas están muy lejos de estar tan claras o su codificación puede resultar prácticamente inviable.

Para intentar resolver el problema de las reglas y su codificación, se planteó que fueran los propios programas informáticos los que descubrieran las reglas, lo que se conoce como aprendizaje automático.⁵ La idea es codificar un programa informático capaz de aprender por sí mismo –algoritmo de aprendizaje automático– y alimentarlo con datos etiquetados –que incluyen las respuestas–, a fin de ajustar un modelo con el que etiquetar futuros datos. Esto permite abordar dos de las tareas de la epidemiología: predicción y clasificación. Por ejemplo, alimentamos un algoritmo con imágenes de tomografía computerizada de pacientes con y sin cáncer de pulmón, junto con otros datos de interés –genéticos, biológicos, clínicos, sociodemográficos, ambientales...–, para que ajuste un modelo con el que clasificar (diagnosticar) a pacientes con sospecha diagnóstica, o alimentamos un algoritmo con datos del pronóstico a 5 años de supervivientes a un infarto de miocardio para que ajuste un modelo con el que predecir el pronóstico de futuros pacientes. A diferencia de la IA simbólica, en el aprendizaje automático proporcionamos los datos y las respuestas al programa, de modo que este puede obtener por su cuenta las reglas que aplicaremos para resolver el problema. El método de aprendizaje automático descrito es de tipo supervisado. Sin embargo, también se pueden alimentar los algoritmos sin proporcionar las respuestas, para que estos exploren por sí mismos, mediante las correlaciones estadísticas presentes en los datos, las características que permiten clasificar el conjunto de datos en categorías separadas. Los sistemas no supervisados permiten aprender cosas nuevas sobre el problema en cuestión. Por ejemplo, detectar matices no perceptibles al ojo humano o nuevas categorías de tumores. Sin embargo, requieren más recursos y son menos

fiables que los supervisados, por lo que apenas se usan en el ámbito de la práctica clínica.

El concepto «inteligencia artificial» se utiliza para referirnos a los sistemas de aprendizaje automático. No hay duda de que son artificiales, pues se trata de una creación humana, pero es mucho más cuestionable que sean inteligentes. Aunque desarrollan tareas de aprendizaje mediante mecanismos similares a la forma en que los niños, y algunas especies de animales no humanos, aprenden tareas básicas, lo cierto es que los sistemas de aprendizaje automático están muy lejos de lo que comúnmente entendemos por inteligencia humana, ya que no pueden generar ningún conocimiento que no esté contenido en los datos con los que son entrenados, ni explicar el porqué de los resultados que generan.⁶ Es decir, carecen de las capacidades básicas para realizar inferencia causal, una tarea esencial de la epidemiología propia de la inteligencia humana. La inferencia causal conlleva un análisis de los posibles sesgos que contienen los datos o que se pueden producir durante el procesamiento y análisis de los mismos, al objeto de prevenir su aparición o, cuando esto no sea posible, controlar sus efectos. Además, la inferencia causal utiliza fuentes de información ajenas a los datos –concordancia entre estudios, verosimilitud biológica, temporalidad, experimentación– a las que es preciso aplicar reglas de decisión de forma independiente, para después integrar todos esos conocimientos en un modelo coherente y verosímil de explicación de los resultados.⁷

Redes neuronales

Las redes neuronales son algoritmos complejos de aprendizaje automático supervisado. Sus unidades elementales, denominadas neuronas, se estructuran en capas neuronales. Cada neurona de la primera capa recibe una o varias entradas de datos, cuya suma ponderada se hace pasar por una función no lineal (función de activación) que produce un resultado que se hace pasar a su vez, junto con los resultados de otras neuronas, a una neurona de la siguiente capa, y así sucesivamente, hasta la capa de salida, donde los resultados de las neuronas confluyen en una única neurona que genera el resultado del modelo.

Durante su entrenamiento, la red neuronal analiza patrones de datos y sus correlaciones estadísticas, hasta dar con los pesos más adecuados para una solución óptima al problema: la predicción final con el mínimo error posible.⁵

Las redes neuronales han supuesto un salto cualitativo en el desarrollo de los sistemas de aprendizaje automático. Aumentar el número de capas neuronales y pasar cantidades enormes de datos a través de las mismas, nos ha permitido obtener resultados que antes parecían inalcanzables, especialmente en los campos de visión artificial y procesamiento del lenguaje natural, este último con un enorme potencial para abordar problemas de salud pública. Los algoritmos basados en redes neuronales con múltiples capas y datos masivos se denominan sistemas de aprendizaje profundo. Aunque suponen un gran avance, no es a coste cero. Al aumentar la profundidad del sistema, los algoritmos de aprendizaje automático terminan por convertirse en cajas negras que, alimentados con datos masivos, nos devuelven predicciones que pueden ser muy precisas, pero mantienen ocultos los mecanismos del modelo para alcanzar su resultado. Esto genera problemas de interpretabilidad y explicabilidad, pues la dificultad para comprender el proceso de toma de decisiones de un modelo (las causas o los porqués de las mismas), limita a su vez la capacidad para comunicar esas decisiones a los demás, especialmente a los profanos. La interpretabilidad tiene muchas implicaciones para la seguridad, porque permite, por un lado, analizar el riesgo de que el objetivo planteado colapse con otros aspectos que no estamos considerando en el estudio y, por otro lado, determinar si el sistema se está fijando en características relevantes o en correlaciones espurias, detectando sesgos en los que no habíamos pensado *a priori*. Las correlaciones espurias son características reales que el sistema usa para tomar su decisión, a pesar de que son irrelevantes para el problema en cuestión. Por ejemplo, un algoritmo de visión artificial diseñado para detectar neumonía en pacientes de COVID-19 alcanzó una sensibilidad cercana al 100%. Analizando las imágenes con un mapa de calor para intentar entender el funcionamiento del algoritmo, se comprobó que el sistema se fijaba en la zona de la clavícula para tomar su decisión, lo cual se debía a que los afectados de neumonía se encogían de hombros por su dificultad para respirar.⁵ Las correlaciones espurias abren la puerta a engañar al sistema y pueden tener consecuencias imprevisibles en el futuro.

Aplicaciones de la IA en salud pública

La IA desempeñó un papel relevante en la gestión de la pandemia de la COVID-19. BlueDot, una empresa emergente canadiense que utiliza un algoritmo de procesamiento de lenguaje natural para la vigilancia de enfermedades transmisibles en el mundo, fue la primera en alertar de un brote de enfermedad respiratoria en la provincia de Wuhan -9 días antes de la notificación del brote por parte de la Organización Mundial de la Salud-,⁸ y fue también la primera en publicar el artículo científico que predijo el potencial de propagación internacional del virus por tráfico aéreo.⁹ BlueDot recopila datos cada 15 minutos, 24 horas al día, a partir de declaraciones oficiales de salud pública, informes de salud en animales, medios digitales, datos de billetes de avión a nivel mundial y datos demográficos de cerca de un centenar de países, que luego son revisados por médicos, epidemiólogos y programadores informáticos. La IA se ha utilizado para rastrear la propagación del SARS-CoV-2, predecir sus impactos, acelerar el desarrollo de vacunas y optimizar la asignación de recursos.¹⁰ Aprendiendo de la experiencia con la COVID-19, los sistemas de IA se pueden refinar y mejorar para prevenir y controlar futuras pandemias. Al entrenar los sistemas de IA con datos pandémicos, podemos desarrollar modelos predictivos más precisos, mejorar los sistemas de detección temprana y la respuesta a futuras crisis de salud. La IA se ha implementado en el diseño de sistemas avanzados de vigilancia y alerta temprana, como el Programa de Monitoreo de Enfermedades Emergentes (ProMED) de la Sociedad Internacional de Enfermedades Infecciosas (<https://promedmail.org/>). ProMED es un sistema global de vigilancia en línea que extrae artículos de noticias, informes oficiales y otras fuentes digitales, para identificar posibles brotes de enfermedades en tiempo real. La IA también puede desempeñar un papel en la previsión de los impactos socioeconómicos de las pandemias, ayudando en estrategias de preparación más integrales.

Las posibles aplicaciones de la IA no se limitan a los sistemas de vigilancia, alerta y respuesta temprana de brotes de enfermedades, o a la preparación frente a futuras pandemias, sino que abarcan prácticamente todos los ámbitos de la salud pública. Sin ánimo de exhaustividad, destacamos las siguientes:¹¹

1. Sistemas de vigilancia de los determinantes de la salud, especialmente los comerciales. Mediante estrategias de control del conocimiento, la narrativa y las reglas de juego relativas al comercio y la publicidad, las empresas que comercializan productos y servicios perjudiciales para la salud han adquirido un poder sin precedentes para modelar los hábitos de consumo de la población mundial.¹²
2. Modelado predictivo para el diagnóstico precoz de enfermedades y otros fenómenos relacionados con la salud: cáncer, enfermedades cardiovasculares, soledad no deseada, salud mental, etc.
3. Identificación de patrones y tendencias ocultas en diversos fenómenos relacionados con la salud y sus determinantes.
4. Análisis y modificación de comportamientos relacionados con la salud –hábitos mentales, salud sexual y reproductiva, alimentación, tabaquismo, actividad física...– por medio de datos generados por el consumidor en redes sociales y aplicaciones informáticas. Con datos de la red social del *Framingham Study*, se observó que la obesidad se transmite a través de contactos sociales, con un patrón similar al que se esperaría de una enfermedad transmisible.¹³
5. Formulación de políticas de salud, desarrollo de intervenciones, optimización de la asignación de recursos y priorización de uso en base a predicciones sobre el balance riesgo-beneficio y la rentabilidad.
6. Mejora de la comprensión de las complejas relaciones entre genética, estilos de vida, medioambiente y fenómenos relacionados con la salud y sus determinantes.

Sesgos en los sistemas de aprendizaje automático

En el ámbito de la IA, habitualmente se entiende por sesgo la falta de coincidencia entre las distribuciones ideales y reales de las etiquetas y los atributos de los usuarios en el entrenamiento y aplicación de un sistema;¹⁴ es decir, que la predicción del modelo no se corresponde con el resultado del parámetro estudiado en la vida real.¹⁵

Otros autores hablan de sesgo cuando el resultado del algoritmo es sistemáticamente menos favorable para individuos de un grupo, sin que haya diferencias relevantes entre grupos que justifiquen esos daños; es decir, como una propiedad potencialmente dañina de los datos, que resulta en una injusticia.¹⁶ Independientemente de la definición usada, para poder identificar y prevenir sesgos resulta esencial su caracterización. A tales efectos, podemos clasificar los sesgos de los sistemas de IA en dos grandes grupos: los sesgos estadísticos y los sesgos inherentes al algoritmo.

Sesgos estadísticos

La mayor parte de la investigación en salud se ha llevado a cabo tradicionalmente en hombres jóvenes, blancos y educados de países industrializados, ricos y democráticos (*Young WEIRD Man*, por sus siglas en inglés).¹⁷ Sin embargo, las personas responden de forma diferente ante el riesgo o el hecho de perder la salud, tanto a nivel biológico como social, en función de características como el sexo, la edad, el nivel de estudios, la clase socioeconómica, la etnia o el lugar de procedencia, entre otras. Por ejemplo, ante un infarto de miocardio, las mujeres presentan con más frecuencia dolor abdominal, dolor de cuello y disnea, síntomas denominados atípicos porque se utiliza como patrón de referencia del infarto de miocardio el cuadro clínico propio de los hombres.¹⁸ A consecuencia de ello, las pacientes, su entorno y hasta los profesionales, tienen más dificultades para identificar un infarto de miocardio, lo que conlleva errores o demoras diagnósticas que empeoran el pronóstico de la enfermedad en las mujeres. Por tanto, si entrenamos un algoritmo para el diagnóstico de infarto de miocardio con datos procedentes de hombres o de muestras donde estos estén sobrerrepresentados, el sesgo está servido. Lo mismo ocurre cuando utilizamos un algoritmo entrenado con personas de piel blanca para el diagnóstico de melanoma en personas de piel negra.⁵ Los patrones de comportamiento relacionados con la salud en un entorno determinado o en respuesta a ciertas intervenciones de salud pública, también presentan diferencias importantes en función de características como la edad, la etnia, el nivel de estudios o la situación económica.¹⁹ Los sesgos estadísticos se producen, por tanto, cuando los datos con los que se entrena un algoritmo –la muestra de estudio– no son representativos de la población en la que se

utilizará dicho algoritmo, lo que en epidemiología se define como un sesgo de selección.

Muchos sistemas de procesamiento de lenguaje natural han utilizado fuentes periodísticas con una larga tradición, cuyos textos han sido elaborados por un perfil muy concreto de personas, típicamente hombres *WEIRD*, por lo que contienen información latente sobre ese grupo demográfico, codificada en señales lingüísticas propias que provocan un sesgo de selección por sobreajuste del modelo. Puesto que el lenguaje de esos periodistas durante las décadas en las que escribieron sus textos difiere del usado por la mayoría de la población en la actualidad, y especialmente por los jóvenes y las minorías, el modelo funciona peor en estos grupos (no comprende bien su lenguaje), que quedan en cierto modo excluidos, lo que puede dar lugar a diversos tipos de discriminación.

Sesgos propios de los algoritmos

Los algoritmos de visión artificial o de procesamiento del lenguaje natural no solo son susceptibles a los sesgos de selección, sino que además plantean desafíos únicos relacionados con sesgos inherentes al algoritmo, que difieren de los sesgos típicos de los estudios epidemiológicos. Los algoritmos de visión artificial utilizan redes neuronales convolucionales, con múltiples aplicaciones en el ámbito clínico, médico y quirúrgico.⁵ Los algoritmos de procesamiento del lenguaje natural, en cambio, utilizan redes neuronales recurrentes no supervisadas, especializadas en trabajar con secuencias. Mediante la búsqueda de asociaciones estadísticas en las secuencias de palabras, detectan patrones en los corpus lingüísticos que contienen información gramatical y semántica. Por sus múltiples aplicaciones en salud pública, desde la vigilancia epidemiológica hasta la formulación de políticas de salud, pasando por el estudio y modificación de los comportamientos relacionados con la salud, en este apartado nos centraremos en los sesgos potenciales de los algoritmos de procesamiento del lenguaje natural.²⁰

Los sistemas de información sanitaria funcionan con distintas fuentes de datos, con formatos diferentes y semánticas antiguas, que no siempre están organizados según un modelo o esquema preestablecido. Cuando los datos

no tienen un lenguaje común o una estructura bien definida, como la información de las historias clínicas –datos no estructurados–, son más propensos a generar errores durante el proceso de recolección y etiquetado. Estos errores, denominados sesgo del etiquetado, son más frecuentes cuando, en lugar de usar a expertos entrenados para recolectar y etiquetar los datos, se encarga esta tarea a un grupo numeroso de personas no especializadas mediante una convocatoria abierta (*crowd outsourcing*). Este sesgo se puede prevenir mediante colaboración abierta de expertos en diferentes campos –salud, ética, lingüística, antropólogos, etnólogos, etc.– y el entrenamiento de los encargados de recolectar y etiquetar los datos.

Los sistemas de procesamiento del lenguaje natural asignan a cada palabra un vector numérico que contiene información semántica, lo que permite asociar palabras insertas en determinados contextos gramaticales. Estas asociaciones de palabras incorporan los sesgos humanos implícitos en los textos que las contienen, que reflejan actitudes discriminatorias hacia determinados grupos, provocando un sesgo semántico o de representación; por ejemplo, la asociación de las palabras «inmigración» y «delincuencia». La prevención de este sesgo mediante la eliminación de determinadas asociaciones de palabras en los textos no es factible, ya que esto supondría suprimir el contexto esencial del texto, necesario para interpretar el mundo que este representa de forma coherente y para hacer uso del mismo en aplicaciones posteriores. Lo que sí puede hacerse es tomar conciencia de las amenazas de la IA para la sociedad y, en base a ese conocimiento, diseñar estrategias para lograr equidad en la toma de decisiones relativa a las aplicaciones posteriores de los sistemas de IA.

Los lenguajes evolucionan continuamente, por lo que incluso un modelo entrenado con una muestra representativa capta solo una instantánea, y los modelos entrenados con datos de muestras no representativas tienden a exacerbar los sesgos cuando se aplican a nuevos conjuntos de datos. La validez de los modelos se puede contrastar con métodos de interpretabilidad y con análisis de sensibilidad para comprobar la robustez de los resultados en varios conjuntos de datos o subgrupos demográficos.

El sesgo de investigación es un sesgo de selección, en la medida en que los grupos vulnerables o marginados no están representados en los equipos de

investigación. Y el hecho de que el inglés sea el idioma más usado por los sistemas de procesamiento natural del lenguaje también puede dar lugar a sesgos. Por ejemplo, el modelo *n-gram* –puramente estadístico–, se basa en la suposición de que la probabilidad de la siguiente palabra en una secuencia depende solo de una ventana de tamaño fijo de palabras anteriores. Este modelo funciona bien con el inglés, pero no tanto con otros idiomas. Aplicar una metodología que tenga presente el riesgo de sesgo y los posibles mecanismos para evitarlo o minimizarlo durante todo el ciclo de la investigación es la mejor estrategia para su prevención.

Injusticia social relacionada con la IA

No es sorprendente que los algoritmos reflejen prejuicios sociales, ya que son una creación de los seres humanos, que somos propensos a una interpretación sesgada de la realidad. La discriminación puede pasar desapercibida porque la mayoría de las personas tenemos dificultad para reconocer nuestros propios sesgos –punto ciego–. En la IA, nuestros propios prejuicios individuales se combinan con los de otras personas a gran escala, se vuelven sistemáticos –deterministas–, se reproducen mediante bucles de retroalimentación y se amplifican en una escala masiva, como cuando se usan los textos generados con aplicaciones de IA sesgadas para entrenar nuevas aplicaciones, exacerbando y acelerando las desigualdades existentes.

Los sesgos en los sistemas de IA provocan con frecuencia injusticias sociales, pero se pueden revertir cuando son detectados y eliminados. En cierta medida, los sesgos se pueden prevenir si se tienen en cuenta durante el diseño de los algoritmos, pero es esencial que los árboles no nos impidan ver el bosque. Además de reproducir y amplificar los prejuicios sociales presentes en la sociedad, las lógicas que impulsan la IA están alimentando prácticas que generan profundas injusticias sociales. La lógica de aumentar la capacidad de las redes neuronales para mejorar el rendimiento de los algoritmos –maximalismo informático– es una práctica común que se basa en un consumo creciente de energía, en la extracción masiva de materias primas –elementos de tierras raras– con gran impacto ambiental, y en el uso de mano

de obra barata, mediante la explotación laboral de personas que trabajan lejos de los grandes centros de computación, con frecuencia en países empobrecidos, extrayendo minerales y etiquetando datos en condiciones precarias.²¹ Además, los sistemas de la IA utilizados por grandes compañías para mejorar su eficiencia se sostienen en base a la deshumanización de sus trabajadores, que son tratados como robots, cerrando así el círculo de la explotación laboral.^{22,23}

Las decisiones sobre la información que alimenta los sistemas de IA y los criterios de clasificación de los datos son de carácter más político y social que técnico, e implican una visión del mundo que tiende a reforzar las estructuras de poder y a exacerbar las desigualdades sociales, ya que los sistemas de IA están diseñados para beneficiar a los Estados y las grandes corporaciones que están detrás de los mismos.²⁴ El uso de algoritmos de reconocimiento emocional en sistemas de seguridad privados y estatales para detectar posibles delincuentes o terroristas, pese a la falta de pruebas científicas sólidas que avalen tal uso,²⁵ es un buen ejemplo de una decisión basada en cuestiones ideológicas, económicas y de control social, con graves consecuencias para millones de personas, que no solo no han dado su consentimiento para la extracción de sus datos, sino que ven vulnerados sus derechos sin la debida justificación.

La lógica de que cualquier dato que se pueda recolectar está ahí para ser extraído, ya sea para un uso concreto o para un potencial uso futuro, y la tendencia actual a publicar datos en abierto, representan una grave amenaza a la privacidad, puesto que abren la puerta a identificar a millones de personas –incluso cuando los datos están anonimizados– y a obtener información sensible de las mismas. El uso masivo de datos para tareas de vigilancia y control de las personas no solo se ha vuelto cada vez más sofisticado, mediante una lógica de puntuación moral –algoritmos de clasificación crediticia–, sino que se ha trasladado del ámbito militar y de inteligencia –vigilancia de potenciales terroristas– al ámbito civil: seguridad de fronteras y sistemas de justicia criminal o penal. En el ámbito militar, por ejemplo, el ejército de Israel está haciendo uso del sistema de IA Habsora –Evangelio–, que permite detectar objetivos de Hamás en la franja de Gaza a un ritmo vertiginoso, lo cual está causando la muerte de civiles en una escala sin precedentes. Esta

mortandad se debe tanto a errores del sistema como a la presión creciente que reciben los militares de Israel para alcanzar el máximo número de objetivos, y se ve favorecida por la continua detección automática de los mismos, asumiendo bajas civiles incluso cuando se trata de objetivos de bajo rango, como simples simpatizantes de Hamás.²⁶ En el ámbito civil, las personas vulnerables que ya sufren algún tipo de discriminación son más proclives a este tipo de vigilancia policial, por lo que el uso de algoritmos de inteligencia artificial para estos fines genera bucles de retroalimentación: las personas que ya están incluidas en un sistema de vigilancia, precisamente por ser vulnerables, son a su vez sometidas a un mayor escrutinio.²⁷ La última vuelta de tuerca en este proceso consiste en usar algoritmos de vigilancia para el acceso a prestaciones públicas, con graves consecuencias para los ciudadanos. En el Estado de Michigan, un algoritmo diseñado para detectar estafadores al seguro de desempleo del Estado identificó erróneamente a más de 40.000 sospechosos de fraude, cuyos salarios fueron embargados, tras imponerles sanciones varias veces superiores a la supuesta deuda que tenían, llevando a muchos a la bancarrota.²⁸ En España, el Ministerio de Seguridad Social implantó en 2018 un algoritmo de IA para rastrear bajas laborales y detectar posibles fraudes.²⁹ La forma en que se ha diseñado y desplegado este sistema, cuyos usuarios califican de opaco, es un ejemplo de *mala praxis*, que va en contra de las guías para un uso responsable de la inteligencia artificial en el sector público:³⁰ sin ningún tipo de consulta previa a las personas implicadas y sin una evaluación rigurosa del sistema, no solo de sus pobres capacidades técnicas –sensibilidad del 65% y especificidad del 25%–, sino de la forma en que se preveía desarrollar y utilizar.

La mentalidad de extracción en relación con los datos se está convirtiendo en un imperativo moral, con menosprecio de valores como el consentimiento y la compensación.³¹ El uso de esos datos masivos en redes neuronales profundas, cada vez más difíciles de interpretar incluso por sus propios creadores, pero supuestamente capaces de reducir la complejidad del mundo y proporcionar soluciones universales a prácticamente todos los problemas de la vida, ha envuelto estos sistemas en un aura mistificadora que aboga por su inevitabilidad. La fantasía de que los sistemas de IA son inmateriales, una especie de cerebros incorpóreos alojados en una nube alejada de los recursos materiales

y humanos que se precisan para producirlos –dualismo de abstracción y extracción–, contribuye a esa aura mística idolatrando la innovación técnica.³² El uso de trabajadores precarios que realizan tareas de etiquetado que, supuestamente, de cara al usuario son realizadas de forma automática por un sistema de IA,²¹ contribuye al imaginario de la inmaterialidad de la IA. Es lo que algunos autores han denominado *determinismo encantado*,³³ para el que importa más la naturaleza innovadora del sistema que su propósito, un fenómeno que oscurece las estructuras de poder subyacentes, dificulta los debates informados e impide la participación de la población en la toma de decisiones, culpando a los algoritmos de los resultados, como si estos fueran agentes autónomos, independientes de las lógicas de extracción y de poder que se ocultan detrás de los mismos, y como si bastase con perfeccionar los modelos para acabar con las injusticias sociales relacionadas con la IA. Estos imaginarios contribuyen a bloquear iniciativas de regulación que resultan imprescindibles para impedir las amenazas de la IA y ponerla al servicio del bien común, asegurando un balance riesgo-beneficio positivo que reduzca las desigualdades sociales contribuyendo a un mundo más justo.

Notas

1. Porta M, Greenland S, Hernán M, dos Santos Silva I, Last JM, editores. A dictionary of epidemiology. Sexta edición. Oxford: Oxford University Press; 2014.
2. Hernán MÁ, Hsu J, Healy B. A Second Chance to Get Causal Inference Right: A Classification of Data Science Tasks. *Chance*. 2009; 32.1:42-9.
3. Turing A. Computing Machinery and Intelligence. *Mind*. 195; 59:433-60.
4. Sevilla J. 25 años de la icónica victoria de un ordenador sobre Kaspárov. *The Conversation*, 15 de febrero de 2021. Disponible en: <https://theconversation.com/25-anos-de-la-icónica-victoria-de-un-ordenador-sobre-kasparov-155225>
5. Cobo M, Lloret L. *Inteligencia artificial y medicina*. Madrid: CSIC, 2023.

6. Larson EJ. El mito de la Inteligencia Artificial. Barcelona: Shakleton Books, 2022.
7. Royo Bordonada MA, Damián J, García López F, editores. Epidemiología básica. Madrid: Escuela Nacional de Sanidad - Instituto de Salud Carlos III - Ministerio de Ciencia, Innovación y Universidades, 2023.
8. Miller E. An AI Epidemiologist Sent the First Warnings of the Wuhan Virus. WIRED, 25 de enero de 2020. Disponible en: <https://www.wired.com/story/ai-epidemiologist-wuhan-public-health-warnings/>
9. Bogoch L, Watts A, Thomas-Bachli A, Huber C, Kraemer MUG, Khan K. Pneumonia of unknown aetiology in Wuhan, China: Potential for international spread via commercial air travel. *J Travel Med.* 2020; 27:2020:taaa008.
10. Wang L, Zhang Y, Wang D, et al. Artificial Intelligence for COVID-19: A Systematic Review. *Front. Med.* 2021;8:704256.
11. García S, D'Agostino M. La inteligencia artificial en la salud pública. Organización Panamericana de la Salud, 2021. Disponible en: <https://iris.paho.org/handle/10665.2/53887>
12. McKee M, Stuckler D. Revisiting the Corporate and Commercial Determinants of Health. *Am J Public Health.* 2018;108(9):1167-70.
13. Christakis N, Fowler JH. The spread of obesity in a large social network over 32 years. *N Engl J Med.* 2007;357(4):370-9.
14. Shah DS, Schwartz HA, Hovy D. Predictive biases in natural language processing models: a conceptual framework and overview. Proceedings of the 58th annual meeting of the Association for computational linguistics; Online. Stroudsburg, PA, USA: Association for Computational Linguistics, 2020:5248-64.
15. Amann J, Blasimme A, Vayena E, et al. 2020. Explainability for artificial intelligence in healthcare: A multidisciplinary perspective. *BMC Medical Informatics and Decision Making.* 2020;20(1): 310.
16. Delgado, de Manuel A, Parra I, et al. Bias in algorithms of AI systems developed for COVID-19: A scoping review. *J Bioeth Inq.* 2022; 19(3):407-19.
17. Henrich J, Heine SJ, Norenzayan A. Most people are not WEIRD. *Nature.* 2010;466:29.
18. Roffi M, Patrono C, Collet J. Ph, et al. 2015 ESC Guidelines for the management of acute coronary syndromes in patients presenting without persistent ST-segment elevation: Task Force for the Management of Acute Coronary Syndromes in Patients Presenting without Persistent ST-Segment Elevation of the European Society of Cardiology (ESC). *Eur Heart J.* 2016;37:267-315.
19. Pampel FC, Krueger PM, Denney T. Socioeconomic disparities in health behaviors. *Annu Rev Sociol.* 2010;36:349-70.
20. Hovy D, Prabhumoye S. Five sources of bias in natural language processing. *Lang Linguist Compass.* 2021;8:e12432.
21. Huet E. The Humans Hiding Behind the Chatbots. Bloomberg, 18 de abril de 2016. Disponible en: <https://www.bloomberg.com/news/articles/2016-04-18/the-humans-hiding-behind-the-chatbots>
22. Ajunwa I, Crawford K, Schultz J. Limitless Worker Surveillance. *California Law Review.* 2017;105: 735-76.
23. Rodríguez P. El algoritmo que penaliza tus datos de productividad por ir al baño: así es el sistema que Amazon va a cambiar ante las denuncias por lesiones laborales. *Xakata*, 2 de junio de 2021. Disponible en: <https://www.xataka.com/pro/algoritmo-que-penaliza-tus-datos-productividad-ir-al-bano-asi-sistema-que-amazon-va-a-cambiar-denuncias-lesiones-laborales>
24. Crawford K. Atlas de IA. Poder, política y costes planetarios de la inteligencia artificial. Ned ediciones; 2023.
25. Heaven D. Why faces don't always tell the truth about feelings. *Nature.* 2020;578(7796):502-4.

26. Abraham Y. Una fábrica de asesinatos en masa. Ctxt, 5 de diciembre de 2023. Disponible en: <https://ctxt.es/es/20231201/Politica/44958/Yuval-Abraham-972-Magazine-Local-Call-gaza-muertos-palestina-habsbora-ia-inteligencia-artificial.htm>
27. Brayne S. Big Data Surveillance: The Case of Policing. *American Sociological Review*. 2017;82:977-1008.
28. Richardson R, Schultz JM, Southerland VM. Litigating Algorithms: 2019 US Report. AI Now Institute, 2019. Disponible en: <https://ainowinstitute.org/wp-content/uploads/2023/04/litigatingalgorithms-2019-us.pdf>
29. Jiménez P, Méndez MA, Márquez R. La Seguridad Social usa una IA secreta para rastrear bajas laborales y cazar fraudes. *El Confidencial*, 17 de abril de 2023. Disponible en: https://www.elconfidencial.com/tecnologia/2023-04-17/seguridad-social-ia-inteligencia-artificial-inss-bajas-empleo-algoritmos_3611167/
30. Leslie D. Understanding artificial intelligence ethics and safety: A guide for the responsible design and implementation of AI systems in the public sector. The Alan Turing Institute, 2019. Disponible en: <https://doi.org/10.5281/zenodo.3240529>
31. Sadowski J. When data is capital: Datafication, accumulation, and extraction. *Big Data & Society*. 2019;6:1-12.
32. Hardt M, Negri A. *Assembly*. Nueva York: Oxford University Press; 2017.
33. Campolo A, Crawford K. Enchanted Determinism: Power without Responsibility in Artificial Intelligence. *Engaging Science, Technology, and Society*. 2020;6:1-19.

Aportaciones de expertos

«Iceberg» de la inteligencia artificial y las desigualdades en la salud

Brenda Biaani León-Gómez y Pere Torán-Monserrat

Fundació Institut Universitari per a la Recerca a l'Atenció Primària de Salut Jordi Gol i Gurina

Los avances en la Inteligencia Artificial (IA) tienen el potencial de transformar la sociedad. Estos cambios pueden ser muy obvios, o no. Algunos cambios obvios pasan por la modificación de las formas de acceso a la salud, el desarrollo de herramientas de diagnóstico preciso o los tratamientos personalizados. La implementación de la IA en la sociedad redefinirá no solo la atención médica, sino que irá mucho más allá y tendrá implicaciones y efectos no tan obvios. Todo ello se verá acelerado por la mejora del tratamiento y manejo de datos que se ha ido produciendo durante décadas. La IA ofrece ahora la oportunidad de un mayor análisis y estudio de los datos, lo que tiene el potencial de ayudar a encontrar nuevas soluciones a problemas de salud históricos –como las pandemias o enfermedades crónicas–.¹

En medio del potencial de innovación que esto representa, surgen planteamientos en torno a cómo la implementación de la IA en el mundo, y no solo en el área de la sanidad, puede tener un efecto en la salud de las personas. La Organización Mundial de la Salud (OMS) ha establecido principios para guiar el desarrollo y la implementación de estas tecnologías en el sector sanitario con el objetivo de que se utilicen de manera beneficiosa para la sociedad.² En torno a esto ya se plantean múltiples desafíos éticos y prácticos.³ Un primer desafío es el riesgo de que la brecha de desigualdad social se haga mayor. Este riesgo puede estar asociado a múltiples causas, entre las que cabe destacar el acceso desigual a la tecnología o la distribución desigual de la integración de estas herramientas en la sociedad. Por ejemplo, en años pasados se ha observado el efecto que ha tenido internet en el aumento de las desigualdades salariales entre distintos grupos poblacionales –por ejemplo entre diferentes países y territorios, urbanos y rurales– en los que el acceso ha sido atemporal. El acceso diferenciado a internet ha condicionado el acceso

diferenciado a la información y a la educación digital, y se ha convertido en un factor fundamental en la definición de los trayectos de vida de pueblos y de comunidades enteras. Abordar estos temas es fundamental para asegurar que los beneficios de la IA sean accesibles a todas las personas, independientemente de su género, ubicación geográfica, estatus socioeconómico, etnia, estatus migratorio, antecedentes culturales, etc. Sin embargo, los desafíos a enfrentar no solo están relacionados con el acceso a la tecnología, sino también al modo en que esta tecnología puede modificar o incidir sobre los diversos determinantes sociales de la salud, definidos por la OMS como «las circunstancias en que las personas nacen, crecen, trabajan, viven y envejecen, incluido el conjunto más amplio de fuerzas y sistemas que influyen sobre las condiciones de la vida cotidiana». Estas fuerzas y sistemas incluyen políticas y sistemas económicos, programas de desarrollo, normas y políticas sociales y sistemas políticos,⁴ y los determinantes sociales incluyen las condiciones en las que las personas nacemos, crecemos, trabajamos y vivimos, y están influenciados por la tecnología.

Estos determinantes de la salud pueden verse modificados ante la implementación de la IA en múltiples áreas, como, por ejemplo, el mercado laboral. La potencial reconfiguración del mercado laboral dada la automatización de procesos por la IA⁵ puede conllevar, entre otras cosas, a la automatización de tareas repetitivas, así como al cambio de habilidades requeridas para el desempeño de trabajos que tradicionalmente se han llevado a cabo de una sola forma. Estos cambios representarán diferencias en el acceso a recursos –educativos y laborales, por ejemplo– que generarán cambios en el mercado laboral. Por ejemplo, en el sector de la salud se requerirán profesionales que puedan usar estas tecnologías. En este sentido, la IA podría aumentar la brecha entre trabajadores cualificados y aquellos con menos habilidades. Más allá de lo anterior, las herramientas de la IA pueden utilizarse para analizar grandes conjuntos de datos y modificar o reformular políticas, como las laborales, que beneficien a grupos sociales. En relación a esto, hay un riesgo de sesgo⁶ debido a la potencial infrarrepresentación de poblaciones en situación de vulnerabilidad en los datos utilizados para entrenar a la inteligencia artificial. Estas poblaciones pueden verse afectadas por la falta de representatividad en los datos utilizados para entrenar sistemas, lo que podría resultar en

información inexacta y conclusiones inexactas o injustas para ciertos grupos. Esto es similar al sesgo de género en los ensayos clínicos. La infrarrepresentación de grupos en situación de vulnerabilidad en los datos de entrenamiento de la IA no solo es injusta, sino que también puede resultar peligrosa, dado que la IA puede generar soluciones para algunos grupos, mientras que para otros no.

Así mismo, la tecnología podría exacerbar la disparidad en las desigualdades, influenciando determinantes intermedios de la salud. En relación con esto, tanto el acceso como la utilización efectiva de las tecnologías emergentes pueden estar desigualmente distribuidos entre distintos estratos. Como resultado de todo ello, las poblaciones con menor acceso a recursos tecnológicos podrían enfrentarse a desventajas adicionales en términos de salud y bienestar, amplificando así la brecha de desigualdad ya existente. Esto es debido a que las tecnologías digitales se han vuelto cada vez más importantes para acceder a recursos como la vivienda, la educación o los cuidados. Por ejemplo, para acceder a múltiples servicios, como citas en la atención primaria a través de una aplicación *online*, las personas necesitan tener dispositivos inteligentes –que tienen un coste–, una conexión a internet estable –que también tiene un coste– y habilidades digitales. Estos requisitos excluyen a una parte significativa de la población, especialmente a aquellos con bajos ingresos o a personas mayores que no están familiarizadas con la tecnología. Ahora imaginemos que existe una limitación previa a este acceso ya limitado *per se*, como por ejemplo que ese sistema de salud tenga que pagar –si es que puede– por la utilización de un servicio de IA para mejorar la experiencia de sus usuarios.

De esta forma, las personas con recursos suficientes –capacidad económica para pagar servicios y habilidades digitales– y ubicadas en sistemas de salud que puedan tener acceso a estos servicios, se beneficiarán de la utilización de la aplicación «mejorada». Estas personas podrán programar citas fácilmente, acceder a su información de salud en cualquier momento, recibir asesoramiento médico, disponer de un servicio de preguntas y respuestas con un chatbot o disfrutar de una mejor gestión de sus servicios por parte de algoritmos sin tener que desplazarse físicamente. Esta accesibilidad mejorará sin duda su experiencia de atención y su bienestar. De igual modo, la falta de

acceso a estas tecnologías reforzará y ampliará la desventaja existente, atrayendo a algunas personas en un círculo vicioso que refuerza las desigualdades en salud.⁷

Declaración de conflicto de intereses

Los autores de este artículo declaran que no tienen conflictos de intereses. Para la realización de resúmenes bibliográficos de ciertos documentos y para criticar el texto de manera interactiva se utilizó GPT-4.

Notas

1. Kumar A, Pujari P, Gupta N. Artificial Intelligence: Technology 4.0 as a solution for healthcare workers during COVID-19 pandemic. *Acta Univ. Bohemiae Merid.* 24; 2021. p. 19-35.
2. World Health Organization. Ethics and Governance of Artificial Intelligence for Health: WHO guidance. World Health Organization; 2021.
3. Guarda P. Ok Google, am I sick?: artificial intelligence, e-health, and data protection regulation. 15, 359-375.
4. PAHO (Organización Panamericana de la Salud. Determinantes sociales de la salud - OPS/OMS. Organización Panamericana de la Salud. *Latinoamerica* 657; 2019).
5. Felten EW, Raj M, Seamans R. The Variable Impact of Artificial Intelligence on Labor: The Role of Complementary Skills and Technologies. *SSRN Electron. J.*; 2019. Disponible en: doi:10.2139/ssrn.3368605.
6. Furman J, Seamans R. AI and the economy. *Innov. Policy Econ.* 19, 161-191; 2019.
7. Baum F, Newman L, Biedrzycki K. Vicious cycles: digital technologies and determinants of health in Australia. *Health Promot. Int.* 29 2; 2014. p. 349-60.

Inteligencia artificial y relación clínica: punto de encuentro

Valle Coronado Vázquez

Profesora de Humanidades Médicas de la Universidad Francisco de Vitoria

La IA ha encontrado en la medicina un campo inmenso en el que desarrollarse. Las empresas de datos tienen el negocio de la salud en el punto de mira. Después de todo, siempre fue una práctica muy rentable. ¿Quién no estaría dispuesto a pagar cualquier precio por estar sano?

El despliegue de la IA en el ámbito sanitario no se ha hecho esperar. Ha desembarcado en la práctica clínica con una nutrida oferta de herramientas para la prevención, el diagnóstico y el tratamiento de las enfermedades y, lo que es aún más rentable, el amplio campo de la promoción de la salud y la medicina predictiva. Mediante dispositivos conectados a nuestro cuerpo, suministramos a las empresas millones de datos sobre distintas funciones biológicas –la glucemia, las horas de sueño, la presión arterial, el ejercicio físico, etc.–, que la IA transforma en una información muy valiosa. Estas potencialidades permitirán una mayor accesibilidad a los servicios de salud, más precisión diagnóstica y terapéutica y la posibilidad de liberar tiempo de los profesionales para practicar una medicina más cercana al paciente. Con esta promesa se hace una llamada a los médicos, para que sean los defensores de la IA. Pero debemos ser cautos, porque, por lo que sabemos, el desarrollo de las tecnologías en la actualidad, lejos de contribuir a establecer una relación más humana con los enfermos y de centrar la medicina en las personas, ha contribuido a instaurar una práctica clínica centrada en las pruebas. Así pues, ¿por qué la IA va a restaurar una relación de confianza entre el médico y el paciente que ya casi se ha perdido?

Podemos imaginar a una IA haciendo diagnósticos con una elevada precisión, sí, pero... ¿podría llegar a sustituir al médico en la relación clínica con los pacientes?

El diálogo con los pacientes durante la consulta y el proceso de deliberación de las decisiones no pueden ser sustituidos por la IA. En estos subyacen valores propiamente humanos, que solo pueden darse en una relación empática, que no podrá ser reemplazada por una máquina.

Además, la IA entraña una serie de problemas éticos relacionados con la privacidad, la confidencialidad y la equidad que no pueden ser obviados.

La posibilidad de extraer y analizar grandes cantidades de datos derivados de la actividad clínica es un problema para la privacidad. El médico ya no es el responsable de la información que le facilitan los pacientes, sino que lo es la industria de lo digital, que genera protocolos accesibles para cualquier persona, con posibilidad de autodiagnóstico y tratamiento: «Doctora, he mirado en internet y tengo todos los síntomas de un autismo leve». Sin duda, se trata de un negocio muy rentable: recolectar datos de las personas a través de dispositivos inteligentes, generar enfermedades y después desarrollar los tratamientos para curarlas.

Además, las fuentes de información de las que se nutre la IA presentan sesgos de información, porque utilizan datos contaminados que pueden perpetuar desigualdades entre las personas. De ahí la preocupación por que determinados grupos de población –que, con frecuencia suelen ser los más desfavorecidos socialmente– se queden fuera de los algoritmos de los sistemas de IA. ¿Quién puede garantizar que esto no ocurra? Si la IA solo beneficia a los que pueden pagar las innovaciones médicas, ¿dónde queda la equidad? La falta de transparencia en los datos es un problema que la IA no tiene resuelto.

Eric Topol, fundador y director del Scripps Research Translational Institute, tiene una visión contrapuesta de todo este asunto y defiende que la IA pone a disposición de las personas más pobres las tecnologías diagnósticas de última generación a través de aplicaciones de teléfonos inteligentes, disminuyendo así las desigualdades. No coincide esta visión optimista con la expresada por la UNESCO en su *Recomendación sobre la Ética de la Inteligencia Artificial*, en la que recomienda promover la justicia social y luchar contra todo tipo de discriminación generada por este tipo de tecnologías.

Luego, como señaló Jordi Vallverdú en una ocasión, se precisa un diálogo entre profesionales de la salud, tecnólogos y bioeticistas para detectar y corre-

gir los sesgos inherentes al uso de la IA en el diagnóstico y la prescripción. Cabe añadir aquí la necesidad de introducir en la formación de pregrado en medicina la enseñanza de la IA, sus beneficios y sus riesgos para la construcción de una medicina más humana.

No hay duda de que la IA puede ser una herramienta de ayuda en las consultas, pero debemos andar con prudencia en este laberinto, no vaya a ser que, deslumbrados por ella, perdamos la esencia de ser médicos.

Inteligencia, artificial y medicina

Bernabé Robles

Neurólogo. Coordinador de la Unidad de Bioética. Parc Sanitari Sant Joan de Déu

Tres palabras: inteligencia, artificial y medicina. A menudo nos preocupan sobre todo las dos primeras, pero quizá deberíamos repensar bien la tercera.

Los y las Sapiens, arrojados por nuestra autoconsciencia de poder, fragilidad y responsabilidad, nos enfrentamos al acrónimo IA –inteligencia artificial– de forma casi inexorable, pero con una mezcla de sensaciones encontradas: desde la curiosidad a la desconfianza.

Ya habíamos sentido emociones similares en el Neolítico, cuando incorporamos a los animales a tareas cotidianas y descubrimos que las hacían con mayor fuerza, rapidez, resistencia –y paciencia– que nosotros. Y ocurrió de nuevo, en mayor grado, en el siglo XVIII, cuando las máquinas comenzaron a sustituir a nuestro sistema osteomuscular y, después, a nuestro sistema extrapiramidal. Las máquinas hacían muchas cosas con más fuerza y rapidez que nosotros, pero también con más precisión. Sentimos entonces un «vértigo» no solo cultural, sino también social y antropológico. En todos estos «hitos» de la civilización, el miedo a ser sustituidos, a convertirnos en seres fútiles, sobrevoló las reflexiones filosóficas y las manifestaciones artísticas, quizá desde un antropocentrismo más instintivo que lógico.

Pero ambos términos, inteligencia y artificial, merecen matizaciones. Esto no va de imitar la inteligencia humana –de hecho, no sabemos qué es eso de la «inteligencia»– sino de resolver problemas o, más bien, de encontrar correlaciones de forma más rápida y eficiente. Por otro lado, no es tan fácil definir «tecnología» o «artificial». Al fin y al cabo, lo artificial está hecho por seres naturales con elementos naturales, y no solo los humanos fabrican edificios o instrumentos. ¿Un dique de castores, o un termitero de tres metros, son naturales o artificiales?

En el fondo, ahora estamos hablando de automatizar decisiones, o, más propiamente, de aportar capacidad de predicción y manejo de datos a las decisiones. La inteligencia artificial permite encontrar patrones que el cerebro humano no puede detectar –o, mejor, no puede «hacer conscientes»–, y nos permite utilizar la estadística «en tiempo real».

No obstante, llegados a este punto es muy importante reparar en que «correlación» no siempre implica «causa». Los algoritmos complejos actuales son extremadamente eficaces detectando correlaciones, pero la relación causal depende de otras atribuciones, más difíciles de programar. Además, en las personas, no solo cuentan las causas, sino también los valores, porque los y las Sapiens participan subjetivamente de la definición de sus fines.

Por tanto, deberíamos intentar no caer en el fundamentalismo de los datos. Los datos son interpretaciones de la realidad y contienen los sesgos y las limitaciones de los aparatos de medida y de las muestras que se obtienen. Los perfiles son representaciones de los sujetos, no son los sujetos y, sobre todo, no representan su subjetividad.

Por supuesto, la potencia estadística en tiempo real de la inteligencia artificial genera la impresión de que estamos desarrollando un instrumento con «caja negra», en el que lo esencial no es transparente. Pero, de hecho, sí que conocemos la fórmula compleja que sustenta los algoritmos, porque la hemos escrito nosotros. Los problemas comienzan cuando no podemos escudriñar el flujo de datos a través de esas fórmulas; en especial, en los sistemas que «autoaprenden» (una denominación, de nuevo, antropocéntrica). Es una fórmula tan complicada que pocos le pueden dar significado a su flujo, pero se trata de una limitación nuestra, no de la red de microperceptrones –a los neurólogos no nos gusta demasiado la desafortunada denominación «redes neuronales»–.

Además, no debemos olvidar que la primera «caja negra» de la inteligencia es nuestro propio cerebro, al menos por el momento. Por otro lado, la subjetividad de nuestra mente es un arma de doble filo, esencial para «personalizar» y detectar valores, pero extremadamente sensible al egoísmo, a los sesgos y a la corrupción. Así pues, la inteligencia artificial, siempre que estuviese democráticamente diseñada, programada y alimentada, podría ser un instrumento

para la justicia. Las decisiones humanas, que ahora interpretamos como el «patrón oro», no son tan ideales, y solo hay que repasar la historia y la sociodemografía humana para entenderlo. La subjetividad puede permitirse particularizar y adaptar las decisiones al contexto, pero también introduce sesgos y conflictos de interés más o menos autoconscientes. Seguramente en muchas decisiones, especialmente las que afectan a la colectividad, una «inteligencia artificial» bien ajustada a aquellos valores colectivos que queramos proteger, sería claramente preferible a las decisiones humanas, a menudo manipulables e interesadas.

Pero la cuestión esencial es el «¿para qué?». Todo instrumento a utilizar para evitar el malestar de los humanos exigiría un consenso social previo sobre usos, límites y oportunidades. Para decisiones sanitarias, son evidentes las aportaciones que podría traernos la «inteligencia artificial». Detecta patrones en océanos de datos, y eso es esencial en medicina. Pero la medicina no va solo de hacer diagnósticos –ni más diagnósticos, ni más precoces–. Los y las Sapiens tenemos muchos defectos, aplicando y repartiendo cuidados, pero los profesionales «humanitarios», deberían combinar calidad y calidez, haciendo diagnósticos y tratamientos «oportunos» y «accesibles».

Sin embargo, basados solo en nuestra mente, como nos recuerda E. Topol, hacemos una medicina superficial, cara e ineficiente. Tenemos profesionales muy preparados, pero solo en algunos aspectos de esta tarea, que trabajan en organizaciones rígidas y fragmentadas, que manejan datos insuficientes de manera poco transparente y que toman decisiones subjetivas y contingentes, más «convenientes» que sustentadas en datos fiables.

Desde el punto de vista individual, a menudo pensamos más en solucionar nuestro conflicto interno «empático» –nuestro sufrimiento al ver padecer a los otros–, que en atender a los verdaderos problemas y necesidades del otro, desde su prisma. Depende de nosotros que la «inteligencia artificial» meramente «sustituya» profesionales sanitarios y devore la medicina, o bien que nos ayude a recuperar el «humanismo» perdido.

Por tanto, como los algoritmos, los diagnósticos también son instrumentos («conocer a través de»). La buena medicina va de hacer diagnósticos y tratamientos oportunos, ajustados a los valores y a la trayectoria vital de cada

persona, y al mismo tiempo de su comunidad, social u ecológica. ¿Están los algoritmos diseñados, alimentados y entrenados desde esa perspectiva?

Por tanto, antes de utilizar «inteligencia artificial» en salud, deberíamos repasar, repensar y redefinir, si fuese preciso, LOS FINES DE LA MEDICINA. Debemos definir si lo que queremos es que la salud se convierta, o no, en un proceso de supervisión/monitorización continua de la preenfermedad, vista esta, hoy por hoy, desde una perspectiva eminentemente biomédica. ¿Basta reparar cuerpos, o hacerles «mantenimiento», para que, hipotéticamente, no «se estropeen», para hacer buena medicina? ¿Puede la inteligencia artificial por sí sola, en el momento actual, generar decisiones «oportunas» basadas en valores?

Humildemente, pienso que, si de verdad queremos generar salud con inteligencia artificial, deberíamos destinar su capacidad de gestión y análisis de datos a mejorar, honestamente, los determinantes sociales de la salud. Y hacer eso puede chocar con los quizá a veces legítimos, aunque no siempre deseables, objetivos empresariales. Las personas que enferman son «sujetos»; no meramente «objetos» de mercado.

No debemos olvidar que los algoritmos, está claro, pueden blanquear, incluso amplificar, sesgos injustos. Pero dichos sesgos han sido producidos, en su origen, por los humanos. La inteligencia artificial aplicada a la salud debe desarrollarse, imbricada con la inteligencia humana, pero con la máxima independencia y transparencia posibles. No debe ser nunca ni un biopoder, ni una estafa moral.

Las personas a las que atendemos se merecen una inteligencia artificial segura, equitativa, accesible y sin intereses espurios, que huya del totalitarismo y de la pura instrumentalización de personas como meros objetos. Si no es así, es que no hemos aprendido nada desde los juicios de Nuremberg.

Por último, y recordando a Warner Slack: *«Si un médico puede ser sustituido por un ordenador, es que se lo merece».*

Relación de autores

- Enrique Bernal-Delgado y Francisco Estupiñán-Romero, Ciencia de Datos para la Investigación en Servicios y Políticas Sanitarias. Instituto Aragonés de Ciencias de la Salud, Zaragoza.
- David Casacuberta, departamento de Filosofía de la Universitat Autònoma de Barcelona.
- Alicia de Manuel, investigadora del Observatorio de Ética e Inteligencia Artificial de Catalunya, Universitat de Girona.
- Francisco Javier García León, epidemiólogo. Asistente de investigación *ad honorem*. Departamento de Metafísica y Corrientes Actuales de Filosofía, Ética y Filosofía Política. Universidad de Sevilla.
- Miguel Ángel Royo, director del Máster de Salud Pública y jefe de estudios de la Escuela Nacional de Sanidad.
- Jordi Vallverdú, investigador ICREA del departamento de Filosofía de la Universitat Autònoma de Barcelona.

Aportaciones de expertos

- Brenda Biaani León-Gómez y Pere Torán-Monserrat. Fundació Institut Universitari per a la Recerca a l'Atenció Primària de Salut Jordi Gol i Gurina.
- Valle Coronado Vázquez, profesora de Humanidades Médicas de la Universidad Francisco de Vitoria.
- Bernabé Robles del Olmo, neurólogo. Coordinador de la Unidad de Bioética del Parc Sanitari Sant Joan de Déu.

Títulos publicados

Cuadernos de Bioética

68. *La necesidad de cuidado: un reto político, social e institucional*
67. *Donación de plasma y altruismo: revisando conceptos*
66. *Eutanasia: los retos jurídicos y administrativos de la LORE*
65. *Vejez, sociedad y salud pública*
64. *Bioética y derecho de la salud pública*
63. *Inteligencia artificial en salud. Retos éticos y científicos*
62. *Soledad no deseada en la era digital*
61. *Discapacidad y equiparación de derechos*
60. *Cuidarse en la sociedad entre pandemias*
59. *La atención a las necesidades sociales y sanitarias, ¿sumamos o dividimos?*
58. *Salud pública y COVID*
57. *Ciudades que cuidan, también al final de la vida*
56. *La salud pública, el género y la ética*
55. *Soledad, envejecimiento y final de la vida*
54. *Sexualidad y diversidad funcional*
53. *Nuevos escenarios en salud e investigación clínica*
52. *Ética y promoción de la salud. Libertad-paternalismo*
51. *Una mirada ética en la gestión de conflictos*
50. *Pensar la maternidad*
49. *Publicidad y salud*
48. *Prioridades y políticas sanitarias*
47. *Ética y donación de plasma: una mirada global*
46. *Comités de Ética y consultores clínicos: ¿complemento o alternativa en la ética asistencial?*
45. *CRISPR... ¿debemos poner límites a la edición genética?*
44. *Crisis y salud mental en niños y jóvenes: ¿causa o consecuencia?*
43. *¿Debemos revisar el concepto de muerte?*
42. *Iatrogenia y medicina defensiva*
41. *Eutanasia y suicidio asistido*
40. *Ethical aspects of research with children*
39. *Discapacidad, nuevos enfoques y retos éticos a la luz de la Convención de la ONU*
38. *Ética, salud y dispendio del conocimiento*
37. *Determinantes personales y colectivos de los problemas de la salud*
36. *Ética y altruismo*
35. *Treinta años de técnicas de reproducción asistida*
34. *Ética de la comunicación corporativa e institucional en el sector de la salud*
33. *Alcance y límites de la solidaridad en tiempos de crisis*
32. *Ética y salud pública en tiempos de crisis*
31. *Transparencia en el sistema sanitario público*
30. *La ética del cuidado*
29. *Casos prácticos de ética y salud pública*
28. *La ética en las instituciones sanitarias: entre la lógica asistencial y la lógica gerencial*
27. *Ética y salud pública*
26. *Las tres edades de la medicina y la relación médico-paciente*
25. *La ética, esencia de la comunicación científica y médica*
24. *Maleficencia en los programas de prevención*
23. *Ética e investigación clínica*
22. *Consentimiento por representación*
21. *La ética en los servicios de atención a las personas con discapacidad intelectual severa*

20. *Retos éticos de la e-salud*
19. *La persona como sujeto de la medicina*
18. *Listas de espera: ¿lo podemos hacer mejor?*
17. *El bien individual y el bien común en bioética*
16. *Autonomía y dependencia en la vejez*
15. *Consentimiento informado y diversidad cultural*
14. *Aproximación al problema de la competencia del enfermo*
13. *La información sanitaria y la participación activa de los usuarios*
12. *La gestión del cuidado en enfermería*
11. *Los fines de la medicina*
10. *Corresponsabilidad empresarial en el desarrollo sostenible*
9. *Ética y sedación al final de la vida*
8. *Uso racional de los medicamentos. Aspectos éticos*
7. *La gestión de los errores médicos*
6. *Ética de la comunicación médica*
5. *Problemas prácticos del consentimiento informado*
4. *Medicina predictiva y discriminación*
3. *Industria farmacéutica y progreso médico*
2. *Estándares éticos y científicos en la investigación*
1. *Libertad y salud*

Informes de la Fundació

6. *La interacción público-privado en sanidad*
5. *Ética y biología sintética: cuatro corrientes, tres informes*
4. *Las prestaciones privadas en las organizaciones sanitarias públicas*
3. *Clonación terapéutica: perspectivas científicas, legales y éticas*
2. *Un marco de referencia ético entre empresa y centro de investigación*
1. *Percepción social de la biotecnología*

Interrogantes éticos

5. *Pedagogía de la Bioética*
4. *Repensar el cuerpo*
3. *La subrogación uterina: análisis de la situación actual*
2. *Afectividad y sexualidad. ¿Son educables?*
1. *¿Qué hacer con los agresores sexuales reincidentes?*

Para más información: www.fundaciongrifols.org

FUNDACIÓ
VÍCTOR
GRÍFOLS
i LUCAS