LÍMITES DE LA INTELIGENCIA ARTIFICIAL EN SALUD DIGITAL

PAULA SUBÍAS-BELTRÁN

EURECAT, CENTRE TECNOLÒGIC DE CATALUNYA,
UNITAT DE DIGITAL HEALTH, BARCELONA

OBSERVATORI DE BIOÈTICA I DRET
UNIVERSITAT DE BARCELONA

WORKING PAPER 3/2021

WORKING PAPERS
JEAN MONNET CHAIR
EUROPEAN PRIVATE LAW

UNIVERSITAT DE BARCELONA

***Abstract**:* Artificial intelligence continues to advance in all aspects of society. In particular, in healthcare, more solutions based on artificial intelligence are used day after day, such as those that support decision-making or those that facilitate telecare. All the problems to be solved have their intricacies and are characterized by their context. Consequently, it is important to identify their complexities to be able to contextualize these solutions in terms of how they should be applied in order to better reason about the expected uncertainties and limitations of their application. This article analyses a number of the concerns related to the limits of the application of solutions based on artificial intelligence, such as technosolutionism, the suitability of the training data sample, data privacy, as well as the evaluation of the limitations of these solutions.

***Title**:* Limits of artificial intelligence in digital health

***Keywords**:* artificial intelligence, ethics, research on digital health, personal data, no discrimination

***Resumen*:** La inteligencia artificial continúa avanzando en todos los aspectos de la sociedad. En la asistencia sanitaria, en particular, día tras día se usan más soluciones basadas en inteligencia artificial, como aquellas que dan soporte en la toma de decisiones o aquellas que facilitan la teleasistencia. Todos los problemas por resolver tienen sus entresijos y se caracterizan por sus contextos. En consecuencia, es importante identificar sus complejidades para luego poder contextualizar dichas soluciones en términos de cómo se van a aplicar para poder razonar mejor sobre las incertidumbres esperadas y las limitaciones de su aplicación. En este artículo se analizan varias de las cuestiones relacionadas con los límites de la aplicación de soluciones basadas en inteligencia artificial, como son el tecnosolucionismo, la idoneidad de la muestra de datos de entrenamiento, la privacidad de estos datos, así como la evaluación de las limitaciones de dichas soluciones.

Título: Límites de la inteligencia artificial en salud digital

Palabras clave: inteligencia artificial, ética, innovación en salud digital, datos personales, no discriminación

SUMARIO

1. INTRODUCCIÓN

Hoy, la inteligencia artificial (IA) es una forma de poder. Se ha utilizado para exponer injusticias, mejorar la calidad de la salud, y avanzar en la innovación. Pero también se ha utilizado para discriminar, controlar, y vigilar. La IA implica oportunidades y riesgos; y no lidiar con estos riesgos conllevaría la propagación de una IA perniciosa. Por eso, debemos impulsar una IA que fortalezca los derechos humanos y nunca los socave.

El término IA se tiende a usar como una sinécdoque haciendo referencia al *todo por una parte*. El todo, en este caso, alude a servicios o soluciones que integran algoritmos de IA cuyo objetivo es dar respuesta a un problema en particular. En la creación de dichas soluciones se deben maximizar los beneficios directos e indirectos para las personas afectadas y se debe minimizar cualquier posible daño a dichas personas. Es necesario tener en cuenta las diferentes fases en las que se trabaja para crear estas soluciones: desde su diseño, donde se define el alcance y las metodologías a aplicar, pasando por el desarrollo, donde se recaban los datos sobre los que se trabajará, hasta su despliegue, donde la solución impacta sobre las personas afectadas.

El campo de la salud digital es paradigmático, ya que los datos de relevancia se consideran sensibles a ojos del Reglamento General de Protección de Datos (RGPD). Es el campo donde las personas son más vulnerables con respecto a sus derechos y donde se magnifican las identidades personales. En este artículo analizaremos algunas cuestiones relacionadas con los límites con los que nos encontramos actualmente en las soluciones basadas en IA con aplicación en el ámbito de la salud digital. En la segunda sección adoptaremos una visión general para abordar las tecnologías emergentes y hablaremos sobre el público al que dan servicio. A continuación, nos centraremos en los cimientos de las soluciones basadas en IA: los datos, y reflexionaremos sobre su representatividad (sección 3), y sobre su privacidad (sección 4). Discutiremos la importancia de evaluar estas soluciones con una visión holística en la sección 5. Y, para acabar, reflexionaremos sobre el futuro de las soluciones basadas en IA y su contexto (sección 6).

2. TECNOLOGÍAS EMERGENTES: ¿INCLUYENTES?

Se acostumbra a usar el término “tecnologías emergentes”, también conocidas como tecnologías convergentes, como referencia a nuevas tecnologías que tienen el potencial de causar una disrupción en el mercado tal como lo conocemos. La definición más conocida es la propuesta por George Day y Paul Schoemaker (Day, Schoemaker, & Gunther, 2000) que las describen como “innovaciones científicas que pueden crear una nueva industria o transformar una existente. Incluyen tecnologías discontinuas derivadas de innovaciones radicales, así como tecnologías más evolucionadas formadas a raíz de la convergencia de ramas de investigación antes separadas”. Se consideran tecnologías emergentes la nanotecnología, la biotecnología, las tecnologías

de la información y la comunicación, la ciencia cognitiva, la robótica, y la IA[1], entre otras. En este artículo nos centraremos en las consideraciones éticas que nacen del uso de soluciones basadas en IA.

Las soluciones basadas en IA, dada su naturaleza instrumental, se embeben en diferentes soportes físicos, como dispositivos móviles, tecnologías vestibles (por ejemplo, relojes inteligentes y cascos virtuales), y otros tipos de computadoras. Este tipo de soluciones presenta oportunidades únicas para diversos sectores, como es el caso de la asistencia sanitaria. La adopción de tecnologías emergentes implica varios retos, pero los beneficios potenciales de nuevas soluciones que puedan mejorar la asistencia ofrecida, así como los procesos actuales deben considerarse seriamente para que este sector siga mejorando y siendo competitivo.

Uno de los retos relacionados con la adopción de soluciones basadas en IA es intrínseco al tecnosolucionismo[2], que considera que las tecnologías digitales son capaces de solucionar de modo efectivo cualquier problema. Un claro contraejemplo que pone en duda la eficiencia que el tecnosolucionismo abraza es el uso de tecnologías digitales en la asistencia sanitaria centrada en las personas mayores.

2.1. PONGAMOS EL FOCO EN LAS PERSONAS MAYORES

En 2050, más de una de cada cinco personas será mayor de 60 años[3]. El envejecimiento de la población puede considerarse un éxito de las políticas de salud pública y el desarrollo socioeconómico, pero también constituye un reto para la sociedad, que debe adaptarse a ello para mejorar al máximo la salud y la capacidad funcional de las personas mayores, así como su participación social y su seguridad. La Organización Mundial de la Salud (OMS) determinó en 2002 estos cuatro como los pilares del envejecimiento activo, al que definió como el "proceso de optimización de las oportunidades de salud, participación y seguridad con el fin de mejorar la calidad de vida a medida que las personas envejecen". Desde entonces han sido muchas las iniciativas que han promovido el desarrollo de soluciones digitales para fomentar el envejecimiento activo[4]. Pero ¿son este tipo de soluciones las más adecuadas para la población objetivo?

[1] La nanotecnología, la biotecnología, las tecnologías de la información y la comunicación, y la ciencia cognitiva se agrupan bajo el término NBIC. NBIC se usa para hablar sobre el estudio interdisciplinario de las interacciones entre sistemas vivos y artificiales y comprende la colaboración sinérgica entre las diferentes disciplinas.

[2] El término tecnosolucionismo lo acuñó Evgeny Morozov y se refiere a querer arreglarlo todo mediante estrategias digitales de cuantificación. Se recomienda la lectura de *Morozov, E. (2015). La locura del solucionismo tecnológico (Vol. 5010). Katz Editores y Capital Intelectual*.

[3] Véase https://www.who.int/es/health-topics/ageing

[4] Referencia a las convocatorias de la Comisión Europea de los últimos años: https://tinyurl.com/H2020-active-ageing

Los posibles beneficios que el uso de nuevas tecnologías puede ofrecer a las personas mayores son muchos y diversos, como, por ejemplo:

a) mantener o aumentar la agilidad mental: la inexistencia de actividad mental está fuertemente ligada a la disminución de la capacidad de aprendizaje en personas mayores. Las nuevas tecnologías facilitan el aprendizaje de nuevos conceptos, lo que aumenta la capacidad cognitiva favoreciendo la agilidad mental de quien las usa, y acaba repercutiendo en la reducción de la incidencia de patologías que cursan con deterioro cognitivo;
b) vencer prejuicios: tanto los propios como los que tiene el resto de la población acerca de la vida de las personas mayores;
c) aumentar la calidad de vida: las nuevas tecnologías ofrecen soluciones que facilitan satisfacer necesidades; y
d) fortalecer la independencia: actualmente hay innumerables recursos ligados a las tecnologías de la información y la comunicación (TIC) que contribuyen a favorecer la autonomía de las personas mayores.

Pero existe un distanciamiento entre las personas mayores y las nuevas tecnologías que se puede deber a las siguientes causas:

a) desconocimiento: muchas personas desconocen para qué pueden utilizarlas y cómo podrían mejorar su día a día;
b) complejidad de uso: a una gran parte de las personas mayores les preocupa que sean demasiado complejas para ellas y no se atreven a utilizarlas por miedo a cometer un error y borrar información o estropear el dispositivo;
c) desinterés e indiferencia: la omisión de las personas mayores de la sociedad de consumo de productos tecnológicos se puede interpretar como que no es un producto dirigido a ellas y generarles, así, desinterés e indiferencia;
d) capacidad económica reducida: el acceso a la última tecnología disponible implica un alto coste que se ve limitado por la capacidad económica insuficiente de las personas mayores.

La evolución de la tecnología ha facilitado nuestra vida en múltiples ámbitos. Sin embargo, ésta no tiene el mismo nivel de aceptación en todos los grupos de edad. Por consiguiente, la Comisión Europea ha invertido recursos en desarrollar soluciones digitales para fomentar el envejecimiento activo con el efecto colateral de reducir la brecha generacional. Estas soluciones digitales nacen de proyectos de investigación e innovación que, por su propia condición, se basan en la última tecnología del mercado, exacerbando, así, el distanciamiento entre las personas mayores y las nuevas tecnologías.

Esta chocante situación nos lleva a preguntarnos: ¿son las tecnologías emergentes incluyentes? El beneficio que estas pueden reportar sobre la sociedad ha sido ampliamente divulgado, pero ¿están todos los sectores de la sociedad preparados para integrarlas en su día a día?

2.2. CONOCIMIENTO A BASE DE INTRUSISMO

Un ejemplo de solución basada en IA son los sistemas de recomendación. Estos sistemas se usan en el ámbito de la asistencia sanitaria como instrumento para conseguir diferentes objetivos. Las dos tipologías de sistemas de recomendación más conocidas son los sistemas que dan soporte a la toma decisiones clínicas (Baig, Gholamhosseini, Connolly, & Lindén, 2015) y las soluciones que permiten facilitar cambios conductuales para mejorar la calidad de vida (Khattak, Pervez, Han, Nugent, & Lee, 2012) (Subías-Beltrán, y otros, 2019), aunque hay muchas más.

Los sistemas de recomendación para personas mayores están a la orden del día. Estos sistemas se basan en la recolección de datos de las personas usuarias para entender sus preferencias, sus rutinas y comportamientos, y su estado de salud (físico, mental, nutricional, etc.). Una vez el sistema conozca a las personas usuarias, este podrá perfilarlas y modelar su comportamiento. De esta manera, el sistema trabajará sobre una abstracción de estas personas que le permitirá adaptarse a ellas con el ulterior fin de maximizar su adhesión al plan de recomendación previsto. Y este es el punto clave subyacente en estos sistemas: la personalización.

La modelización del comportamiento de las personas no es trivial. Las personas somos variantes en nuestras rutinas. Hay muchos factores que influyen en nuestro día a día y que tienen un impacto en las actividades que realizamos: surgen imprevistos personales, no siempre tenemos la misma energía, e incluso las condiciones meteorológicas pueden alterar nuestros planes. Debido a lo cual es imposible que una solución basada en IA sea capaz de modelizar nuestro comportamiento si no tiene acceso a toda esta información.

Pensemos en un hipotético sistema de recomendaciones que pretenda fomentar el seguimiento de una dieta saludable. Para que este sistema ofrezca recomendaciones personalizadas será necesario que el modelo conductual nutricional que el sistema abstraiga de la persona de interés sea lo suficientemente representativo de su comportamiento. De otra manera, el sistema no será capaz de caracterizar a la persona y las recomendaciones ofrecidas no serán las más apropiadas.

Si se pretendiera crear un sistema capaz de sugerir recomendaciones 100% personalizadas, este debería de recolectar información detallada sobre la persona. Por ejemplo, le preguntaría ¿qué le gusta comer? ¿Sigue una dieta especial? ¿Qué presupuesto tiene? ¿Tiene intolerancias alimenticias? ¿Ve la televisión mientras come? ¿Come en compañía? ¿Toma suplementos alimentarios? ¿A qué horas come? ¿Cuánto tiempo querría dedicar a cocinar? Y otras cuestiones más. Para que un sistema cuyo fin se basa en un modelado preciso de la persona usuaria se ajuste a sus rutinas, comportamientos, y preferencias, se necesita mucha información. Pero ¿debería de ser este un requerimiento del sistema para la persona usuaria? El nivel de intrusismo al que se la somete debería de depender de ella misma. Por esta razón, es importante

poner en cuestión la balanza entre precisión e intrusismo. El principio de proporcionalidad[5] debería vertebrar susodicho análisis y este debería dar respuesta a preguntas como: ¿se puede conseguir el mismo objetivo con menos datos? ¿Existe una medida o procedimiento menos invasiva para conseguir el objetivo marcado con la misma eficacia? ¿Se derivan más beneficios para la persona interesada que perjuicios?

3. MUESTRA REPRESENTATIVA: ¿UTOPÍA O REALIDAD?

Las soluciones basadas en IA se soportan por algoritmos. Y estos, a su vez, se alimentan de datos. Los algoritmos analizan estos datos buscando relaciones, patrones, similitudes y diferencias, para luego generalizar. Esta búsqueda de reglas se conoce como la fase de entrenamiento. Los algoritmos se entrenan para resolver una tarea en particular, aprenden de un conjunto de datos concreto, y luego intentan resolver esa misma tarea en un entorno "un poco diferente", es decir, usando un conjunto de datos "un poco diferente" al de entrenamiento. Dicho de otra manera, los dos conjuntos de datos usados no deberían de diferir de forma significativa. El sector de la población que represente este segundo conjunto de datos debería estar también representado en el de entrenamiento. De no ser así, el algoritmo daría una respuesta para estas personas que estaría fomentada en un sector de la población diferente. Este es el caso de varios algoritmos de reconocimiento facial desarrollados por grandes empresas norteamericanas que fueron prácticamente solo entrenados sobre rostros de hombres de tez blanca. La evaluación del comportamiento de esta solución mostró que se alcanzaba un error menor al 1% al probarla sobre rostros de hombres de tez blanca, pero se obtenía un error entre el 20% y 35% en el reconocimiento de mujeres de tez oscura (Buolamwini & Gebru, 2019).

En atención a lo cual, es importante acotar correctamente la población objetivo de la solución basada en IA que se quiere desarrollar y ceñirse a ello. A continuación, se propone un ejemplo práctico que tiene como objetivo acotar la población que beneficiaría de una hipotética solución basada en IA. En particular, nos centraremos en la supuesta creación de un sistema que ayude a la identificación de nódulos de pulmón usando radiografías.

3.1. DETECCIÓN DE NÓDULOS DE PULMÓN MEDIANTE RADIOGRAFÍAS

En los últimos años, en el ámbito de la asistencia sanitaria ha habido un auge en el número de soluciones basadas en IA que dan soporte en la interpretación de contenidos multimedia. En particular, en el campo de la imagen médica. Estos sistemas se conocen como diagnósticos asistidos por ordenador[6] y buscan patrones en los datos que indiquen posibles anomalías.

[5] El principio de proporcionalidad se recoge en el RGPD en el artículo 5 como uno de los principios relativos al tratamiento de forma que los datos objeto de tratamiento sean adecuados, pertinentes y limitados en relación con los fines para los que son tratados.

[6] En inglés conocidos como *computer-aided diagnosis*, *CAD*.

Para crear nuestro hipotético sistema que asista en la detección de nódulos de pulmón mediante radiografías, la primera pregunta que nos deberíamos hacer es: ¿el sistema incorpora alguna restricción? Lo primero que debemos saber es que en enero de 2014 la Unión Europea aprobó una nueva legislación por la que se establecen normas de seguridad básicas para la protección frente a los peligros derivados de la exposición a radiaciones ionizantes[7]. Esta ley considera la exposición de infantes a este tipo de radiación como una práctica especial. Es decir, que siempre que se pueda, se deben buscar otras alternativas más seguras que no radien, como la ecografía o la resonancia magnética. En consecuencia, nos encontraremos con escasas radiografías de personas de corta edad.

Si analizamos qué población se puede ver afectada por esta enfermedad vemos que el cáncer de pulmón principalmente afecta a personas de edad avanzada. De hecho, menos del 15% de los casos acontecen en menores de 30 años, y la edad promedio en la que se detecta son 60 años (Rubin, 2003). El conocimiento de la enfermedad pone sobre la mesa que la población menor no se vería beneficiada por esta solución en demasía.

Una vez hemos analizado sobre qué sectores de la población podríamos conseguir este tipo de datos y cuál es el perfil de las personas que se podrían ver beneficiadas por esta solución, debemos analizar qué factores debemos tener en cuenta para conseguir una muestra representativa. En este ejemplo particular, debemos tener en cuenta qué propiedades pueden tener los nódulos que analicemos. Una de sus características es el tamaño. Para asegurarnos de que nuestra solución dará respuesta ante todo tipo de nódulo, es relevante asegurarnos de que en los datos de entrenamiento haya una representación significativa de los diferentes tamaños de nódulos.

Este es un mero ejemplo práctico que pretende alimentar la reflexión sobre las preguntas que nos debemos hacer para asegurarnos de que la muestra con la que trabajaremos describe de forma significativa la realidad que queremos modelar.

3.2. SITUACIÓN ACTUAL

Actualmente, nos encontramos con que los algoritmos se entrenan en un entorno y luego se despliegan en un entorno significativamente diferente. Este es el caso de algunas herramientas de reconocimiento facial -como la citada anteriormente-, o la herramienta usada durante unos meses por Amazon para filtrar currículos que discriminaba a las mujeres que solicitaban trabajo ya que no encajaban con la plantilla de Amazon, compuesta básicamente por hombres caucásicos[8]. En este segundo

[7] Véase directiva 2013/59/EURATOM para más información.

[8] Véase https://www.theguardian.com/technology/2018/oct/10/amazon-hiring-ai-gender-bias-recruiting-engine.

ejemplo, se asumió que los datos de entrenamiento representaban de forma adecuada la realidad que se quería modelar. Y aquí está el error: la desigualdad estructural.

La sociedad en la que vivimos está plagada de sesgos, como son: los sesgos cognitivos, los sesgos estadísticos, y los sesgos contextuales. Eliminar todos los sesgos de un sistema de generalización puede parecer un objetivo deseable, pero el resultado no lo es. La capacidad de un sistema de aprendizaje se basa en la identificación de diferencias para clasificar nuevas instancias. Así pues, los sesgos no son el problema, los prejuicios y la discriminación sí. El conflicto es que los sistemas basados en IA pueden perpetuar las formas existentes de desigualdad estructural incluso cuando funcionan según lo previsto.

Una muestra representativa es aquella que tiene una fuerte validez externa en relación con la población objetivo que la muestra debe representar. Siempre será relevante conocer con qué limitaciones nos encontramos en la recolección de datos, cuál es la población afectada, así como definir con anterioridad la pluralidad a la que queremos que dé respuesta nuestra solución. Por esta razón, deberíamos adoptar como buenas praxis las siguientes indicaciones:

a) Si la solución basada en IA no fue probada, no se debería confiar en ella.
b) Si la solución basada en IA no fue probada en [entorno], no se debería confiar en ella en [entorno].
c) Si la solución basada en IA no fue probada con [población de personas], no se debería confiar en ella con [población de personas].
d) Si la solución basada en IA no fue probada con [población de datos], no se debería confiar en ella con [población de datos].

4. RESPETO POR LA PRIVACIDAD

Los beneficios que podemos obtener gracias a la explotación de datos son numerosos. Pero es importante que esto se consiga preservando los derechos humanos, como es el derecho a la privacidad y, en consecuencia, la protección de los datos personales. En este contexto, la disociación de los datos identificativos de las personas adquiere valor como el vehículo que puede asegurar el avance de las soluciones basadas en la explotación de datos sin menoscabar el respeto por la privacidad de las personas.

La anonimización de los datos implica la disociación de los datos identificativos de los datos personales con el objetivo de impedir su asociación a la persona titular de los mismos. Dicho con palabras de la AEPD, la anonimización debe producir “la ruptura de la cadena de identificación de las personas”[9].

Aunque la anonimización siempre sea preferible, en el ámbito de la salud no siempre es posible. Por ejemplo, los datos genéticos son por naturaleza identificables

[9] Véase https://www.aepd.es/sites/default/files/2019-09/guia-orientaciones-procedimientos-anonimizacion.pdf.

si estos caracterizan de forma única a una persona. En esta situación, hablar de anonimización carece de sentido. Existen otros escenarios, como el de subcontratar a entidades externas para realizar un análisis específico, que nos empujan a explorar otras opciones. En este ejemplo, querríamos enviar datos anonimizados sin perder los datos identificatorios de las personas titulares. En consecuencia, deberíamos de seudonimizar nuestro conjunto de datos.

La seudonimización implica el tratamiento de datos personales de manera que estos no puedan atribuirse a la persona titular sin usar información adicional, siempre que esta información adicional resida en otro espacio de alojamiento. En otras palabras, la seudonimización consiste en separar los datos personales de los identificativos y almacenar los datos identificativos en otra base de datos diferente a la original. La conexión entre ambas bases de datos se podría realizar mediante un código compartido, aunque esta conexión no se debería de realizar para así mantener la privacidad de las personas. Así, podríamos facilitar el conjunto de datos anonimizado a la entidad subcontratada sin que esta pueda identificar a las personas de la muestra.

Existen varias metodologías que permiten preservar la privacidad de las personas alterando los datos originales. Sea cual sea la metodología que convenga aplicar en nuestro contexto, es importante adoptar la buena praxis de calcular el riesgo de identificación de las personas. La eficacia del resultado obtenido se puede conocer analizando algunas de las propiedades del conjunto de datos final, como son: la k-anonimidad (Sweeney, 2002), la L-diversidad (Aggarwal & Philip, 2008), o la T-proximidad (Aggarwal & Philip, 2008).

5. PARA UNA EVALUACIÓN HOLÍSTICA

¿Estamos evaluando de forma correcta las soluciones basadas en IA? Aunque en general se prefiera escuchar un solo número para resumir una evaluación, en la mayoría de las situaciones es mejor informar de múltiples medidas. Centrémonos en el caso de una clasificación binaria, como sería el diagnóstico de una enfermedad (o la tienes o no). La precisión estadística analiza si las respuestas de la solución coinciden con la realidad. Por ejemplo, si se utiliza una solución basada en IA como soporte en el diagnóstico de X, una medida simple de precisión sería la cantidad de casos que se clasificaron correctamente como X como una proporción de todos los casos que se analizaron. En un supuesto en el que el 95% de los casos fueran categorizados de forma correcta como X, podríamos crear una solución que categorizara todos los casos como X y obtendríamos la misma precisión. ¿Sería esta una buena métrica? Claramente no. La evaluación de este ejemplo mejoraría con la incorporación de dos conceptos importantes que cobran especial relevancia en las pruebas médicas: la sensibilidad y la especificidad.

La sensibilidad mide la frecuencia con la que una prueba da un resultado correcto positivo para las personas que tienen la enfermedad de interés. También se

conoce como la tasa de "verdaderos positivos". La especificidad, en cambio, mide la capacidad de una prueba para dar un resultado negativo de forma correcta para las personas que no tienen la enfermedad del estudio. También se conoce como la tasa de "verdaderos negativos".

Estas tres métricas quizás sean las más conocidas, pero hay muchas más que pueden ser de interés en función del contexto. Sea como fuere, debemos ganar consciencia sobre que difícilmente un único valor describirá el comportamiento de nuestra solución. Evaluar susodicha solución mediante diferentes métricas nos permite comprender su comportamiento de una forma más holística. También es relevante conocer el grado de incertidumbre que la acompaña para evaluar su nivel de confianza, en otras palabras, es preferible acompañar las métricas reportadas de su intervalo de confianza.

Por otra parte, es importante destacar que las métricas no son importantes por sí mismas y es necesario tratarlas como aproximaciones a aquello que queremos estimar. Como dijo Goodhart "cuando una medida se convierte en el objetivo, deja de ser una buena medida"[10]. De lo contrario, el desarrollo de la solución basada en IA se centrará en intentar manipularla.

6. ¿HACIA DÓNDE VAMOS?

La IA avanza a un ritmo vertiginoso y, en consecuencia, las soluciones basadas en IA cobran un papel más importante en nuestras vidas día tras día. Estas soluciones tienen el potencial de ayudarnos a maximizar nuestra calidad de vida. Pero, al mismo tiempo, pueden conducirnos hacia una sociedad distópica. Encontrar el equilibrio adecuado entre el desarrollo tecnológico y la protección de los derechos humanos es una cuestión urgente.

Para asegurarnos de que el futuro de las soluciones basadas en IA no solo preserva, sino que fortalece los derechos humanos, se debe promover el conocimiento y la comprensión de la IA en las instituciones gubernamentales, las estructuras judiciales, y entre el público en general. Una alfabetización en IA nos permitiría como sociedad comprender los riesgos y el alcance de las soluciones basadas en IA, información ahora enmascarada. Así, la sociedad tendría el conocimiento suficiente para someter a escrutinio los avances y las aplicaciones de la IA.

Para que la alfabetización en IA cobre el máximo sentido, es necesario tener acceso a la información de interés. En otras palabras, se debe seguir fomentando la ciencia abierta y la transparencia. Por ejemplo, todo proyecto o producto subvencionado con financiamiento público debería de estar en abierto, facilitando, así, el acceso a quien tenga interés en escrutar la solución en cuestión.

[10] En inglés conocida como "When a measure becomes a target, it ceases to be a good measure".

El escrutinio algorítmico incrementaría su impacto si se realizara desde una conjunción de perfiles diversos que permitan analizar la solución de interés desde diferentes perspectivas para cubrir el análisis de todas sus posibles implicaciones éticas, sociales, legales, etc. La transdisciplinariedad connota una estrategia de investigación que atraviesa límites disciplinarios para crear un enfoque holístico. Tal como en el ámbito de la investigación en salud existen los comités de ética que se encargan de evaluar la protección de los derechos, seguridad y bienestar de las personas que participan en proyectos de investigación, recientemente han comenzado a ver la luz comités de ética formados por perfiles también tecnológicos para evaluar la investigación en otros sectores[11]. Deberíamos de seguir avanzando en esta línea para definir como sociedad qué tipo de soluciones queremos y cómo queremos que estas soluciones se integren en nuestras vidas.

La IA supone un futuro de posibilidades excitante, pero o lo acompañamos de más información, más transparencia, lo hacemos más comprensible, y facilitamos su escrutinio, o empezamos a olvidarnos de un beneficio para el público global en pro de un beneficio privado y sectorizado.

7. BIBLIOGRAFÍA

AGGARWAL, C. C., & PHILIP, S. Y. (2008). A general survey of privacy-preserving data mining models and algorithms. En *Privacy-preserving data mining* (págs. 11-52). Springer.

BAIG, M. M., GHOLAMHOSSEINI, H., CONNOLLY, M. J., & LINDÉN, M. (2015). Advanced decision support system for older adults. *pHealth*, (págs. 235-240).

BUOLAMWINI, J., & GEBRU, T. (2019). Gender shades: Intersectional accuracy disparities in commercial gender classification. *Conference on fairness, accountability and transparency* (págs. 77-91). PMLR.

DAY, G. S., SCHOEMAKER, P. J., & GUNTHER, R. E. (2000). *Wharton on Managing Emerging Technologies.* JohnWiley &Sons.

KHATTAK, A. M., PERVEZ, Z., HAN, M., NUGENT, C., & LEE, S. (2012). DDSS: Dynamic decision support system for elderly. En *2012 25th IEEE International Symposium on Computer-Based Medical Systems (CBMS)* (págs. 1-6). IEEE.

[11] Véase como ejemplo la creación del comité de ética de la UPC, formado el 01/04/2020 https://cutt.ly/xv3TLgf.

RUBIN, P. (2003). *Oncología clínica: enfoque multidisciplinar para médicos y estudiantes.* Elsevier España.

SUBÍAS-BELTRÁN, P., ORTE, S., VARGIU, E., PALUMBO, F., ANGELINI, L., ABOU KHALED, O., . . . CAON, M. (2019). A decision support system to propose coaching plans for seniors. *2019 IEEE 32nd International Symposium on Computer-Based Medical Systems (CBMS)* (págs. 592-595). IEEE.

SWEENEY, L. (2002). k-anonymity: A model for protecting privacy. *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems, 10*(5), 557-570.