Guía de autoevaluación ética de sistemas de Inteligencia Artificial para aplicaciones de la salud

Grupo de investigación del proyecto
“Detección y eliminación de sesgos
en algoritmos de triaje y localización
para la COVID-19”

Las siguientes personas son autoras (por orden alfabético) del presente documento, todas ellas pertenecientes al Grupo de investigación del proyecto “Detección y eliminación de sesgos en algoritmos de triaje y localización para la COVID-19”, financiado por la Fundación BBVA, Convocatoria ayudas a equipos de investigación científica SARS-COV-2 y COVID-19, Humanidades, 2020.

Txetxu Ausín
David Casacuberta
Maite Cruz
Janet Delgado
Ariel Guersenzvaig
Joaquín Hortal
Alicia de Manuel
Cristian Moyano
Iris Parra Jounou
Angel Puyol
David Rodríguez-Arias
Jon Rueda

Octubre del 2022.

Índice

4 **Rationale**

4 ¿Cuál es el objetivo de este documento?

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

9 ¿Es posible eliminar estos problemas éticos siguiendo las recomendaciones de este documento u otros similares?

11 ¿Qué entendemos por sesgo algorítmico?

12 ¿Qué mecanismos pueden producir sesgos algorítmicos?

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

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

17 ¿Cómo debería presentar sus diagnósticos, predicciones y recomendaciones un algoritmo?

18 Sobre el uso de este documento de autoevaluación ética de sistemas de IA para aplicaciones de salud

20 **Documento de autoevaluación ética de sistemas de Inteligencia Artificial para aplicaciones de la salud**

20 Justicia, equidad y dinámicas participativas

24 Errores de juicio (fiabilidad, falsos positivos)

26 Adecuación con la práctica clínica

28 Privacidad

Rationale

¿Cuál es el objetivo de este documento?

El objetivo del rationale es exponer y justificar los argumentos y supuestos que hay tras las recomendaciones del *Documento de autoevaluación ética de sistemas de Inteligencia Artificial para aplicaciones de la salud* (sección 2). Estas recomendaciones sobre el diseño y uso de sistemas de Inteligencia Artificial (en adelante, IA) están planteadas en forma de preguntas y ayudarán al lector o lectora a hacer una autoevaluación ética. Además, en el rationale se explican cuestiones generales como qué son los sesgos e injusticias algorítmicas, cómo se producen, por qué afectan a la equidad en salud y cómo se pueden evitar o minimizar.

El valor añadido del presente escrito es doble. En primer lugar, su misión principal es que el lector o lectora tome conciencia de los problemas éticos del uso de la IA en salud, que es una tarea previa e imprescindible para ir más allá de la visión centrada en lo que se exige por ley. Por ello, no pretende ser una verificación del cumplimiento de la legalidad vigente. En segundo lugar, el texto pone el énfasis en las injusticias algorítmicas y la equidad en salud.

Las guías al uso sobre ética de la IA en salud se centran en problemas de privacidad y confidencialidad de los datos personales. Sin embargo, este documento destaca la precaución de que los sistemas de IA aplicados a la salud no generen ni contribuyan a acentuar la discriminación a las personas ni a acrecentar desigualdades en salud. Se trata de un aspecto menos abordado, pero igual o más importante desde un punto de vista ético. Esta precaución es necesaria porque la equidad en salud se puede ver seriamente comprometida con la injusticia algorítmica.

Queremos contribuir a evitar que un diseño y/o un uso inadecuado de la IA aplicada a la salud refuercen una mayor inequidad en salud. Por esa razón, hemos tenido en cuenta no solo las posibles discriminaciones (por razón de género, etnia, discapacidad, edad, etc.) de los sistemas de IA, sino también la presencia de los determinantes sociales de la salud (DSS), que son los responsables de buena parte de las desigualdades sociales en salud y que la IA podría exacerbar si no se toman las medidas oportunas tanto en su diseño como en su uso.

La guía fue pensada, inicialmente, para el uso de herramientas informáticas basadas en IA y aplicadas al contexto de la pandemia COVID-19 (gestión de recursos escasos en la Unidad de Cuidados Intensivos [UCI] y *apps* de trazabilidad). Sin embargo, muchas de estas reflexiones van más allá y permiten evaluaciones éticas de carácter general que pueden aplicarse a cualquier herramienta basada en IA y aplicada en salud (por ejemplo, un sistema de diagnóstico de cáncer).

Las preguntas de autoevaluación incluidas en el documento no pretenden ser un cuestionario cerrado sino abierto, capaz de integrar inquietudes éticas futuras o que el texto no ha captado adecuadamente en su formulación actual.

Aclaración preliminar: para facilitar la comprensión del presente escrito, cuando hablamos de "algoritmos" nos referimos exclusivamente a algoritmos de IA que utilizan técnicas de aprendizaje automático y no a los algoritmos tradicionalmente utilizados en medicina, como, por ejemplo, el árbol de decisión para determinar la etiología del dolor abdominal.

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

Buena parte de los algoritmos de IA usados en medicina tienen como función principal bien asignar la probabilidad de que una persona padezca una enfermedad o cualquier otro estado psicofísico que se pretenda identificar, bien la probabilidad de un mejor o peor pronóstico en pacientes ya diagnosticados.

En otras palabras, se trata de determinar, a partir de información clínica y de otro tipo (por ej., variables sociodemográficas), si esa persona sufre o no una determinada enfermedad o cuál es el riesgo de que su salud se agrave. Si bien en la mayoría de los casos analizados esta asignación probabilística 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 (por ejemplo, el "sobreajuste" u *overfitting*, relativo a los datos utilizados para entrenar el modelo). Podemos así tener un falso positivo (el o la paciente no tiene la supuesta enfermedad, pero el algoritmo diagnostica que sí) o bien 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 intervenciones fútiles que además conllevan riesgos asociados (primer caso, donde lo indicado sería no realizar una intervención, según el principio de no maleficencia) o la omisión de tratamientos necesarios (segundo caso). Por otra parte, estos errores de juicio automatizado podrían afectar a la autonomía decisional de los pacientes al no permitirles participar en la toma de decisiones sobre el tratamiento. Además, podrían generarse "perfiles" de pacientes a partir de los datos que utiliza el algoritmo, de manera que se automatizaran las decisiones en función de estos grupos y se perdiera el componente humano en la individualización de decisiones en cada caso.

2. **Errores de diagnóstico:** este tipo de falsos positivos o falsos negativos, a diferencia de los que pueden ocurrir 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 pacientes como al propio personal

sanitario, porque enturbian 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 haya un sistema de revisión o de cuestionamiento del resultado que arroja el algoritmo. Debe poder mantenerse el principio de confianza. Dado que el algoritmo automatizado de IA puede condicionar parte del procedimiento asistencial, este debe complementarse con la participación del profesional.

3. **Privacidad:** independientemente de si el resultado algorítmico es acertado o no, la información queda registrada en una base de datos, accesible a terceros si no se adoptan medidas adecuadas para la seguridad de los datos incluidos. Tenemos así un riesgo para la privacidad que obliga a asegurar que la información proporcionada por la IA solo pueda ser consultada por personas autorizadas.
Imaginemos el caso de una *app* de bienestar mental y emocional ofrecida a empleados y empleadas por parte de la mutua de salud contratada por la empresa u organismo público donde trabajan. Esos datos, que aparentemente solo benefician a los empleados, podrían ser usados por la empresa para vigilar sin consentimiento explícito la salud mental de sus trabajadores y/o a tomar decisiones operativas (un despido o una promoción) en base a ellos.

4. **Fragmentación y parcelación 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; es decir, situados en contextos sociales, culturales y con una visión tan holística como sea posible de la salud.

5. **Relación médico-paciente:** 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 sobresaturación por alertas del punto 2). Ello puede desembocar en una transformación 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 IA puede suponer la intervención de nuevas entidades (por ejemplo, empresas) en procesos tradicionalmente definidos, coordinados y ejecutados por profesionales sanitarios y gestores públicos. ¿Cómo garantizar que los criterios que guían la práctica médica sean también los que guíen la gestión de una entrada masiva de proveedores de servicios de IA? ¿Cómo modifica, de manera indirecta, la atención equitativa y justa de los pacientes?

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 (equilibrio entre predicción y prevención cuaternaria). Esto contribuye a un mayor riesgo de iatrogenia, uno de los grandes males de los sistemas sanitarios más avanzados.

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 tanto individual como colegiado. 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 IA. La sobrevaloración y el exceso de confianza en las potencialidades de la IA pueden agravar los errores y problemas arriba mencionados, tanto en su dimensión técnica como ética, cuando el equipo médico confía ciegamente en la propuesta generada por el algoritmo y prescinde de su propio criterio.

9. **Justicia y equidad:** problemas relacionados con los datos de entrenamiento del algoritmo que tienen poca o inexistente representación de ciertos grupos sociales o determinadas características de la condición bajo estudio exacerban las disparidades ya existentes en salud y llevan a juicios sesgados. Por ejemplo, la baja cantidad de imágenes de pieles oscuras para entrenar un *software* de detección de melanomas provoca peores resultados en el análisis de personas afrodescendientes. Otro ejemplo es la presencia de "correctores raciales" en la priorización de acceso al trasplante 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 (ver *How algorithm blocked kidney transplants black patients*). Si el algoritmo se emplea para gestionar el acceso a recursos sanitarios, tendremos otro problema de sesgo y discriminación desigual. También podemos tener sesgos si, por ejemplo, usamos en un país de forma acrítica un algoritmo que ha sido entrenado en el contexto específico de otro país con una composición poblacional diferente y diferentes hábitos de salud. Como pone de manifiesto el proyecto *AI-Fora*, los algoritmos de 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 equidad han sido tradicionalmente infravalorados en un contexto de salud. Nuestra intención es darles el protagonismo que se merecen a tenor de las enormes consecuencias que pueden tener en la inequidad y en las desigualdades de salud de la población.

   Ante el riesgo de vulneración de derechos en un ámbito tan sensible como la salud, que un algoritmo sea técnicamente fiable y genere resultados para lo que ha sido entrenado no es suficiente. Tenemos que

establecer también 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. Es especialmente importante la necesidad de que haya transparencia algorítmica respecto al código fuente del programa con el fin de que cualquier persona interesada pueda verificar su diseño y funcionamiento.

¿Es posible eliminar estos problemas éticos siguiendo las recomendaciones de este documento u otros similares?

A la hora de la verdad, cualquier sistema desarrollado por el ser humano es falible y abierto a sesgos. Pensar que podemos crear un sistema completamente libre de errores y justo es una quimera.

Siempre habrá alguna brecha de seguridad en la base de datos que una persona maliciosa pueda aprovechar para apoderarse de información sensible de terceros, por ejemplo. Tampoco podemos esperar que una sociedad con injusticias de base, en la que todas las personas con capacidad para intervenir en el proceso de creación de IA son susceptibles de tener prejuicios o de actuar sesgadamente, sea capaz de generar un algoritmo absolutamente justo. En línea con esto, y al igual que ocurre en el campo de la filosofía, las ciencias de la computación tampoco disponen de métricas o nociones unánimemente aceptadas para la justicia o la equidad; algunas de ellas, incluso, no acaban de ser compatibles entre sí.

Sin embargo, el riesgo añadido del empleo de algoritmos es que se puede generar la impresión errónea de que son infalibles o moralmente neutros; de que no se basan en decisiones que implican priorizar ciertos valores sobre otros. Por otro lado, desde una perspectiva de equidad, el trato estrictamente igual a las personas en un contexto social de desigualdades también podría considerarse injusto.

Si los sistemas algorítmicos a los que nos referimos demuestran su utilidad clínica, superan estudios de validación, y satisfacen a la vez los requerimientos legales y morales, resulta inadmisible desde el punto de vista ético no utilizarlos en situaciones en las que su uso sería

provechoso. Por eso es muy importante ser conscientes de sus problemas; saber detectarlos, reconocer sus causas y ponerles remedio —o, cuando esto no es posible, al menos minimizarlos— siguiendo las indicaciones de este documento u otros similares.

En un futuro cercano, el personal sanitario que trabaje con la IA para diagnósticos y decisiones clínicas debería, como mínimo, poder preguntarse con normalidad si en relación con una paciente concreta (pongamos una mujer de sesenta y cinco años, obesa y de etnia gitana) la IA utilizada para realizar un diagnóstico y un pronóstico de su enfermedad ha sido adecuadamente entrenada con datos que no infrarrepresenten a su grupo social de referencia o a las personas con sus mismas características sociales y clínicas.

Hay razones técnicas y metodológicas que imposibilitan adoptar un criterio que elimine totalmente los problemas éticos. En el proceso de creación de un algoritmo de IA, la fiabilidad del algoritmo y la justicia en el proceso de decisiones basadas en él son incompatibles muchas veces. Cuanto más nos esforcemos en que el algoritmo capture la situación real que queremos modelar, con una representación proporcional de los colectivos afectados, más ganará en fiabilidad y validez en la identificación y predicción. Sin embargo, también entrarán con más facilidad los sesgos ya presentes en la sociedad. Por esta razón, la cuestión ética no puede resolverse de manera permanente y genérica. Cada aplicación requiere de consideraciones, valoraciones y decisiones que son histórica y socialmente contextuales y, por tanto, dinámicas.

En general, los algoritmos deben tratarse como herramientas de ayuda a la asistencia sanitaria, es decir, como sistemas de apoyo a las decisiones (*DSS decision support systems*). Se debe alentar un enfoque dinámico de revisión de los algoritmos en el que la intervención del profesional de la salud siga siendo imprescindible. Los profesionales sanitarios deben cultivar una visión crítica que contemple estos sesgos y errores y procure minimizarlos. Esto no significa que debamos prohibir el uso de algoritmos de IA o que debamos aceptarlos tal y como son ahora. La realidad es que, aunque los riesgos no se puedan eliminar por completo, sí se pueden minimizar sustancialmente. La responsabilidad consiste en generar y usar algoritmos más sensibles a las desigualdades sociales, que eviten reproducirlas en sus análisis, que cometan menos errores derivados de sesgos éticos, y que se apliquen de forma suficientemente segura para garantizar la privacidad.

¿Qué entendemos por sesgo algorítmico?

La palabra "sesgo" tiene múltiples significados, y ello produce dificultades a la hora de decidir cuándo un algoritmo es injusto y qué podemos hacer para remediarlo. Una lectura demasiado matemática del término "sesgo" lleva a considerar solamente problemas asociados a la exactitud del algoritmo y a proponer soluciones puramente matemáticas que no consideran las injusticias estructurales que el algoritmo puede potenciar. Hay sesgos estadísticos, humanos y sistémicos. En este documento nos centraremos en sus resultados, que llevan a la discriminación de algunos grupos minoritarios o vulnerables mientras que otros se benefician de ello.

Encontraréis una definición de cada uno de los términos y conceptos técnicos usados en el documento en el glosario de la página web del grupo de investigación.

En este documento, entendemos *sesgo* como la tendencia a beneficiar a un grupo y/o perjudicar a otro de manera injusta o éticamente inadmisible. En otras palabras, decimos que un sistema *contiene sesgos* 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, donde la mayoría son todavía hombres. También 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 de pecho y brazo izquierdo frente al dolor en epigastrio en mujeres) debido a una subrepresentación de las mujeres en el modelo algorítmico. Los sesgos a los que nos referimos son primordialmente sesgos *no intencionados*, es decir, son sesgos 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. Las condiciones sociales, ambientales y económicas de las personas determinan su salud a través de los

determinantes sociales de la salud (DSS). Cuando estos DSS afectan de manera desigual a las personas y estas no los pueden controlar, decimos que su influencia es injusta. 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 del 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.

¿Qué mecanismos pueden producir sesgos algorítmicos?

Resumiendo, podemos establecer cuatro tipos de mecanismos principales:

1. **El sesgo codificado en los datos.** Sucede cuando las personas responsables de generar, o al menos de etiquetar el conjunto de datos, lo hacen de forma sesgada. Debido a que los algoritmos de aprendizaje automático están diseñados con precisión para ajustarse a los datos que reciben, inevitablemente reproducirán el sesgo ya presente en los datos y sería ilusorio creer que el algoritmo por sí solo puede deshacerse de tales sesgos.
   Así, en el desarrollo de un algoritmo para decidir si un paciente ha de ser ingresado en la UCI o no, si los datos se toman de decisiones anteriores en las que se discriminaba a las personas de edad avanzada, esto se traducirá probablemente en un sesgo. El algoritmo tendrá una precisión perfecta, pero a la vez injusta. No obstante, es necesario matizar que no siempre debe haber una representación homogénea de los distintos estratos para cerciorarse de que el sistema es justo, puesto que dependerá del problema que el algoritmo aborda. No tendría sentido, por ejemplo, que se desarrollara un algoritmo de detección prematura de osteoporosis con una base de datos que representara por igual a personas mayores y jóvenes, ya que la osteoporosis es prevalente en personas mayores.

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. 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 acarrear que colectivos que ya son discriminados al recibir menos atención sean doblemente discriminados al subestimarse su riesgo y se priorice a grupos que ya se benefician de un mayor acceso a prestaciones sanitarias (ley de cuidados inversos).

3. **Crear un algoritmo que se ajusta excesivamente 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 anteriormente mencionado; el algoritmo lo hace relativamente bien con personas de origen caucásico, pero comete errores importantes con personas de otras etnias.

4. **Las poblaciones son sistemas dinámicos.** Cada decisión tomada por una persona o un algoritmo producirá cambios en el conjunto que tendrán que procesarse. Para confirmar que un algoritmo ha generado una predicción o recomendación correcta, se crea una tensión entre lo que es más útil para la persona a nivel individual y lo mejor para establecer la fiabilidad y la exactitud del algoritmo. Si algunas minorías están subrepresentadas en el conjunto de datos o el sistema no es lo suficientemente preciso con quienes están dentro de un contexto cultural diferente, la distancia entre lo que el sistema necesita y lo que es mejor para los pacientes aumentará y, por lo tanto, generará una situación injusta para pacientes de contextos minoritarios.

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

En primer lugar, debemos asegurar que el algoritmo representa a todos los grupos sociales relevantes en nuestra base de datos en un número suficiente para evitar el etiquetado incorrecto de miembros de minorías por la falta de casos suficientes.

En la práctica, esto implica, en muchas ocasiones, la necesidad de sobrerrepresentar a las minorías en las muestras utilizadas para diseñar el algoritmo. Siguiendo esta regla, los algoritmos de aprendizaje automático que diagnostican enfermedades de la piel mejorarían si en lugar de tener una sobrerrepresentación de hombres blancos aumentaran la diversidad de tonos de piel en la muestra hasta alcanzar un número suficiente para que todos los tipos de piel estuvieran bien representados.

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. Obtener y clasificar los datos cuando se quiere modelizar comportamientos sociales elaborados —como, por ejemplo, "entrar de forma ilegal en un edificio"— es muy complejo. En el ámbito social no contamos con un punto de partida unívoco (*ground truth*) que permita determinar qué datos empíricos harán falta para entrenar el sistema de manera idónea. Así que es fácil que el algoritmo no procese datos menos frecuentes, pero relevantes, de factores socialmente significativos y que al procesar otros que son frecuentes, aunque no relevantes, acabe fijándose exclusivamente en los datos más estables e intente generalizar a partir de ellos.

Para prevenir 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 de 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 utilizar hasta 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, tenemos que saber 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. Para esto, 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.

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

Tradicionalmente, un algoritmo de IA se diseña con una única métrica en mente: la fiabilidad y exactitud de la predicción. Dicho de otro modo, se busca que el algoritmo se acerque de la manera más precisa al resultado correcto (positivos y negativos verdaderos) y se minimice la diferencia entre el resultado real y el predicho por la máquina. Nada impide, sin embargo, incluir otras métricas que permitan juzgar la bondad del algoritmo desde otras perspectivas.

De la misma forma, podemos incluir una métrica que asegure una mayor equidad a la hora de representar los resultados. Se le puede dar un peso específico a los DSS, aunque en el proceso perdamos algo de exactitud en

el cómputo final. Por ejemplo, usar un algoritmo que tenga en cuenta que determinados colectivos tienen un mejor estado de salud debido a DSS desiguales e injustos podría resultar útil para tomar decisiones más justas a la hora de priorizar a personas en una UCI con recursos limitados. Eso implica que la decisión de priorización estará informada por criterios éticos y no sólo clínicos, con el fin de evitar decisiones que perpetúan ciertas injusticias.

También podemos sacrificar cierta exactitud para proteger la privacidad. La *privacidad diferencial* es un modelo en el que pueden incluirse datos falsos de personas ficticias, o modificar de manera aleatoria algunos datos que forman la base de datos ("ruido", en términos técnicos). Se hace de forma aleatoria para que, en el caso de que la base de datos acabe en manos de terceros no deseados, la presencia de ese ruido imposibilite la identificación de ninguna persona de la base de datos. Asimismo, existe la posibilidad de combinar datos "sintéticos" (es decir, datos artificiales generados por el propio algoritmo) con datos reales. Así, el proceso de reidentificación se hace prácticamente imposible. También se puede optar por usar solo datos sintéticos, con lo que no se implica la privacidad de nadie.

El precio de incluir ruido a través de datos ficticios son 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 una UCI 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, nuestro contraargumento es que también en una UCI 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.

¿Cómo debería presentar sus diagnósticos, predicciones y recomendaciones un algoritmo?

Normalmente, los algoritmos de IA de aprendizaje automático muestran sus resultados con valores numéricos que han de interpretarse como probabilidades. Lo más común es que la predicción aparezca como un valor puntual con un intervalo de confianza predeterminado.

Un algoritmo que calcule la probabilidad de supervivencia de una persona enferma de COVID-19 si entra en la UCI devolverá un valor entre 0 y 1, expresado habitualmente en tanto por ciento. Un supuesto: se nos diría, por ejemplo, que la probabilidad de supervivencia es del 78%, con un intervalo de confianza (habitualmente del 95%) entre 76% y 80%. Ese número puede ir acompañado de una barra de estado que representa gráficamente ese intervalo.

Este tipo de representación de resultados basada en la probabilidad numérica es problemático. El ordenador determina un número y el personal sanitario asume que debe aceptarlo y aplicarlo, sin necesidad de entender el razonamiento que se ha usado o los posibles sesgos que arrastra, y sin posibilidad de replicarlo.

En consecuencia, cada vez más instituciones, equipos investigadores y colectivos demandan otra forma de presentar las predicciones en algoritmos; una en la que la interpretación sea un valor central. Es decir, que el programa no se limite a determinar un número, sino que además indique las razones de las que se deduce el resultado. Esto contribuiría a mantener y aumentar las habilidades de los profesionales de la salud y fomentaría una actitud crítica hacia el juicio automatizado.

Inevitablemente, para crear un algoritmo que ofrezca explicaciones la búsqueda de relevancia estadística y de localización de patrones debe ser más estable. Esto dificulta hacer ajustes finos y afecta negativamente a la exactitud total del algoritmo. Se trata de una cuestión técnica. Para mejorar la exactitud de un algoritmo, tenemos que entrar en sus “tripas”, y esto disminuye su explicabilidad. Pero, insistimos, cuando hay varios valores en juego, es inevitable tener que negociar entre ellos para asegurarse de que el resultado final se adapte lo mejor posible a una pluralidad de intereses. La exactitud no es la única meta éticamente posible y deseable detrás de un algoritmo de IA. La explicabilidad y la transparencia son otros valores importantes en justicia algorítmica.

Sobre el uso de este documento de autoevaluación ética de sistemas de IA para aplicaciones de salud

Las siguientes preguntas se refieren al empleo de algoritmos y otras herramientas de IA para cualquiera de sus aplicaciones en el campo de la salud, incluyendo la gestión de crisis sanitarias, como la de COVID-19 y análogas. Dichas preguntas están elaboradas con la intención de que se responda con una afirmación o una negación.

La creación de esta guía de autoevaluación ética, así como de las preguntas que contiene, forman parte de una investigación realizada en varias etapas. La primera, un estado de la cuestión sintetizado en el artículo *"Bias in algorithms of AI systems developed for COVID-19: A scoping review"* (Delgado et al. 2022). Nuestro objetivo era conocer lo que la literatura científica decía sobre los sesgos que se producen en los algoritmos de IA utilizados tanto para triaje como en aplicaciones de rastreo de contactos durante el COVID-19. A partir del análisis de la bibliografía disponible, llegamos a una serie de conclusiones que incluía tanto problemas y errores técnicos como debates éticos. Gracias a los resultados obtenidos durante la investigación pudimos realizar un análisis más profundo sobre los problemas y debates éticos detectados e incluirlos en este *rationale* [consultables en el artículo "*Ethical assessments and mitigation strategies for non-racial biases in AI-systems used during the COVID-19 pandemic*" (de Manuel et al. 2022)].

Esta primera investigación nos mostró la necesidad de avanzar en torno a los modelos de autoevaluación. Realizamos, pues, una búsqueda sobre modelos de *checklists*, certificaciones éticas y sistemas de autoevaluación —como pueden ser *"Ethical by Design: Principles for Good Technology"* de *Ethics Center*, *"Technical methods for regulatory inspection of algorithmic systems in social media platform" de Ada Lovelace Institute*, *"Ethics Guidelines for Trustworthy AI"* de la Comisión Europea, *"Draft text of the Recommendation on the Ethics of Artificial Intelligence"* de UNESCO, *"AI systems classification framework"* de la OECD, o en un contexto más próximo, el sistema de autoevaluación "*modelo PIO*" desarrollado por el Observatori d'Ètica en Intel·ligència Artificial de Catalunya. Basándonos en estos modelos de *checklists* actualmente operativos pudimos desarrollar este modelo de autoevaluación que permite a las empresas, organizaciones y agentes conocer el estado en el que se encuentra un determinado sistema de IA desde el punto de vista de la equidad y la justicia.

Nuestro objetivo es que la persona que responda a las preguntas planteadas pueda autoevaluar su nivel de conocimiento y compromiso con la detección y eliminación de sesgos éticos en los algoritmos utilizados en el campo de la salud. Cuantas más respuestas afirmativas, mayor compromiso. El fin último de este documento es que sirva de herramienta para tomar conciencia de la importancia de prevenir, detectar y eliminar sesgos que vuelven injustos tanto el diseño como la aplicación de los algoritmos usados en el campo de la salud y, de este modo, contribuir al desarrollo de otros más justos.

Este documento ha sido diseñado especialmente para equipos desarrolladores y aplicadores de los algoritmos en salud, pero puede ser igualmente útil para cualquier persona interesada en la prevención, detección y eliminación de la injusticia algorítmica. Igualmente, partimos de la convicción de que cualquier profesional que diseñe o use dichos algoritmos tiene la responsabilidad de hacer un uso éticamente adecuado de ellos con independencia de su nivel de implicación profesional y de su categoría laboral.

Hay que tener en cuenta que el uso de la IA conlleva el riesgo de diluir las responsabilidades éticas de su utilización. Aparece la falsa idea de que la decisión es tomada por la IA, como si ella misma fuese un agente moral al que se le pudiese pedir responsabilidad. Muy al contrario, precisamente para contrarrestar esa inercia, hay que estar más precavidos que nunca y asumir la responsabilidad moral personal en el uso de ese tipo de tecnología. La IA nunca puede sustituir a la autoridad clínica del profesional sanitario que toma decisiones clínicas y, por tanto, este tampoco puede eludir su responsabilidad moral en ellas, incluso cuando haya utilizado la IA como apoyo.

Dada la extensión del documento, existen otras dimensiones con clara relevancia ética que no han podido ser incluidas o analizadas en detalle. Algunas de ellas son: las condiciones laborales de quienes participan en el procesamiento de los datos o en el entrenamiento del sistema, el impacto medioambiental del propio sistema, el sesgo colonialista, los sesgos culturales, o la ciberseguridad, entre otros. Se trata de cuestiones a escala macro que exceden el foco de estas recomendaciones y requieren un tratamiento más específico.

El documento de autoevaluación se puede contestar individualmente, pero recomendamos que sea todo el equipo que intervenga en el desarrollo o en la aplicación de la IA quien responda.

Documento de autoevaluación ética de sistemas de Inteligencia Artificial para aplicaciones de la salud

Usted, desarrollador o implementador de este algoritmo basado en Inteligencia Artificial, ha considerado:

Justicia, equidad y dinámicas participativas

1. **¿Si el público destinatario que será usuario del algoritmo ha participado en las distintas fases del ciclo de vida del proyecto (de diseño a implantación)?**

   Es importante contar con la participación en todas las fases del ciclo de vida del proyecto (desde el diseño a la implantación) de una representación de aquellos colectivos que vayan a ser usuarios del algoritmo y/o afectados por el mismo. De esta manera, además, promovemos la transparencia algorítmica dentro del sistema. Por ejemplo, se puede valorar la participación de estos colectivos mediante metodologías específicas: grupos de discusión, cocreación, metodología Delphi, grupos nominales, etc.

2. **¿La posibilidad de incorporar mecanismos externos e independientes de auditoría del código fuente del sistema de IA?**

   Se recomienda promover la transparencia algorítmica, sobre todo si existe un riesgo de que se puedan producir sesgos, todo ello sin perjuicio de otros intereses comerciales, industriales, etc.

3. **¿Si se han documentado todos los pasos de diseño y evaluación, así como las decisiones tomadas durante el ciclo de vida del proyecto?**

   Se recomienda documentar todos los pasos de diseño y evaluación para poder detectar con mayor facilidad los sesgos que se hayan podido

introducir y facilitar la rendición de cuentas. Para ello, recomendamos la utilización de técnicas como *Model Cards for Reporting* o el Modelo de Explicaciones Justificativas desarrollado en el marco de este proyecto.

4. **¿Si el sistema de IA contempla un mecanismo para recibir retroalimentación de cualquier parte interesada o afectada, así como para dar una respuesta a esos usuarios y/o personas interesadas?**

   Se recomienda que la organización genere un espacio accesible para recibir *feedback* de personas usuarias interesadas y/o afectadas, de esta manera, además, ayudamos a promover la transparencia en el sistema. Por ejemplo, a través de su página web, un apartado específico en la aplicación móvil o una cuenta de correo determinada.

5. **¿Que se hayan tomado precauciones para minimizar la posibilidad de que el sistema de IA pueda discriminar a las personas por alguno de estos motivos: sexo, género, nacionalidad, raza, origen étnico o social, características genéticas, idioma, religión o creencia, opinión política o de cualquier otro tipo, pertenencia a una minoría, lugar de residencia, discapacidad, edad y/u orientación sexual?**

   Bajo ningún concepto un sistema de IA debería tratar de manera injusta a las personas usuarias por alguno de los motivos anteriormente mencionados. Una buena forma de detectar estos problemas al inicio es realizar un análisis estadístico con los datos que alimentan al algoritmo para comprobar que la población está correctamente representada, y no hay sobrerrepresentación o subrepresentación de algunos colectivos. Valore cuidadosamente la posibilidad real que tiene su sistema de producir estos sesgos, particularmente teniendo en cuenta sus resultados. Por ejemplo, es posible que un sistema pueda ser utilizado en un país o área geográfica con poblaciones distintas a las que se usaron durante el entrenamiento y desarrollo del sistema. Por ello, y para prevenir el riesgo de malinterpretar el sistema como universal, recomendamos avisar sobre las condiciones bajo las que el sistema ha sido creado, por ejemplo: "el sistema contempla una población por etnicidad, o por país, entre 20 y 70 años... No utilizar bajo otras condiciones". Por otra parte, el hecho de que características como el sexo, género o etnia no estén incluidas dentro de los datos utilizados en el sistema no significa que no puedan aparecer como *proxies* generados por el algoritmo. Por ejemplo, una base de datos puede no incluir datos como la etnia, pero a través de datos como el código postal, el programa podría reconstruir esa información en el caso de que estos datos estén altamente correlacionados.

6. **¿Que, además de las características anteriores, se hayan tomado precauciones para minimizar la posibilidad de discriminar a las personas por su condición socioeconómica?**

   En ningún caso, un sistema de IA debería discriminar a las personas usuarias por su condición socioeconómica. Es importante recordar que

datos aparentemente neutros, como el código postal o la obesidad, son de hecho poderosos informadores de la condición socioeconómica, indicando que la persona vive en un barrio rico o uno pobre o que pertenece a una clase social determinada. Es fundamental tener en cuenta esta información con el fin de evitar o minimizar que el sistema reproduzca o acentúe las desigualdades en salud derivadas de los determinantes sociales de la salud injustos. Por ejemplo, un algoritmo que, ante recursos limitados, prioriza a un enfermo porque tiene un beneficio sanitario esperado mayor medido en AVACs (Años de Vida Ajustados por Calidad de vida) puede estar tratando injustamente a los enfermos que más sufren las desigualdades injustas en salud y que, por esa razón, no pueden tener una mayor puntuación en AVACs, por ejemplo, aquellos que sufren obesidad mórbida. Este tipo de injusticia se puede producir sin la intervención de la IA, pero esta tiene el triple efecto negativo de perpetuar la injusticia inicial, de aumentarla (al negar nuevas expectativas de salud que van a parar a enfermos que ya se beneficiaban de las desigualdades sociales en salud) y de ocultarla bajo la falsa percepción de que se trata de una decisión éticamente neutral al haber sido aconsejada por la impersonalidad de la IA. Es posible que el profesional sanitario que tiene que aplicar la IA no disponga de herramientas para evitar ese tipo de injusticia, pero eso no le exime de ser consciente de ella.

7. **¿Complementar el uso de IA con ayudas de otro tipo (por ej., guías o árboles de decisión tradicionales) para apoyar el proceso de toma de decisiones en la dictaminación de la distribución final de los recursos sanitarios limitados?**

   Es recomendable que la estimación concreta de la IA no sea el *único* criterio a tener en cuenta en la asignación de recursos escasos (como la ventilación mecánica, los órganos para trasplante, las camas de UCI, etc.). Esto es especialmente importante en los casos de racionamiento, esto es, cuando se le deniega el recurso escaso a alguien que lo necesita para asignarlo a otra persona.

8. **¿Si el algoritmo ofrece una explicación comprensible para las personas usuarias y/o para aquellas afectadas por los resultados que arroja el sistema?**

   Se recomienda que la explicación de los resultados que arroja el sistema sea accesible, comprensible y se encuentre disponible en varios idiomas para todas las personas usuarias o que puedan ser afectadas por el sistema de IA. Se ha de garantizar tanto el derecho a saber de la persona usuaria como el derecho a no saber, siguiendo lo dispuesto en la normativa vigente, especialmente en lo que se refiere a los resultados que arroje el sistema en términos de salud.

9. **¿Si, en caso de que el algoritmo se haya desarrollado en un contexto de crisis sanitaria que requiera una rápida respuesta, el equipo ha prestado la misma atención y no una atención menor a ciertos parámetros que en situación de normalidad?**

   Bajo ningún concepto, incluso en casos de crisis o emergencia sanitaria, o con tiempos muy cortos de respuesta, deben ser comprometidos los parámetros y estándares mínimos de calidad del sistema de IA con relación a aspectos técnicos o legales (estándares de privacidad, la gestión de datos, etc.), ni pueden admitirse resultados que sean prima facie injustos (i.e., discriminatorios) con ciertos colectivos.

10. **¿Si los resultados arrojados por el algoritmo del sistema son cuestionables por personas usuarias y/o por el personal sanitario?**

    Los resultados arrojados por el algoritmo deben poder ser siempre cuestionables y reclamables por personas usuarias y por el personal sanitario de manera rápida y fácilmente accesible. El sistema debe ofrecer una funcionalidad para lograr este propósito.

11. **¿Si los indicadores o métricas de desempeño se han definido con personas expertas del ámbito de la salud, aunque no sean usuarios?**

    Se recomienda que los indicadores o métricas del desempeño del sistema sean consensuadas y definidas juntamente con expertas y expertos en el ámbito de la salud.

12. **¿Tomar precauciones para evitar que, una vez implementado el sistema, diferencias en el grado de adopción (por ejemplo, en diferentes áreas geográficas o en colectivos sociales) puedan terminar generando recomendaciones y/o predicciones sesgadas?**

    Es importante considerar que un sistema que al inicio de su funcionamiento genera resultados justos a primera vista puede generar sesgos con el uso. Por esta razón es importante implementar mecanismos de control periódicos. Un sistema de IA, una vez entrenado e implementado, seguirá "aprendiendo" y puede ocurrir que, en el caso de que se produzcan diferentes grados de adopción por parte de diferentes colectivos o en zonas geográficas los nuevos datos que se generen pueden contribuir a sesgar el funcionamiento del sistema. Así, si un sistema es ampliamente utilizado en un barrio con ciertas características (por ejemplo, una baja prevalencia de una enfermedad determinada), los datos generados podrían sesgar su funcionamiento en detrimento de otras áreas geográficas donde la enfermedad sí sea prevalente, pero la adopción del sistema sea menor. El sesgo en los datos obtenidos durante el uso puede también provenir de otras fuentes, por ejemplo, un sistema pensado para ser utilizado directamente por los

y las pacientes puede generar datos sesgados si el sistema no funciona según el tipo de *smartphone* (por ej. si funciona bien en IPhone, pero "se cuelga" en Android, las personas usuarias no lo usarán) o si no se han desarrollado las funcionalidades de accesibilidad para personas con necesidades especiales.

13. **¿Que el desarrollo, implementación y uso del sistema de IA genera un daño medioambiental?**

    Se recomienda valorar si es posible realizar una evaluación del impacto medioambiental del sistema de IA y sus fases de desarrollo, implementación y uso para reducir su huella ecológica. Para almacenar y procesar todos los datos que puede manejar un sistema de IA se utiliza una gran cantidad de materias primas escasas y diversos recursos energéticos. Sería conveniente detectarlos, así como conocer las emisiones de contaminantes o de gases de efecto invernadero derivados.

Errores de juicio (fiabilidad, falsos positivos)

14. **¿Establecer una estrategia o un conjunto de procedimientos durante la fase de desarrollo para detectar si el sistema puede generar sesgos injustos?**

    Es importante generar diferentes procedimientos y estrategias para detectar los sesgos injustos. Por ejemplo, a través de simulaciones que permitan comprobar qué predicciones haría el algoritmo con datos de usuarios que no representen a la mayoría de la población. También es importante tener en cuenta que la población para la que ha sido diseñado el sistema de IA pueda cambiar, por ejemplo, debido a un cambio migratorio.

15. **¿Establecer, una vez implementado el sistema, un mecanismo o estrategia para relacionar los errores de predicción del algoritmo de IA con posibles sesgos injustos?**

    Se recomienda implementar un plan de seguimiento de los errores de predicción para detectar si están produciendo sesgos injustos. De esta manera, los errores de predicción detectados no solo adquieren una relevancia técnica, sino también ética.

16. **¿Si el origen de los datos de entrenamiento y validación que utiliza el sistema refleja la realidad del contexto de aplicación?**

    Es importante conocer la información detallada del origen de los datos de entrenamiento para poder reducir y prevenir la existencia de sesgos injustos. Si los datos de entrenamiento se han conseguido en un entorno geográfico diferente al de su uso, con diferentes poblaciones y condiciones socioeconómicas diferentes, las predicciones pueden ser erróneas, o puede provocar que determinados grupos queden

sobrerrepresentados o infrarrepresentados. Una opción es exigir a los desarrolladores o comercializadores del sistema con IA utilizado que informen detalladamente del origen de los datos de entrenamiento y validación para determinar si son realmente compatibles con el contexto en que queremos aplicarlo.

17. **¿Si, en el caso de procesos de toma de decisiones con un alto grado de automatización, se ha establecido un mecanismo de supervisión humana (lo que se conoce como *human-in-the-loop*)?**

    Se deben establecer mecanismos que permitan a una persona responsable tomar el control de las decisiones producidas por un algoritmo en cualquier fase del proceso de toma de decisiones. Esto incluye la capacidad de revertir las decisiones adoptadas de modo automático.

18. **¿Si se han establecido mecanismos de control para mitigar y reducir los riesgos en integridad de la base de datos y garantizar la seguridad de los datos utilizados?**

    Es necesario establecer mecanismos de seguridad que permitan proteger de ataques las bases de datos y evitar que los datos utilizados sean comprometidos o utilizados con fines diferentes de los previstos. Es importante tener en cuenta que las violaciones de ciberseguridad también podrían introducir sesgos.

19. **¿Si el equipo de diseño y desarrollo tiene suficiente conocimiento en el ámbito de la salud específicamente vinculado al sistema de IA?**

    A la hora de evitar la inclusión de sesgos en el diseño del sistema de IA se recomienda que los equipos de diseño y desarrollo tengan conocimiento suficiente o incluyan profesionales de la salud en el ámbito de la aplicación de sistemas de IA.

20. **¿Si el proceso de diseño y desarrollo ha sido supervisado y revisado por profesionales de la salud del ámbito de aplicación de la IA?**

    A la hora de evitar y prevenir la aparición de sesgos en el diseño del sistema, es necesario que, en los procesos de diseño y desarrollo, participen profesionales de la salud en el ámbito de la aplicación de sistemas de IA.

21. **¿Establecer mecanismos de revisión y actualización del sistema, con especial interés en momentos de emergencia sanitaria?**

    Se deben revisar y actualizar los sistemas de IA cada vez que se introduzca o se actualice cualquier elemento que pueda afectar o modificar al resultado arrojado por el sistema. Esto es especialmente determinante en momentos de emergencia sanitaria debido a la rapidez

con la que se desarrollan los sistemas que puede favorecer la aparición de ciertos riesgos como los sesgos algorítmicos. De manera análoga, aquellos sistemas que resultaban adecuados en momentos de normalidad pueden no serlo en situaciones de emergencia sanitaria.

Adecuación con la práctica clínica

22. **¿Si se incluyen predicciones o recomendaciones generadas por el sistema que vayan más allá del valor numérico?**

    Es necesario que los usuarios del sistema entiendan cuál es su propósito, y que las predicciones o recomendaciones generadas incluyan una explicación que permita a esos usuarios entender mejor esa predicción o recomendación y así juzgar si tiene sentido o no. Si, por ejemplo, una dermatóloga lee simplemente que "la probabilidad de melanoma es del 89%", su única opción es decidir si se cree lo que dice la máquina o no, y lo más fácil es que se lo acabe creyendo, pues las personas tendemos a asignarle un valor de objetividad a predicciones generadas mecánicamente. En cambio, si esa dermatóloga recibe además del valor numérico una serie de indicaciones explicando por qué es un melanoma (la mancha no es redonda, sino que tiene una forma muy irregular y su color no se corresponde a otras manchas de la piel del paciente) la dermatóloga tiene ahora una explicación que sí puede valorar, conectándola con su experiencia, y dispone de más criterios para decidir si confía en el diagnóstico automático o no.

23. **¿Si las personas usuarias del sistema conocen las condiciones de funcionamiento asumidas por los y las diseñadores del producto?**

    A veces sucede que el modelo que tienen en mente los desarrolladores del sistema no está alineado del todo con el problema clínico real, y eso se traduce tanto en funcionamientos erróneos del sistema como en la imposibilidad de interpretar correctamente las métricas de rendimiento o de cuantificación estadística de sesgos. Los escenarios de uso tenidos en cuenta durante la fase de diseño deben ser congruentes con la casuística real de la práctica médica. También por esta razón es importante integrar, al menos de manera consultiva, a profesionales sanitarios en los equipos de diseño y desarrollo.

24. **¿Si las recomendaciones o predicciones propuestas por el algoritmo son *prima facie* congruentes con la experiencia profesional sanitaria y estado del arte de la práctica clínica y se han comparado con otros métodos diagnósticos convencionales?**

    Se debe valorar si el sistema ha contado o es posible contar con personal sanitario capacitado con experiencia clínica previa para evaluar la congruencia de los resultados con los estándares de corrección y excelencia de la práctica médica. Asimismo, es importante que la

comunicación de estos sea adecuada, rigurosa y veraz, explicando su significado a personas expertas en la materia y a cualquier ciudadano que desee comprender la recomendación o predicción realizada

25. **¿El funcionamiento global del sistema puede explicarse en términos operativos generales a las diferentes partes implicadas de modo que este pueda ser evaluado por ellas?**

    Se recomienda desarrollar una explicación lo más sencilla posible dirigida a las partes implicadas sobre el funcionamiento del sistema. Esta explicación no debe contener excesivos términos técnicos y debe clarificar qué hace el sistema y cómo lo hace.

26. **¿Las razones que fundamentan las decisiones clave que se han tomado durante el proceso de diseño y desarrollo (acerca de recolección de datos, definición de parámetros, diseño y validación de modelos) son públicas y comprensibles por cualquier persona que quiera revisarlas?**

    Se recomienda hacer pública esta información y acercar el lenguaje especializado a las personas no expertas. Esta información puede acompañar a la documentación técnica del sistema.

27. **En el caso de que las personas usuarias necesiten un conocimiento previo específico para utilizar el sistema de IA o interpretar los datos que este arroja, ¿se provee de esta formación previa?**

    En el caso de que las personas usuarias necesiten conocimiento específico para interpretar los resultados se debe valorar la posibilidad de incorporar algún tipo de formación dirigida al público en general previa a la utilización del sistema.

28. **¿Está justificado el uso de la IA por no haber métodos convencionales que lleguen a resultados equivalentes?**

    Es necesario justificar de manera convincente la necesidad de utilizar un sistema con IA. Se debe valorar la pertinencia o no del uso del sistema de IA para el fin propuesto. La utilización de un sistema con IA puede estar justificada, por ejemplo, por ser los métodos convencionales muy costosos o porque el problema a resolver es complejo de evaluar por medios tradicionales.

29. **En el caso de que el sistema de IA haya sido auditado, ¿es posible acceder al informe o a sus conclusiones?**

    Es importante que la documentación técnica del sistema incluya eventuales auditorías o evaluaciones de relevancia ética que se hayan realizado durante el desarrollo, y que estas sean accesibles para su consulta.

Privacidad

30. En el caso de que se utilicen datos personales, ¿se han usado las medidas e instrumentos que exige la ley de protección de datos para preservar la privacidad de las personas implicadas?

Se debe revisar la legislación vigente en el país de aplicación del sistema de IA para determinar las medidas necesarias para gestionar la privacidad de los datos utilizados.

31. ¿Se ha contactado con la persona delegada de protección de datos (DPD/DPO)?

Las leyes para la protección de datos personales son complejas de seguir. Es una buena idea contactar con la persona responsable de esos temas en la institución de aplicación de la IA para así evitar problemas y errores posteriores.

32. Más allá del estricto cumplimiento de la legalidad vigente ¿Se está realmente protegiendo la privacidad y el derecho al consentimiento informado de las personas implicadas?

El hecho de que la legislación permita ciertas acciones no significa necesariamente que la acción sea éticamente correcta. Por ejemplo, en el caso de querer reutilizar los datos previamente obtenidos, tal vez algunas personas no se sentirán cómodas en participar en una nueva investigación y por ello es recomendable volver a obtener su permiso, si es posible.