Uso ético y responsable de inteligencia artificial en salud

Clara M. Mosquera Lopez[a], O. Lucía Quintero Montoya[b], Daniela Marín Ramírez[c], José J. Garcés[b], Carolina Sánchez Vásquez[d], Verónica M. Echeverri Salazar[e], Paula Roldán Maya[f], Jose A. Toro Valencia[d], Jaime Rugeles Ortíz[f], Juliana Ortiz Marín[c], Jesús F. Vargas Bonilla[g], María I. Vélez Agudelo[a]

[a] Centro para la Cuarta Revolución Industrial de Colombia, afiliado al Foro Económico Mundial
[b] Departamento de Ciencias Matemáticas, Universidad EAFIT
[c] Innovación EAFIT, Universidad EAFIT
[d] Departamento de Derecho, Universidad EAFIT
[e] Facultad de Derecho y Ciencias Políticas, Universidad de Antioquia
[f] Corporación Ruta N
[g] Facultad de Ingeniería, Universidad de Antioquia

Este documento fue revisado, analizado y modificado en algunos apartes teniendo en cuenta la visión y recomendaciones del Comité de Inteligencia Artificial de la Asociación Colombiana de Radiología (ACR).

Resumen

El uso de tecnologías de la Cuarta Revolución Industrial (4RI), particularmente de la inteligencia artificial (IA), se está expandiendo cada vez más rápido. Uno de los sectores en los que esta tecnología tiene mayor potencial de generar un impacto positivo es el de la salud humana. El rango de aplicaciones de la IA en ese contexto es amplio, dado que puede implementarse en múltiples actividades, incluyendo la toma de decisiones en la práctica clínica, intervenciones de salud pública, investigación biomédica, descubrimiento de nuevos medicamentos y tratamientos, así como en labores administrativas complementarias. Aunque la integración de la IA en el sistema de salud brinda la oportunidad de mejorar la eficiencia y la calidad de la prestación de los servicios, su implementación trae consigo riesgos que se deben entender y minimizar. Estos riesgos incluyen no sólo los posibles daños relacionados con la salud de los usuarios del sistema, sino también amenazas a la privacidad y confidencialidad, el consentimiento informado y la autonomía de los pacientes, y la profundización de inequidades presentes en la sociedad.

El desarrollo y uso de la IA ha propiciado espacios de discusión en el ámbito internacional a partir de los cuales se han publicado estrategias para el desarrollo y adopción de la tecnología y marcos para guiar su implementación ética y responsable. Canadá, por ejemplo, lanzó la primera estrategia nacional sobre IA en 2017, seguida poco después por países líderes en el desarrollo de la tecnología como Japón y China. La Organización para la Cooperación y el Desarrollo Económicos (OCDE), por su parte, presentó en 2019 un compendio de principios para la administración responsable de la IA. Otras iniciativas encaminadas a proponer reflexiones y recomendaciones para el uso responsable de la IA incluyen la llamada de Roma para la ética de la IA en febrero del 2020; la resolución del Parlamento Europeo expedida en octubre del 2020 con recomendaciones destinadas a la Comisión Europea sobre un marco de aspectos éticos de la

IA, la robótica y las tecnologías conexas; The Partnership on AI to Benefit People and Society; The Montreal Declaration for a Responsible Development of Artificial Intelligence; y la Declaración de Barcelona, entre otras. Todas estas iniciativas plantean diferentes objetivos, alcances, niveles de profundidad y sobre todo tienen diferentes entendimientos alrededor del desarrollo de marcos éticos, no obstante, constituyen un referente necesario para abordar el uso de la IA en áreas específicas, en este caso la salud humana. Actualmente varios países se han comprometido a crear consejos de ética de IA, entre ellos Alemania, Reino Unido, India, Singapur y México.

En Colombia se han dado algunos avances de políticas públicas sobre la IA. Específicamente, en el Plan Nacional de Desarrollo (PND) 2018-2022 se establece la necesidad de contar con un marco regulatorio que vaya en consonancia con los principios fundamentales del Gobierno, lo que posteriormente, condujo a la publicación del CONPES 3975 de 2019 y, en agosto de 2020, a la publicación de un documento para discusión sobre el Marco Ético para la IA en Colombia, el cual se ocupa de propiciar y orientar una discusión nacional sobre los principios éticos que deberán guiar el desarrollo de la IA en Colombia, y proponer herramientas para su implementación teniendo en cuenta los principios de privacidad, responsabilidad, seguridad, explicabilidad, justicia, control humano y respeto a los derechos humanos. Sin embargo, no existe un documento específico para el sector salud que contemple las particulares implicaciones éticas de la implementación de la IA sobre los derechos de los pacientes y usuarios del sistema de salud.

El presente documento, construido por un equipo interdisciplinario compuesto por profesionales de la salud, abogados, profesionales especializados en políticas públicas, matemáticos e ingenieros del Centro para la Cuarta Revolución Industrial de Colombia, la Universidad de Antioquia, la Universidad EAFIT y Ruta N, busca generar un mayor entendimiento de las implicaciones éticas que surgen al momento de desarrollar o implementar tecnologías de IA para aplicaciones en sector salud. A su vez, pretende ofrecer pautas para la creación de un marco regulatorio en el país que promueva el uso ético y responsable de la IA en dicho sector, teniendo como propósito superior el uso de la IA como un medio para el mejoramiento de la calidad de vida de los ciudadanos garantizando sus derechos fundamentales.

Inteligencia artificial y sus aplicaciones en el sector salud

La IA puede definirse como un conjunto de técnicas donde los algoritmos descubren o aprenden asociaciones de poder predictivo a partir de los datos. La forma más tangible de IA es el aprendizaje automático, que incluye el aprendizaje profundo que se basa en múltiples capas de representación de datos y, por lo tanto, pueden representar relaciones complejas entre los datos de entradas y la respuesta del sistema (Panch, Szolovits, & Atun, 2018).

El aprendizaje automático es un subconjunto de la IA que permite que los programas aprendan sin ser programados explícitamente. Estos programas aprenden de conjuntos de datos, identifican patrones dentro de ellos y usan la información para hacer predicciones sobre datos a los que no han estado expuestos anteriormente. Dentro del aprendizaje automático existe un subconjunto relativamente nuevo de técnicas denominadas aprendizaje profundo que utiliza algoritmos basados en redes neuronales más profundos en estructura y función (Helm et al., 2020). El aprendizaje profundo requiere mayores cantidades de datos que conducen a niveles más altos de precisión en algoritmos y esta es una

de las principales razones por las que el aprendizaje profundo ha revolucionado el campo de la IA en la actualidad.

El aprendizaje automático se divide en aprendizaje supervisado y no supervisado. En el aprendizaje supervisado, los algoritmos usan datos etiquetados para encontrar un modelo que, dadas las variables de entrada, asigne la etiqueta de salida adecuada. En los métodos de aprendizaje no supervisado un modelo se ajusta a las observaciones e identifica características estructurales de los datos que permitan agrupar dichas observaciones. Algunos ejemplos de métodos de aprendizaje no supervisados es la agrupación de puntos de datos bajo una métrica de similitud y la reducción de dimensionalidad de los datos; por ejemplo, la predicción de la enfermedad cardiovascular por agrupación de datos. Por otro lado, un ejemplo clásico de aprendizaje supervisado es la clasificación de nódulos pulmonares en malignos o benignos. Cada uno de estos métodos poseen ventajas y desventajas que influyen en su selección para diferentes tareas en el sector salud.

Medidas de desempeño de sistemas de IA

Algunos indicadores básicos de desempeño de modelos entrenados usando aprendizaje automático, particularmente los modelos para resolver problemas en que se requiere detectar la presencia o ausencia de una enfermedad o condición médica incluyen la sensibilidad, especificidad y el área bajo la curva de característica de operación del receptor (ROC por sus siglas en Inglés). La sensibilidad es la probabilidad de que un modelo clasifique una muestra/imagen/señal como positiva si la condición bajo estudio está presente (paciente enfermo o con patrón de referencia positivo). También se puede definir como la proporción de los casos positivos correctamente detectados por el algoritmo respecto al total de casos positivos. La especificidad es la probabilidad de que un modelo clasifique una muestra/imagen/señal como negativa si la enfermedad está ausente (paciente sano o con patrón de referencia negativo). También se puede definir como la proporción de casos correctamente clasificados como negativos por el algoritmo respecto al total de sujetos sanos. El área bajo la curva ROC (AUC) representa la validez global de un modelo de clasificación y mide el área bidimensional por debajo de la curva ROC. Un clasificador que discrimina perfectamente entre los dos grupos de pacientes (por ejemplo, clasificación normal vs. anormal), describiría una curva cuyo AUC es 1.0. La curva ROC de un clasificador totalmente inútil tiene un AUC menor o igual a 0.5. Entonces, en la práctica, un algoritmo debe tener un AUC entre 0.5 y 1.0. El valor mínimo valor de la AUC para demostrar un desempeño aceptable de las aplicaciones de IA en medicina sigue siendo motivo de debate, varía de acuerdo con su caso de uso en la práctica clínica. Sin embargo, entre más cercano es a 1.0 mejor es el desempeño del sistema.

Aplicaciones de la IA en el sector salud

En la atención en salud, las aplicaciones de IA pueden agruparse de acuerdo con la tarea que realizan: predicción, diagnóstico u optimización, en las cuales la tecnología actúa en conjunto con los profesionales de la salud o el personal administrativo. Algunos ejemplos de aplicaciones de IA incluyen: detección y clasificación de enfermedades a través de imágenes de radiología e imágenes histopatológicas, cirugía robótica, medicina de precisión, personalización de tratamientos, descubrimiento de nuevos medicamentos y terapias, sistemas de soporte de decisiones clínicas, sistemas de monitoreo remoto a pacientes, y pronóstico y predicción de enfermedades. Otras aplicaciones relacionadas con la investigación clínica incluyen la asistencia para encontrar pacientes y

candidatos para participar en investigaciones. De otro lado, la IA también puede aplicarse a labores administrativas complementarias incluyendo auditoría clínica, programación de citas, consultas y procedimientos médicos, codificación de información, facturación, pagos automáticos y detección de fraudes en el sistema de salud.

Contextualización

Las tecnologías de la información y la comunicación (TIC) y las tecnologías de la 4RI, entre ellas la IA, constituyen herramientas para promover la equidad en la prestación de servicios de salud a los ciudadanos.

En Colombia, como en la mayoría de los países en América Latina, existen marcadas diferencias en la calidad de la prestación de servicios de salud, situación que se hace más compleja en las zonas rurales y de difícil de acceso. Por ejemplo, los servicios médicos de diagnóstico y tratamiento de segundo y tercer nivel de complejidad se ofrecen principalmente en los centros urbanos, limitando el acceso de la población en zonas remotas a servicios de salud de primer nivel (International Telecommunication Union, 2014).

Un estudio comparativo del acceso a los servicios de salud en Colombia entre 1997 y 2012 evidenció que, pese a los esfuerzos por aumentar la cobertura del Sistema General de Seguridad Social en Salud (SGSSS), el acceso a los servicios se redujo. En aquellas regiones que presentaron una mayor demanda por servicios, también se tuvo una menor disponibilidad de sus prestadores y, como consecuencia, se presentó menor acceso durante el periodo estudiado (Ayala-García, 2014). Aunque en los últimos cinco años se ha incrementado la cobertura del SGSSS, con una afiliación del 95% de los ciudadanos[1], lo anterior refleja que existe una brecha entre la oferta y demanda de los servicios de salud que impide un acceso equitativo al sistema de salud.

La generación de grandes cantidades de datos resultantes de la creciente adopción de tecnologías digitales, incluyendo el uso de historias clínicas electrónicas centralizadas[2], la ubicuidad de dispositivos y aplicaciones móviles, biosensores que tienen la capacidad de recolectar datos relacionadas con diferentes aspectos de la salud humana, así como la disponibilidad de mejores herramientas computacionales para el análisis de los datos y el incremento del capital humano entrenado para su desarrollo y uso, representan una gran oportunidad para desarrollar modelos de predicción y sistemas de soporte de decisiones clínicas basados en IA que pueden ser usados para mejorar la calidad y cobertura de los servicios de salud. Estas tecnologías podrían democratizar el acceso de los pacientes al sistema de salud e incrementar la disponibilidad de recursos médicos al hacer más eficientes los procesos del cuidado de la salud mediante sistemas de monitoreo remoto, servicios de telemedicina, sistemas de soporte de diagnóstico, entre otros, los cuales permitirían acercar el conocimiento especializado que se encuentra concentrado en las zonas urbanas a las zonas más remotas del país.

[1] Información del Sistema integral de información SISPRO con corte a Diciembre de 2017.
[2] En Colombia, un hospital urbano de tamaño mediano genera anualmente cerca de 6.000 entradas de historias clínicas de hospitalización, más 95.000 imágenes radiológicas y cerca de 350.000 estudios de laboratorio clínico (Clínica CardioVid Medellín, 2019).

Según lo anterior, aunque el impacto de la implementación de la IA puede ser profundamente positivo en el sistema de salud colombiano, existen múltiples retos asociados a la implementación y uso generalizado de la IA. Uno de los retos más importantes tiene que ver con el manejo, la calidad y la suficiencia de los datos que son usados para el desarrollo de modelos de IA. Las aplicaciones médicas exigen altos estándares de privacidad y seguridad de los datos de pacientes (o participantes en estudios clínicos), sin embargo, al mismo tiempo existen líneas difusas en la definición de la propiedad de los datos a lo largo de las diferentes etapas de su cadena de procesamiento (por ejemplo: captura, anonimización o seudonimización, análisis, etc.).

Existen también preocupaciones éticas relacionadas con los sesgos y comportamientos discriminatorios que pueden ser incorporados durante el entrenamiento y reproducidos durante el uso de los sistemas de IA. La transparencia e interpretabilidad de las decisiones y resultados de los sistemas de IA, así como los criterios de optimización de los algoritmos presentan cuestiones éticas que necesitan ser resueltas para una adopción exitosa de la IA en el sector salud.

De otro lado, desde el punto de vista legal, no hay consenso sobre las líneas de actuación y el nivel de autonomía de los sistemas basados en IA y la responsabilidad civil que genera su utilización, lo cual genera una subjetividad legal que debe abordarse para poder establecer las responsabilidades jurídicas asociadas al uso de esta tecnología.

Considerando los riesgos señalados, es importante plantear la necesidad de un marco ético que ayude a gestionarlos para evitar afectaciones a derechos fundamentales.

Dado que en Colombia existen lineamientos para el establecimiento de un marco ético para la implementación de la IA en general, publicados en agosto de 2020 (Guío, 2020), no existe un marco ético y regulatorio exhaustivo que guíe el desarrollo de la IA más allá de las consideraciones técnicas y que promueva su implementación y uso responsable en el sector salud. Este documento aborda el análisis de algunos de los múltiples riesgos de la implementación de la IA en el sector salud y se propone como una herramienta complementaria, producto del consenso entre diferentes actores con conocimiento e interés en la adopción y uso ético de la IA y los datos en el sector salud, que plantea recomendaciones para llenar vacíos regulatorios y lineamientos que permitan transferir de manera efectiva los desarrollos en IA a aplicaciones en el sector salud con los estándares éticos necesarios para asegurar que su implementación agrega valor a los pacientes y a la sociedad.

Retos e implicaciones éticas de la implementación de la IA en el sector salud

Como se plantea en el Marco Ético para la Inteligencia Artificial en Colombia (Guío, 2020), la implementación de la IA presenta retos que se deben abordar de forma tal que se puedan balancear los beneficios y los riesgos que presenta la tecnología. En el sector salud, los retos aplicables incluyen los siguientes: (i) amenazas a la autonomía de los pacientes, (ii) responsabilidad de los profesionales de la salud; (iii) sesgos, discriminación y exclusión; (iv) pobre calidad, cantidad y relevancia de los datos usados para el entrenamiento de sistemas de IA; (v) generación de evidencia clínica; y (vi) amenazas a la privacidad de los pacientes.

Las amenazas al principio de autonomía del paciente se presentan cuando éste no es informado sobre la presencia de un sistema de IA guiando la toma de decisiones relacionadas con su salud. Existen

preocupaciones relacionadas con la vulneración de derecho de los pacientes de ser informados y decidir libremente sobre cómo son tratados dentro del sistema de salud. Por ejemplo, en Estados Unidos, las decisiones sobre planes de alta de miles de pacientes hospitalizados en uno de los hospitales pertenecientes a uno de los sistemas más grandes del estado de Minnesota fueron informadas con la ayuda de un sistema de IA, pero muy pocos pacientes fueron informados sobre el involucramiento de este sistema en el cuidado de su salud. Así mismo, existe un número creciente de hospitales y clínicas que usan herramientas de apoyo a decisiones clínicas basadas en IA, muchas de ellas no probadas, para predecir si es probable que los pacientes hospitalizados desarrollen complicaciones o se deterioren, si están en riesgo de readmisión y si es probable que mueran pronto. No obstante, a estos pacientes y sus familiares a menudo no se les informa ni se les pide su consentimiento para el uso de estas herramientas en su cuidado. A algunos médicos les preocupa que la decisión de no mencionar estos sistemas de IA aparte de amenazar el principio de autonomía de los pacientes podría ser contraproducente si se presentan daños en su salud que puedan ser atribuidos al uso de estos sistemas de IA (Robbins & Brodwin, 2020).

El segundo reto señalado, tiene que ver con la responsabilidad de los profesionales de la salud, quienes deben estar en capacidad de comprender el funcionamiento de los sistemas de IA y sus limitaciones, de forma que puedan evitarse escenarios en los que se presenten afectaciones a la salud de los pacientes por decisiones médicas basadas en una confianza excesiva en las recomendaciones de sistemas de IA.

Un tercer reto tiene que ver con el sesgo algorítmico*.* Los modelos de IA pueden reproducir heurísticas prejuiciosas y prácticas no ortodoxas así como profundizar desigualdades existentes presentes en los datos usados para el entrenamiento de los algoritmos o por la selección de alternativas de modelado incorrectas (Chen, Szolovits, & Ghassemi, 2019). En general, el sesgo algorítmico ocurre cuando un sistema de IA toma decisiones que tratan a individuos en situaciones similares de manera diferente cuando no hay una justificación para tales diferencias, independientemente de la intención. Algunos ejemplos de este problema incluyen algoritmos de predicción de la mortalidad hospitalaria con precisión variable según el origen étnico (Chen, Johansson, & Sontag, 2018) y algoritmos que pueden clasificar imágenes de lunares benignos y malignos con precisión similar a la de los dermatólogos (Esteva et al., 2017a, 2017b) (Haenssle et al., 2018), pero con bajo rendimiento en imágenes de lesiones en la piel de pacientes con piel oscura debido a que el entrenamiento del algoritmo se realizó usando conjuntos de datos abiertos de pacientes predominantemente de piel clara (Kelly, Karthikesalingam, Suleyman, Corrado, & King, 2019). Los pacientes que reciben respuestas erróneas por parte de sistemas de IA usados para asignación de riesgo o diagnóstico de enfermedades tienen menos probabilidades de poder acceder al estándar de atención necesario y mayores probabilidades de experimentar efectos adversos como resultado de que se les retrase o se les niegue la atención adecuada.

El cuarto reto señalado tiene que ver con la calidad, cantidad y relevancia de los datos para entrenamiento y validación del desempeño de los sistemas de IA. Si un algoritmo se entrena en un conjunto de datos que contiene un tipo específico de pacientes y éste se aplica a otros tipos de pacientes puede generar resultados erróneos, ya que el algoritmo no ha aprendido las características de la nueva población donde se aplica.

Un quinto reto es la generación de evidencia clínica a partir de las aplicaciones de IA. Idealmente, el estudio clínico controlado de asignación aleatoria es el método de referencia en salud para validar con datos reales los sistemas de IA y los criterios primarios de valoración del estudio deben ser los

resultados clínicos por encima de los niveles de precisión del modelo de IA. Esto es importante dado que se puede tener un sistema de IA con predicciones muy precisas, pero que no tenga impacto significativo en los resultados de salud de los pacientes (Keane & Topol, 2018).

Finalmente, existen retos asociados a la privacidad. Las preocupaciones en este sentido van por dos caminos. Primero, el proceso de recopilación de información y el intercambio de información entre las entidades del sistema de salud y los desarrolladores de sistemas de IA que podrían llegar a representar violaciones de la privacidad de los pacientes. Segundo, la IA podría implicar violaciones a la privacidad cuando ésta es capaz de predecir información privada sobre los pacientes, aunque el algoritmo nunca haya recibido esa información.

Los retos ejemplificados anteriormente ponen de manifiesto la necesidad de abordar integralmente los posibles beneficios y los riesgos asociados a la implementación de la IA en el sector salud a través de un marco ético y regulatorio con consideraciones especiales para el sector.

Principios éticos para la implementación de la IA en el sector salud

Los principios para la implementación ética de la IA en el sector salud guardan entre sí una estrecha relación, en tanto son complementarios entre ellos, para garantizar el respeto de los valores sociales y los derechos humanos ya reconocidos y protegidos en las sociedades modernas. En esta línea hacen parte del entramado ético de la IA, principios de diferente orden. En un primer orden, se deben situar los principios en términos genéricos, esto es, que rigen en cualquier ámbito de aplicación de esta tecnología. Aquí, sobresalen los principios éticos de la trasparencia y rendición de cuentas, la justicia, la no discriminación y la privacidad (Fjeld, Nele, Hilligoss, Nagy, & Srikumar, 2020) (Organisation for Economic Co-operation and Development, 2020), los cuales, son complementados en un segundo orden, en el campo de la salud, con los principios de no maleficencia, beneficencia, autonomía y justicia (Floridi & Cowls, 2019).

A continuación, se presentan las definiciones de los principios de segundo orden (Beauchamp & Childress, 1979) desde los cuáles se abordan las cuestiones éticas asociadas al uso de la IA en el sector salud, los cuales amplían los principios de privacidad, seguridad, transparencia y explicación, justicia y no discriminación, responsabilidad, control y supervisión humana y beneficio social considerados en Marco Ético para la Inteligencia Artificial en Colombia en su versión preliminar.

El **principio de no maleficencia** es considerado el más importante en el área de la salud humana, y expresa que cualquier acto médico debe pretender en primer lugar no hacer daño alguno, de manera directa o indirecta.

El **principio de beneficencia** está cercanamente relacionado con el principio de no maleficencia, y se refiere a que los actos médicos deben tener la intención de producir un beneficio para el paciente.

El **principio de autonomía** tiene que ver con el derecho del paciente de decidir por sí mismo sobre los actos que se practicarán en su propio cuerpo y que afectarán de manera directa o indirecta su salud, su integridad y su vida.

El **principio de justicia** se refiere a la equidad en la distribución de cargas y beneficios. El criterio para saber si una actuación es o no ética desde el punto de vista de la justicia, es valorar si la actuación es equitativa, es decir, debe ser posible para todos aquellos que la necesiten. Esto incluye el rechazo a la discriminación por cualquier motivo.

Controles técnicos y éticos para la implementación de la IA en el sector salud

En esta sección se proponen controles técnicos y éticos para la implementación de la IA en el sector salud desde el punto de vista de los datos, los algoritmos y la práctica, como una herramienta para mitigar riesgos asociados a la vulneración de derechos fundamentales de los usuarios del sistema de salud y minimizar posibles afectaciones a la salud de los pacientes.

Desde el punto de vista de los datos

Controles técnicos

En sistemas basados en IA, las garantías de aprendizaje se cumplen con base en características de los datos que se usen para tal propósito, su cantidad y calidad. Con base en lo anterior, para lograr una correcta representación del fenómeno que se está modelando por medio IA, un correcto muestreo de los datos garantiza la potencial inclusión de todos los modos de operación del sistema o proceso; esencialmente, mitigará la eliminación de descriptores o características asociadas con sesgos o discriminación. Un incorrecto uso de la densidad de probabilidad para recoger dichas muestras puede potencialmente traer problemas en los algoritmos desarrollados. Además, el número de muestras debe ser cuidadosamente elegido con el fin de mitigar el riesgo de sobreajuste en los algoritmos[3] y desarrollar un modelo capaz de generalizar correctamente nuevas instancias.

En el área de la salud, esto se traduce en la relevancia práctica del diseño de los estudios clínicos y los protocolos de recolección de datos tomando en cuenta factores etnográficos y epidemiológicos que, a la luz de la IA y la teoría de aprendizaje, deben ser revisados por el grupo matemático e ingenieril de modo tal que se logre satisfacer las condiciones para el problema de optimización que se resuelve por la de vía supervisión u otros métodos.

Cumplidas las cotas de aprendizaje y guardadas las consideraciones estadísticas del muestreo que mitigan los sesgos de los datos, aparecen otros elementos que sirven de insumos en algoritmos de supervisión y son esencialmente procesos de etiquetado. En la práctica, para el desarrollo de sistemas de IA para aplicaciones en el sector salud, se requiere conocimiento experto que sea representativo del área que se desea emular. Idealmente se propone que exista un panel de expertos que vía triangulación o consenso establezcan las etiquetas que puedan servir para procesos de clasificación. El proceso de

[3] El sobreajuste se refiere a un modelo que memoriza los datos con que ha sido entrenado, incluyendo el ruido presente en los datos. Esto tiene un impacto negativo en el rendimiento del modelo cuando se aplica a datos nuevos. En términos prácticos, un modelo sobreajustado tiene un desempeño casi perfecto en los datos de entrenamiento y un desempeño significativamente inferior cuando es aplicado a nuevos casos a los que el modelo no ha sido expuesto durante el proceso de aprendizaje.

triangulación puede evitar potenciales sesgos cognitivos o sesgos de discriminación que puedan hacer parte del proceso de etiquetado.

En el caso de sistemas de IA que sean entrenados por vía de algoritmos de no supervisión debe establecerse una metodología de supervisión de los resultados de los algoritmos que satisfaga los teoremas de muestreo (error real). En este caso, las metodologías cualitativas permiten hacer un correcto análisis de los resultados y las metodologías cuantitativas favorecen la calificación de los resultados.

Controles éticos

Al momento de incorporar datos e información de seres humanos y en especial de pacientes en potencial condición de vulnerabilidad, el grupo desarrollador del sistema de IA debe garantizar el diseño de un protocolo claro de la investigación o implementación, que incluya los aspectos clínicos/médicos y que especifique el propósito para el cual serán usados los datos. Este protocolo debe contener un sustento técnico robusto que cubra aspectos de privacidad, seguridad y confidencialidad de la información de los pacientes que permita garantizar el respeto los derechos humanos relacionados con la privacidad y la protección de datos personales. Cada momento y actividad dentro de la cadena de valor de los datos que incluye la recolección, procesamiento y uso de los mismos, requiere responsabilidad y custodia, que en muchos casos puede lograrse mediante el uso de técnicas de anonimización, seudonimización o de-identificación que blindan los datos para que terceros no puedan identificar a personas ni vulnerar sus derechos. Sin embargo, es muy difícil garantizar que a partir de algunos datos que han sido desprovistos de su identificación no se pueda llegar a saber de quién son al cruzarlos con otras bases de datos. En este contexto cobra sentido la implementación de tecnologías de encriptación o *blockchain.*

Para garantizar que se cumplan los principios de no maleficencia, beneficencia y justicia, los protocolos relacionados con el diseño e implementación de sistemas de IA en el sector salud deben ser aprobados por el comité de ética de las instituciones involucradas en la investigación o despliegue de estos sistemas o un organismo que garantice que la implementación cumple con los estándares éticos aplicables.

Finalmente, para operacionalizar el principio de autonomía en cuanto el uso de datos para fines de diseño, implementación y pruebas de sistemas de IA, se debe realizar el procedimiento de consentimiento informado a pacientes o participantes de estudios clínicos a través del cual se garantiza que éstos han recibido y comprendido la información sobre el propósito del uso de sus datos, sus beneficios y posibles riesgos.

Desde el punto de vista de los algoritmos

Controles técnicos

Para que un sistema de IA pueda ser implementado en la práctica clínica debe lograr una generalización confiable, es decir, debe tener un desempeño adecuado cuando procesa nuevos datos que no han sido utilizados durante el proceso de entrenamiento del algoritmo. Un modelo no generalizable puede tener

puntos ciegos que pueden producir decisiones erróneas que conlleven, por ejemplo, a un error de diagnóstico que a su vez produzca la selección incorrecta del tratamiento que se debe aplicar a un paciente. La generalización puede ser difícil debido a diferencias en las características de la población, diferencias técnicas entre instituciones, así como variaciones en las prácticas clínicas y administrativas locales.

Para mitigar los riesgos asociados a errores de generalización de los algoritmos es necesario evaluar exhaustivamente el desempeño del sistema de IA, lo cual requiere una validación adecuadamente diseñada que implique la prueba del sistema utilizando conjuntos de datos de tamaño adecuado, recopilados en instituciones distintas de las que proporcionaron los datos para el entrenamiento del modelo.

Adicionalmente, para la implementación de la IA para aplicaciones en el sector salud, es importante realizar ajustes al sistema de IA con datos de la institución donde se va a implementar la solución para adaptar el modelo de forma que responda adecuadamente a las características específicas de la nueva población si esta es diferente a la que se usó para el entrenamiento del sistema; proporcionar métodos para detectar entradas que están fuera de la distribución de los datos usados para el entrenamiento y ajuste del sistema de IA, así como también una medida del nivel de confianza del sistema frente a sus predicciones.

Para mitigar los riesgos de discriminación y daño asociados a sesgos algorítmicos se debe realizar un análisis cuidadoso del desempeño de sistemas de IA a la luz de la etnografía, es decir, por subgrupos de población incluyendo edad, etnia, sexo, estrato socioeconómico, entre otros (por ejemplo, analizando diferencias estadísticamente significativas en los errores producidos por el modelo cuando es evaluado en los subgrupos de interés). Es clave asegurar que el sistema de IA tenga un desempeño comparable en cada uno de los subgrupos y si este no es el caso, documentar o corregir las limitaciones del sistema identificando e incluyendo variables adicionales para mejorar su desempeño en poblaciones donde tiene un desempeño subóptimo.

Ante la posible combinación de consecuencias de resultados erróneos, el alto riesgo de sesgos no cuantificados difíciles de identificar a priori y el potencial riesgo de uso de variables tangenciales que pueden causar confusiones en el diagnóstico o recomendaciones dadas por un sistema de IA, la explicabilidad y la transparencia son claves para la verificación del sistema. Esto mejora la capacidad de los expertos para reconocer errores del sistema, detectar resultados basados en razonamiento inadecuado, e identificar el trabajo requerido para eliminar sesgos. En otras palabras, la explicabilidad y la transparencia guardan una estrecha relación con la necesidad de verificar que un sistema IA pueda prevenir eficazmente la distorsión, la discriminación, la manipulación y otras formas de uso indebido, para lo cual se sostiene que los sistemas de IA deberían proporcionar detalles suficientes sobre sus operaciones con el fin de que las mismas puedan ser validadas[4].

En general, para mejorar la transparencia de los sistemas de IA, se deben proveer mecanismos de explicación de cada una de sus salidas. Algunos algoritmos diseñados para tal fin incluyen Local

[4] En la actualidad, existe una tensión entre desempeño y explicabilidad. Los modelos de IA con mejor rendimiento (por ejemplo, aquellos basados en aprendizaje profundo) son a menudo los menos explicables, mientras que los modelos con desempeño más bajo (por ejemplo, regresión lineal, y árboles de decisión) son los más explicables.

Interpretable Model-Agnostic Explanations (LIME) (Ribeiro, Singh, & Guestrin, 2016), SHapley Additive exPlanations (SHAP) (Lundberg & Lee, 2017) y explicaciones contrafactuales (Wachter, Mittelstadt, & Russell, 2017), entre otros. Las explicaciones deben permitir que el personal médico utilice objetivamente la ayuda de modelos de IA para tomar sus propias decisiones, no que confíen ciegamente en el modelo.

Además de explicar el porqué de sus salidas, los sistemas de IA deben reportar las métricas comunes de desempeño de los modelos (por ejemplo: AUC, sensibilidad, especificidad, valores predictivos) de forma tal que los profesionales de la salud puedan comprender cómo los algoritmos propuestos podrían impactar al paciente dentro del flujo de atención usando mecanismos de teoría de decisión. Enfoques potenciales sugieren el uso de curvas de decisión, que tienen como objetivo cuantificar el beneficio neto de utilizar un modelo para guiar intervenciones clínicas y acciones posteriores (Vickers, Cronin, Elkin, & Gonen, 2008) (Mosquera-Lopez et al., 2020).

Adicionalmente, se debe proveer a los sistemas de IA con mecanismos de trazabilidad que permitan, bajo demanda, recuperar las salidas del sistema, las entradas que produjeron dichas salidas y la explicación dada por el sistema sobre el razonamiento que produjo los resultados. Esto implica mantener y monitorear las versiones y actualizaciones del sistema de IA, particularmente si este se adapta con el tiempo.

Finalmente, los algoritmos y sistemas de IA aplicados en el sector salud deben ser seguros, fiables y sólidos (Organisation for Economic Co-operation and Development, 2020); y además deben ser diseñados contemplando la posibilidad de ciberataques, ataques por manipulación de entradas (Finlayson et al., 2019) y fallos técnicos, de forma que se puedan prevenir afectaciones a la salud de los pacientes. Para mitigar los riesgos asociados a este tipo de ataques es importante proveer los sistemas con mecanismos de verificación y estandarización de las entradas, de forma que no se tomen decisiones basadas en entradas incoherentes o de baja calidad. Esto puede poner en riesgo la garantía de los principios de no maleficencia y beneficencia si se materializan riesgos asociados por ejemplo a diagnósticos erróneos o asignación equivocada de niveles de riesgo de pacientes que conlleven a tratamientos inadecuados.

Controles éticos

Para garantizar que se cumplan los principios éticos desde el mismo diseño y desarrollo de los algoritmos, es importante capacitar a los diseñadores de algoritmos de IA en aspectos relacionados no sólo con su entrenamiento (por ejemplo, selección de variables y tipo de modelo, algoritmo de entrenamiento, optimización de hiper parámetros, entre otras consideraciones) y testeo exhaustivo, sino también en las implicaciones éticas del uso de estos algoritmos en el área de la salud. Además, se debe empoderar a los médicos y personal de la salud para participar críticamente en el diseño y desarrollo de los sistemas de IA para guiar a los desarrolladores a garantizar que se toman los pasos correctos para cuantificar el sesgo algorítmico e identificar posibles escenarios de discriminación, así como validar los mecanismos de explicabilidad que habilitan la transparencia de los sistemas de IA.

Otra consideración ética importante es poner a la comunidad y las necesidades del usuario final en el centro del desarrollo de los algoritmos. En este sentido, es importante que quienes van a ser impactados por la implementación de sistemas basados en IA participen activamente desde el diseño y planeación de la implantación de sistemas de IA, de forma que el equipo de desarrollo aborde todas las consideraciones éticas desde la protección de sus derechos hasta la evaluación de las consecuencias no

deseadas (Celi et al., 2020) (MacPherson & Pham, 2020). En línea con el planteamiento anterior, grupos como el MIT Critical Data recomiendan un enfoque al que se refieren como ética embebida o integrada en el desarrollo e implementación de la IA (McLennan et al., 2020). La ética embebida es un enfoque altamente colaborativo e interdisciplinario que permite la formulación de estándares de práctica para establecer un mínimo de calidad metodológica en la formalización de la ética en IA como disciplina.

Desde el punto de vista de la práctica

Controles técnicos

Los sistemas de IA no son perfectos y pueden cometer errores que tengan como consecuencia daños al paciente u otros problemas en la atención médica. Por supuesto, muchos problemas en la atención de pacientes ocurren debido a errores médicos en el sistema de atención actual, incluso sin la participación de la IA, aún así, hoy se aceptan los riesgos en la atención médica por errores humanos pero para el caso de la IA se espera, de algún modo, la perfección. Esto se explica porque los errores asociados a la IA en la práctica médica podrían tener un impacto negativo mucho mayor si el sistema de IA es de uso masivo, resultando en daño a muchas personas en lugar del número limitado de pacientes afectados por un error de una sola persona (Price, 2019). Un ejemplo de este escenario es cuando un sistema se utiliza para cuantificar el riesgo de los pacientes de desarrollar enfermedades crónicas. Errores en este caso pueden conducir a que ciertos pacientes reciban atención médica de estándares más bajos, por ser considerados de bajo riesgo por un algoritmo de IA.

Para mitigar el riesgo de posibles daños a la salud del paciente causados por asignación errónea de tratamientos médicos, se propone que los sistemas de IA deben ser supervisados y entendidos por los humanos a un nivel consecuente con la gravedad de las consecuencias en la salud de los pacientes resultado de un error de diagnóstico por parte de un algoritmo de IA. Por ejemplo, sistemas de alto riesgo como aquellos que realizan diagnóstico de enfermedades en disciplinas como la radiología y la patología que definen el curso de tratamiento para el paciente, deben tener el menor nivel de autonomía para tomar decisiones clínicas, garantizando así el respeto a los principios de no maleficencia, beneficencia y justicia.

Controles éticos

Desde el punto de vista del principio ético de autonomía, considerando que es requerido en los consentimientos informados, los pacientes deben ser informados sobre la inclusión de sistemas de IA dentro del flujo de la atención en salud, el propósito de su uso, su nivel de autonomía, sus posibles beneficios, limitaciones y riesgos y las alternativas de tratamiento. Esta práctica genera confianza en el paciente y promueve la adopción de la tecnología (Schiff & Borenstein, 2019).

Por otro lado, los avances en la tecnología de IA hacen que los sistemas sean cada vez más complejos. Esta complejidad, que en general va ligada a mejores niveles de desempeño (por ejemplo, el caso de los algoritmos basados en redes neuronales de múltiples capas), puede eventualmente requerir conocimientos técnicos avanzados por parte del personal de la salud, no sólo para entender las salidas de los sistemas, sino también para poder explicarlas a sus pacientes. En consecuencia, la educación

médica deberá preparar a los proveedores de servicios de salud para manejar, evaluar e interpretar los sistemas de IA que encontrarán en el entorno de atención médica que se encuentra en constante evolución.

Finalmente, la integración de sistemas de IA en la práctica clínica requiere la construcción de una relación mutuamente beneficiosa entre el personal de la salud y la IA, en la cual ésta última ofrece mayor eficiencia a cambio de exposición para aprender a gestionar casos clínicos más complejos con el tiempo. Durante todo el proceso, es fundamental garantizar que la IA no oscurezca el rostro humano de la medicina, de forma que se pueda lograr la construcción de confianza en la IA como un elemento en la nueva relación médico-IA-paciente (Buch, Ahmed, & Maruthappu, 2018).

Antecedentes regulatorios sobre sistemas de IA con aplicaciones en el sector salud

Referente internacional

Para el análisis de la reglamentación y normativa a nivel internacional será abordado como referente la Administración de Medicamento y Alimentos (FDA por sus siglas en inglés) de los Estados Unidos, por sus avances en materia de reglamentación en salud digital, sus aportes en el IMDRF y su historial en aprobación de dispositivos basados en IA, que se ha visto acelerado gracias a su Centro de Dispositivos y Salud Radiológica (CDRH) creado en el año 2017 (U.S. Food and Drug Administration, 2017). Actualmente, la FDA evalúa los sistemas basados en IA dentro de la categoría *software como dispositivo médico* (SaMD). Sin embargo, la misma entidad enfatiza la necesidad de revisar y modernizar la regulación aplicada a sistemas de IA dado que la IA y en especial el aprendizaje automático difieren de SaMD en que tiene la capacidad de “aprender” y puede adaptarse con el tiempo. Por lo tanto, la regulación de dispositivos médicos no aplica para esta tecnología, pues una vez autorizado su uso está pueden tener cambios significativos en sus resultados debido al proceso de aprendizaje, es por ello que desde el 2019 se ha empezado a debatir un marco regulatorio aplicable a tecnologías basadas en IA para llenar los vacíos reglamentarios anteriormente mencionados (U.S. Food and Drug Administration, 2019).

Para la creación de este nuevo enfoque reglamentario para la IA en salud basado en el riesgo que representa el sistema de IA para los usuarios y pacientes, se siguen las recomendaciones y las definiciones del Foro Internacional de Reguladores de Dispositivos Médicos (IMDRF según sus siglas en inglés), que define software como un dispositivo médico como software destinado a ser utilizado para uno o más propósitos médicos que cumplen estos enfoques sin ser parte de un dispositivo médico de hardware. Para el marco normativo de este tipo de dispositivos, se establece un espectro del impacto para los pacientes que es categorizado según el nivel del riesgo, donde se identifica el uso previsto de la información proporcionada por el mismo (International Medical Device Regulators Forum, 2014). En conjunto, los factores que describen el uso previsto del dispositivo son clasificados en cuatro categorías en el caso del software como dispositivo médico, que van desde el menor (I) hasta el mayor riesgo (IV), y reflejan el peligro asociado con la situación clínica y el uso del dispositivo, como lo muestra la Tabla 1.

Estado de la situación o condición de atención médica	Relevancia de la información entregada por el SaMD para decisiones médicas		
	Trata o diagnostica	Conduce el manejo clínico	Informa el manejo clínico
Crítico	IV	III	II
Serio	III	II	I
No serio	II	I	I

Tabla 1. Categorización de riesgos de Software como dispositivo médico según IMDRF, 2014.

Referente nacional

En la actualidad la reglamentación colombiana no hace referencia explícita a dispositivos médicos basados en IA, dejando vacíos relacionados con la adopción de estas tecnologías en el sistema de salud. Sin embargo, los sistemas basados en IA pueden ser asignados inicialmente dentro de la categoría de software como dispositivo médico basado en IA, según las definiciones establecidas por el marco normativo internacional IMDRF, siempre y cuando cumpla los usos de: *"**diagnóstico, prevención, supervisión, tratamiento o alivio de una enfermedad...**"* según el artículo 27 del decreto 4725 de 2005 por el cual se reglamenta el régimen de registros sanitarios, permiso de comercialización y vigilancia sanitaria de los dispositivos médicos para uso humano. Esta clasificación enmarca una serie de requisitos técnicos para la certificación de registro sanitario basados en el nivel del riesgo que representa el dispositivo, clasificado en 4 categorías, siendo la Clase I la de menor riesgo y la clase III de alto riesgo, sujetos a controles adicionales (ver Tabla 2).

Requisitos técnicos	Clase I	Clase IIA	Clase IIB	Clasel II
Descripción del Dispositivo Médico	✔	✔	✔	✔
Estudios Técnicos y comprobaciones analíticas. •Verificación y validación de diseño. •Certificado de análisis del producto terminado.	✔	✔	✔	✔
Método de esterilización	✔	✔	✔	✔
Método de desecho o disposición final	✔	✔	✔	✔
Estudios de biocompatibilidad, estabilidad, citotoxicidad, seguridad eléctrica.		✔	✔	✔
Análisis de Riesgos		✔	✔	✔
Descripción de medidas de seguridad		✔	✔	✔
Estudios Clínicos			✔	✔
Certificación de Compromiso: entregarán al usuario final el manual de operación o usuario los cuales se encuentran disponibles en idioma castellano y tendrá disponibles los manuales de mantenimiento y operación		✔	✔	✔

Tabla 2. Requisitos técnicos para dispositivos médicos en Colombia. Fuente: Instituto Nacional de Vigilancia de Medicamentos y Alimentos (INVIMA), 2005.

Los requisitos técnicos son dados de manera general para todos los dispositivos médicos, y es allí donde se hace necesario establecer algunos parámetros diferenciales para el software basado en la IA y

aprendizaje automático como dispositivo médico que permitan incorporar la seguridad, privacidad y en general, mitigar los potenciales riesgos éticos desde las etapas de diseño, desarrollo y fabricación de estos dispositivos.

Consideraciones para establecer un marco regulatorio para la implementación de la IA en el sector salud en Colombia

Ambiente regulatorio para la IA en Colombia

El gobierno colombiano suscribió la Recomendación del Consejo de OCDE sobre IA en marzo de 2019, con lo cual adhiere a un conjunto de recomendaciones directrices de políticas intergubernamentales sobre IA que pretenden promover la implementación de una serie de principios para una administración responsable y confiable de la IA, respetando los derechos humanos y los valores democráticos.

En Colombia se está generando un ambiente propicio para la implementación de la IA a través de la creación de marcos éticos y regulatorios para la IA. Algunos resultados importantes de esta apuesta que impactan al sector salud son: (i) la reglamentación de la historia clínica electrónica interoperable (Ley 2015 del 31 de Enero de 2020); (ii) la definición de los requisitos de interoperabilidad de los datos consignados y el planteamiento de la necesidad de establecer un marco ético regulatorio flexible que incentive el desarrollo e implementación de tecnologías de IA en el sector salud, al tiempo que se estimule la adopción de mecanismos de experimentación regulatoria en el CONPES 3975 de 2019 para la transformación digital con énfasis en IA; y (iii) la publicación de la versión preliminar de un marco ético para la IA por parte de la Consejería Presidencial para asuntos económicos y de transformación digital en agosto de 2020.

Consideraciones para establecer un marco regulatorio para la IA en el sector salud

Para la implementación ética y responsable de la IA en el sector salud, el mayor reto que enfrentan los reguladores es analizar y diseñar medidas que permitan mejorar la experiencia y los resultados clínicos de los pacientes, mejorar la salud de las poblaciones, disminuir el costo de salud per cápita en beneficio de las comunidades. Adicionalmente, dado que este documento se ocupa de la implementación de la IA en salud, es importante también que se creen políticas que promuevan la innovación, pero que a su vez brinden reglas claras sobre el rol y responsabilidad de las empresas desarrolladoras de tecnología, los mecanismos para la protección de los derechos de los pacientes, y a intervención de los organismos de control. Un marco regulatorio claro debe ofrecer seguridad jurídica y garantizar el respeto por los derechos de los ciudadanos.

Las medidas de regulación deben ser acordes con el riesgo que representa el uso de sistemas basados en IA para el bienestar de los pacientes y de la sociedad en general. En este sentido, es posible proponer algunos criterios respecto de cuáles ámbitos (ver Tabla 3) y aplicaciones (ver Tabla 4) deberían estar sujetos a regulación y a autorregulación. Por ejemplo, los escenarios en los que se requeriría regulación externa son aquellos aspectos más críticos para garantizar el respeto a los derechos humanos o aquellos que pueden comprometer la seguridad física de las personas. Por su parte, los asuntos que se autorregulan son aquellos que no tienen implicaciones directas sobre la vida del paciente.

La autorregulación es el conjunto de procesos a través de los cuales los actores -y en especial los aglutinados en cierto sector e industria- construyen y fijan procedimientos reglas y normas, formales e informales, orientadas a fijar las conductas a seguir en una materia, tema o sector determinado. La autorregulación se genera en un espacio intermedio entre el Estado y el mercado e involucra el conocimiento especializado de gremios, centros de pensamiento y expertos. La autorregulación parte del supuesto que quienes la promulgan son las autoridades en la materia y por ello tienen la capacidad de prescribir las mejores conductas, prácticas y procesos para un sector o materia específico. En esta materia se llevaría a cabo por los desarrolladores y operadores de los distintos dispositivos y aplicaciones médicas basados en inteligencia artificial y sus instituciones aglutinadoras. La clasificación de los distintos dispositivos médicos es un criterio para determinar el nivel de riesgo de los mismos y puede servir como criterio para determinar si en este caso es sujeto a regulación o autorregulación.

Aspectos a regular	Ejemplo	Regulación	Autorregulación
Afectación de derechos fundamentales	Negación de servicios de remisión a tercer nivel	X	
Cuestiones y decisiones técnicas	Utilización de servidores vs. servidores en la nube		X
Responsabilidad por daños causados por IA	Error médico debido a falso positivo generado por detección de nódulo pulmonar	X	
Selección de autoridades competentes	Sanción de comité de ética médica vs. demanda civil por errores relacionados con IA	X	
Competencias de las empresas y la sociedad civil en la IA	Limitaciones de las empresas	X	X
Competencias de los pacientes en cuando a decisiones relacionadas con IA	Decisión de no aceptar manejo de IA en las conductas clínicas	X	

Tabla 3. Escenarios sujetos a regulación externa o autorregulación.

Aplicaciones	Ejemplo	Regulación	Autorregulación
Diagnóstico de enfermedades	Detección de nódulos pulmonares	X	
Tratamiento de enfermedades	Uso de IA para control de sistemas como páncreas artificial para el tratamiento de diabetes tipo 1	X	
Trámites administrativos	Cancelación de citas		X
Monitoreo remoto de pacientes	Sistemas de apoyo en tele-radiología		X
Apoyo en toma de decisiones clínicas	Puntajes de riesgo en unidades de cuidado intensivo (UCI)	X	

Tabla 4. Aplicaciones de la IA en el sector salud sujetas a regulación o autorregulación.

Cuando el establecimiento de una regulación externa sea apropiado, se puede proponer una regulación *ex ante* para evitar que se concreten riesgos en cuanto a derechos fundamentales (protección de los datos personales de los pacientes, privacidad y derecho a la no discriminación) y en cuanto a la seguridad (derecho a que no se cause daño). Para esta protección *ex ante*, es preciso operacionalizar los controles técnicos, éticos y regulatorios revisados a lo largo de este documento dentro de un marco que tenga como principio base el control humano que, con su carácter instrumental, funge como una herramienta para preservar el principio de la responsabilidad en el desarrollo de los sistemas de IA, ya que mientras exista la presencia humana en el diseño, desarrollo e implementación de los sistemas de IA, se facilita la identificación de los actores responsables de las decisiones derivadas de estos sistemas.

En este punto, cabe hacer hincapié en la importancia que supone el régimen de responsabilidad tanto civil como del Estado, a la hora de definir el marco de regulación de la IA, de manera particular cuando la misma es usada en el sector salud. En este sentido, a continuación, se exponen algunas consideraciones relevantes para abordar la responsabilidad por daños ocasionados por la IA.

Régimen de responsabilidad civil por el uso de la IA

Lo primero que se debe advertir en este punto, es que detrás del régimen de responsabilidad subyace la pregunta sobre quién será el responsable de las decisiones que ya no son tomadas por humanos, sino por sistemas de IA, así como cuestionamientos sobre quién asumirá las consecuencias o costos de los daños asociados al uso de la IA y los impactos de la tecnología. En este contexto, la responsabilidad también supone un contenido legal que se refiere a la necesidad de garantizar la vinculación y resarcimiento de las personas o entidades que son responsables de los daños ocasionados por la implementación de la IA (Fjeld et al., 2020).

Actualmente, la reparación ha suscitado problemáticas profundas en la responsabilidad civil, pues la alta complejidad y opacidad que han alcanzado los sistemas de IA dificulta hacer un seguimiento claro a sus modelos de decisión, realizar una auditoría contundente a los datos utilizados para el entrenamiento del sistema o hacer un seguimiento de las actualizaciones y sus implicaciones para el funcionamiento del sistema. Las anteriores cuestiones, complican el establecimiento de relaciones causales entre las operaciones de los sistemas de IA y los daños causados a las víctimas, lo que dificulta que estas puedan acceder a una indemnización integral. Esto supone igualmente ajustes a las regulaciones existentes y creación de nuevas regulaciones en relación con la responsabilidad civil por daños ocasionados con sistemas de IA y otros sistemas posibles de compensación a las víctimas como fondos de reparación o esquemas de seguros.

El sistema que actualmente se utiliza para la responsabilidad por los daños causados por la IA y las tecnologías asociadas, es un sistema de responsabilidad civil objetiva, en el que se prescinde del juicio de culpabilidad. Ello significa que la víctima de los daños debe probar que el sistema es defectuoso, el daño y el nexo de causalidad entre el defecto y el daño, sin necesidad de probar la culpa. Se considera que el sistema legal que se implemente debe así mismo consagrar un esquema de carga dinámica de la prueba, de tal manera que, en el proceso de indemnización a la víctima, esta tenga una protección especial, en materia sustancial, procesal y probatoria. De igual manera, se debe continuar con el esquema de

responsabilidad solidaria de los actores de la cadena de producción y comercialización de los productos que utilizan sistemas de IA para el resarcimiento de los daños. Sobre estos temas se discutirá en un nuevo documento de investigación.

En este punto, se debe advertir la complejidad que representa la idea de instaurar una regulación sobre el régimen de responsabilidad de la IA. Así, se debe señalar que las reglas sobre la responsabilidad no provienen en su mayoría de fuentes normativas o codificadas, sino más bien jurisprudenciales, en tanto, ha sido la tradición casuística la que ha permitido la construcción de los regímenes de responsabilidad, por lo que se debe decidir si ajustar la jurisprudencia ya existente a los nuevos casos relacionados con la IA o si, más bien, es necesario, la construcción independiente de un nuevo régimen para esta materia.

En igual sentido, surge la pregunta sobre las autoridades responsables de crear y aplicar dicho régimen de responsabilidad. Tal cuestionamiento se enmarca en la naturaleza deslocalizada de la IA, en tanto, el desarrollo, distribución, uso y adaptación de esta se puede presentar en diferentes países, lo que complejiza aún más la tarea. Igualmente, surge en este punto la necesidad de crear un órgano de monitoreo, supervisión y en todo caso, responsable de hacer efectivo tal régimen de regulación, frente a lo que surgen preguntas sobre la naturaleza de dicho órgano, pudiendo ser público, privado o, hasta mixto.

De esta manera, se evidencia la necesidad de establecer controles para la implementación de la IA en el sector salud, la responsabilidad supone la necesidad de crear un organismo de monitoreo para la supervisión técnica de los sistemas de IA. Lo anterior, implica la necesidad de una nueva organización o estructura organizacional (por ejemplo, crear una nueva dirección o división dentro de una organización/entidad existente, o actualizar las funciones de algunos órganos) para crear y supervisar estándares y mejores prácticas a nivel ético y técnico en el contexto de la IA, con el propósito que la misma no infrinja los derechos humanos, las libertades y la dignidad de quienes hacen uso de ella.

Finalmente, se advierte que desde el punto de vista del principio de responsabilidad es necesario que tanto los diseñadores, productores y proveedores de servicios de salud que usan sistemas de IA creen protocolos conjuntos de mitigación de riesgos donde se dé claridad de las responsabilidades del fabricante y de las entidades que van a utilizar o desplegar los dispositivos. Los fabricantes deben ser los responsables de permanecer atentos a la identificación de riesgos y peligros asociados con sus dispositivos médicos y establecer protocolos de mitigación de los riesgos susceptibles a prevención. Las Instituciones prestadoras de salud deben evaluar la seguridad de su red y proteger sus sistemas hospitalarios.

Recomendaciones

Esta sección plantea algunas recomendaciones para desarrollar un marco ético y regulatorio que pueda aplicarse para la implementación de la IA en el sector salud centrado en el ser humano.

Las herramientas que se plantean en este documento para garantizar el respeto de los derechos de los pacientes y todos los usuarios del sistema de salud son las siguientes:

1. **Garantizar desempeño adecuado de los sistemas de IA mediante auditoría de los algoritmos**
   En el contexto de la auditoría de los algoritmos, se recomienda que para todos los sistemas de IA con aplicaciones en el sector salud, se garanticen desde el punto de vista técnico el uso

correcto de métodos estadísticos, probabilísticos y de teoría de información para seleccionar muestras y para evaluar/medir/controlar sesgos algorítmicos de modo que se respete el principio de justicia; la completa y correcta documentación técnica que permita evaluar el funcionamiento de los sistemas, así como sus limitaciones a través de métricas claras de desempeño. En este sentido se recomienda promover una política de calidad que incluya procedimientos efectivos de monitoreo de las soluciones de IA antes y después de que han salido al mercado para certificar de manera continua que los sistemas de IA y los ajustes que se realicen durante la vida útil de la solución cumplen con todas los requisitos técnicos, éticos y normativos revisados en este documento para su operación en el sector salud.

2. **Mitigación de riesgos asociados a sesgos algorítmicos**

   Aunque hay varias fuentes que pueden producir resultados sesgados en sistemas de IA en el sector salud incluyendo sesgos históricos, sesgo de representación, sesgo de medición y sesgo de codificación (el sesgo introducido por las personas que desarrollan los algoritmos), en general la causa más importante de los sesgos está asociada con los datos usados para el entrenamiento de los algoritmos. En este sentido, y entendiendo que las minorías son la población potencialmente más afectada por sesgos algorítmicos, se recomienda incluirlas como parte central de los equipos interdisciplinarios de desarrollo e implementación de sistemas de IA en salud. Adicionalmente, se recomienda considerar el uso de los datos para entender mejor las problemáticas de las minorías y promover la inclusión de estas poblaciones subrepresentadas a través de la apertura responsable de datos relacionados con minorías que puedan complementar las bases de datos con las que se entrenan sistemas de IA aplicados al sector salud.

3. **Validación clínica**

   Previo a la inclusión de sistemas de IA en el flujo de trabajo clínico, es indispensable su validación clínica. Aunque la validación de sistemas de IA se haya realizado con un número importante de pacientes y una amplia evaluación comparativa frente al desempeño de expertos, la gran mayoría de estos estudios se realizan retrospectivamente. Para una adecuada validación del impacto clínico es importante tener en cuenta la realización de estudios prospectivos y prescriptivos para evaluar la utilidad y usabilidad de los sistemas de IA en la práctica médica antes de su implementación. Es importante anotar que la realización de estudios clínicos en seres humanos debe guiarse por los protocolos y los principios enunciados en la actual Declaración de Helsinki y la Guía de Buenas Prácticas Clínicas (GCP, por su sigla en Inglés) emitidas por el Consejo Internacional de armonización de los requisitos técnicos para el registro de medicamentos de uso humano; además deben ser sometidos a un aval por parte de los comités de ética de las instituciones participantes.

4. **Normatividad técnica sobre dispositivos basados en IA para uso en el sector salud**

   Se sugiere la adopción de normas internacionales como la ISO 14971: 2019 sobre riesgos en dispositivos médicos, la cual especifica el proceso que debe seguir el fabricante para identificar los peligros, estimar y evaluar los riesgos relacionados con la seguridad de los datos, los sistemas, la electricidad, las piezas móviles, la radiación y la usabilidad del dispositivo. Adicionalmente, se sugiere la adopción de normas complementarias como la IEC 62304 que

proporciona un marco seguro para los procesos del ciclo de vida en el desarrollo de software desde la etapa de diseño hasta la de mantenimiento seguro del software.

Para el caso de sistemas de IA y su uso en el sector salud, se sugiere la adopción proactiva de normas que se encuentran en proceso de validación como a ISO / IEC CD 23894, que se enfoca en mecanismos para la identificación de causas, métodos de control y medidas correctivas que pueden ser adoptadas en relación con los riesgos inherentes al uso de la IA. La adopción temprana de este tipo de normas se podría implementar usando mecanismos como los sandboxes regulatorios u otro tipo de mecanismos de regulación flexible.

5. **Soporte técnico y actualizaciones de sistemas de IA en el sector salud**

   Para el desarrollo efectivo de los dispositivos de uso médico se deben seguir la reglamentación vigente del INVIMA con algunos ajustes recomendados para el caso de dispositivos médicos basados en IA. Estos ajustes enfatizan la necesidad de considerar, en el proceso de posventa, controles sobre las modificaciones que se pueden presentar por la naturaleza de la tecnología como cambios en el rendimiento, ingreso de nuevos datos y uso previsto del dispositivo, por tratarse de un sistema adaptativo o que puede continuar aprendiendo en el tiempo, de forma que se garantice la seguridad de los pacientes a lo largo de la vida útil del sistema médico.

6. **Capacitación**

   El uso responsable de la IA está condicionado por la capacidad de los desarrolladores y usuarios de entender sus alcances y limitaciones técnicas, así como también sus implicaciones éticas. Por esta razón se recomienda que el gobierno y las instituciones prestadoras de servicios de salud provean mecanismos para la formación de talento humano involucrado en los equipos multidisciplinarios de desarrollo e implementación de la IA mediante programas de entrenamiento técnico y de uso ético y responsable de la IA. Estas habilidades deben ser certificables por una entidad de educación formal y requeridas como parte de los requisitos básicos para avalar la comercialización y aplicación de sistemas basados en IA en el sector salud.

   Esta recomendación está en línea con el planteamiento del Consejo Profesional Nacional de Ingeniería (COPNIA), en el que expresa que, durante el proceso de formación de los futuros ingenieros, las Instituciones de Educación Superior, revisen sus planes académicos e incorporen por un lado la técnica adecuada de cada disciplina, junto a una propuesta de construcción permanente de valores, en beneficio del ejercicio profesional.

   Adicionalmente, dado el carácter disruptivo de la IA en el sector salud, se recomienda que el Gobierno nacional, en conjunto con las facultades de Ciencias de la Salud del país y la Asociación Colombiana de Facultades de Medicina ASCOFAME, diseñen adiciones o adaptaciones al currículo de los programas académicos, de forma que los nuevos profesionales del sector estén formados para participar activamente en los equipos de desarrollo y evaluación de herramientas basadas en IA; colaborar y manejar sistemas de diagnóstico y soporte clínico basados en IA; entender las salidas de los sistemas médicos basados en IA y cómo aplicarlas en el proceso de toma de decisiones con pacientes y familiares; y finalmente, cultivar una

excelente relación médico-paciente cuando se usan sistemas automatizados en la en la nueva era de la medicina aumentada por la IA.

7. **Control humano**
   Se recomienda que la regulación de la IA en el sector salud tenga como principio el *control humano* como instrumento para preservar los demás principios ya señalados en el desarrollo de los sistemas de IA, toda vez que mientras haya presencia humana en el diseño, desarrollo e implementación de los sistemas de IA, se facilita la identificación de los actores responsables de las decisiones derivadas de estos sistemas y se promueve la defensa de los demás principios éticos aquí propuestos.

8. **Reglas sobre el consentimiento informado**
   El consentimiento informado en Colombia está regulado por la ley 23 de 1981, la resolución 4343 de 2012 y la resolución 8430 de 1993. Cuando un profesional utilice una herramienta de soporte para la toma de decisión que tenga incluido un algoritmo de IA, que derive en un diagnóstico o una conducta que tenga impacto positivo o negativo en la salud del paciente, debe informarle a éste acerca de ello, obtener su consentimiento, y diligenciar el respectivo consentimiento informado según la normatividad vigente para ello. Debe describirse en el consentimiento informado algunas métricas claras sobre el desempeño del sistema de IA y cómo el sistema está usando dichas métricas en su proceso de toma de decisiones. En general, el consentimiento informado debe estar escrito en un lenguaje claro y entendible para el paciente, también debe describirse si el uso de los datos relevantes (información clínica), quedará almacenada o será eliminado una vez el algoritmo de IA procese la información suministrada. También debe existir un espacio donde el paciente manifieste su disentimiento sobre el uso de tecnología con IA.

9. **Autorregulación**
   En el caso de los sistemas de IA para aplicaciones en las que se sugiere la autorregulación, se recomienda que los diseñadores, productores y proveedores de servicios de salud que usan sistemas de IA creen protocolos conjuntos de mitigación de riesgos donde se dé claridad de las responsabilidades del fabricante y de las entidades que van a utilizar o desplegar los dispositivos y además que monitoreen de forma continua la calidad de los sistemas y sus indicadores de desempeño. Es posible que sean necesarios mayores esfuerzos de supervisión por parte de organizaciones médicas profesionales para garantizar la calidad de los sistemas que están por fuera de la normativa del ente regulador, como aquellos desarrollados al interior de las organizaciones que no son comercializados.

Conclusiones

La IA aplicada a la salud representa un campo amplio de trabajo con un significativo potencial para mejorar los resultados clínicos, la salud de las poblaciones y optimizar el funcionamiento del sistema de salud en Colombia. Sin embargo, existe poco sustrato normativo sobre la forma de incorporar la IA en el sector salud de forma que se minimicen los riesgos asociados a su uso, mientras se maximizan sus potenciales beneficios para la sociedad. En este contexto, los avances en el marco ético para la IA en

Colombia y otros referentes mundiales proporcionan lineamientos generales para el uso ético y responsables de la IA, los cuales retomamos en el presente documento como punto de partida para establecer recomendaciones ajustadas a las particularidades de las aplicaciones de la IA en el sector salud.

El presente documento presenta un análisis reflexivo y detallado de los riesgos de la IA aplicada en el sector salud y plantea la necesidad de crear un marco ético y regulatorio para su integración en este sector. Este trabajo presenta recomendaciones desde el punto de vista ético que se deben considerar al momento de diseñar herramientas de política pública para la implementación de la IA en el sector salud.

En primer lugar, este trabajo propone controles éticos y técnicos para garantizar la operacionalización de los principios aplicables de la IA contenidos en el marco para la IA en Colombia y de los cuatro principios bioéticos básicos de no maleficencia, beneficencia, autonomía y justicia, así como también la creación de un marco ético que tenga como principio base el control humano como herramienta para preservar el principio de la responsabilidad en el desarrollo de los sistemas de IA.

En segundo lugar, este trabajo plantea recomendaciones que están enfocadas en establecer estándares y mecanismos que permitan comprender, evaluar, validar y supervisar el desarrollo e implementación ética y responsable de la IA aplicada al sector salud dentro de un marco que considere medidas de autorregulación y regulación externa acordes con el riesgo que representa el uso de los sistemas basados en IA para el bienestar de los pacientes y de la sociedad en general. El presente documento enfatiza la necesidad del uso responsable de los datos de los pacientes y usuarios del sistema de salud; la responsabilidad de ajustar la estructura del consentimiento informado para comunicar adecuadamente a los pacientes sobre el uso de herramientas basadas en IA dentro de su proceso de atención; la importancia de prevenir y detectar sesgos algorítmicos que puedan resultar en la discriminación de grupos minoritarios; y la necesidad de ajustes en el sistema educativo para formar a quienes están involucrados en el desarrollo y uso de sistemas de IA en nuevas habilidades que les permitan entender los sistemas de IA no sólo desde el punto de vista técnico, sino también sus implicaciones éticas en la sociedad.

Finalmente, el presente documento propone consideraciones para establecer un marco regulatorio para la IA en el sector salud que abarca elementos para abordar la responsabilidad por daños ocasionados por la IA. Potenciales alternativas implican la creación de órganos de control multidisciplinarios con facultades legales, técnicas y éticas; que cumplan finalidades de monitoreo y vigilancia de forma que se implemente un marco regulatorio balanceado en el que la IA se use como una herramienta para mejorar la experiencia y los resultados clínicos de los pacientes, mejorar la salud de las poblaciones y disminuir el costo de salud per cápita en beneficio de las comunidades respetando los derechos fundamentales de los ciudadanos.

Referencias

Ayala-García, J. (2014). *La salud en Colombia: Más cobertura pero menos acceso*. (1692-3715). Retrieved from https://www.banrep.gov.co/sites/default/files/publicaciones/archivos/dtser_204.pdf.

Beauchamp, T., & Childress, J. (1979). *Principles of Biomedical Ethics* (2 ed.): Oxford University Press.

Buch, V. H., Ahmed, I., & Maruthappu, M. (2018). Artificial intelligence in medicine: current trends and future possibilities. *Br J Gen Pract, 68*(668), 143-144. doi:10.3399/bjgp18X695213

Celi, L.A., Majumder, M.S., Ordóñez, P., Osorio, J.S., Paik, K.E. and Somai, M., (2020). Leveraging Data Science for Global Health. Springer Nature. doi: 10.1007/978-3-030-47994-7

Chen, I., Johansson, F., & Sontag, D. (2018). *Why is my classifier discriminatory?* Paper presented at the International Conference on Neural Information Processing Systems, Montreal.

Chen, I., Szolovits, P., & Ghassemi, M. (2019). Can AI Help Reduce Disparities in General Medical and Mental Health Care? *AMA J Ethics, 21*(2), E167-179. doi:10.1001/amajethics.2019.167

Esteva, A., Kuprel, B., Novoa, R. A., Ko, J., Swetter, S. M., Blau, H. M., & Thrun, S. (2017a). Corrigendum: Dermatologist-level classification of skin cancer with deep neural networks. *Nature, 546*(7660), 686. doi:10.1038/nature22985

Esteva, A., Kuprel, B., Novoa, R. A., Ko, J., Swetter, S. M., Blau, H. M., & Thrun, S. (2017b). Dermatologist-level classification of skin cancer with deep neural networks. *Nature, 542*(7639), 115-118. doi:10.1038/nature21056

Finlayson, S., Bowers, J., Ito, J., Zittrain, J., Beam, A., & Kohane, I. S. (2019). Adversarial attacks on medical machine learning. *Science, 363*, 1287-1289.

Fjeld, J., Nele, A., Hilligoss, H., Nagy, A., & Srikumar, M. (2020). Principled Artificial Intelligence: Mapping Consensus in Ethical and Rights-based Approaches to Principles for AI. Retrieved from https://dash.harvard.edu/handle/1/42160420

Floridi, L., & Cowls, J. (2019). A Unified Framework of Five Principles for AI in Society. *Harvard Data Science Review, 1*(1). doi:10.1162/99608f92.8cd550d1

Guío, A. (2020). *Marco ético para la inteligencia artificial en Colombia: Documento para discusión*. Retrieved from https://dapre.presidencia.gov.co/AtencionCiudadana/DocumentosConsulta/consulta-marco-etico-IA-Colombia-200813.pdf.

Haenssle, H. A., Fink, C., Schneiderbauer, R., Toberer, F., Buhl, T., Blum, A., . . . Zalaudek, I. (2018). Man against machine: diagnostic performance of a deep learning convolutional neural network for dermoscopic melanoma recognition in comparison to 58 dermatologists. *Ann Oncol, 29*(8), 1836-1842. doi:10.1093/annonc/mdy166

Helm, J. M., Swiergosz, A. M., Haeberle, H. S., Karnuta, J. M., Schaffer, J. L., Krebs, V. E., . . . Ramkumar, P. N. (2020). Machine Learning and Artificial Intelligence: Definitions, Applications, and Future Directions. *Curr Rev Musculoskelet Med, 13*(1), 69-76. doi:10.1007/s12178-020-09600-8

International Medical Device Regulators Forum. (2014). "Software as a Medical Device": Possible Framework for Risk Categorization and Corresponding Considerations. Retrieved from http://www.imdrf.org/docs/imdrf/final/technical/imdrf-tech-140918-samd-framework-risk-categorization-141013.pdf

International Telecommunication Union. (2014). *Mejores Prácticas de liderazgo, innovación y gestión pública en e-salud: Los casos de Brasil, México y Perú*. Retrieved from https://www.itu.int/en/ITU-D/Regional-Presence/Americas/Documents/PBLCTNS/20140331-ehealth-SP.pdf.

Keane, P. A., & Topol, E. J. (2018). With an eye to AI and autonomous diagnosis. *NPJ Digit Med, 1*, 40. doi:10.1038/s41746-018-0048-y

Kelly, C. J., Karthikesalingam, A., Suleyman, M., Corrado, G., & King, D. (2019). Key challenges for delivering clinical impact with artificial intelligence. *BMC Med, 17*(1), 195. doi:10.1186/s12916-019-1426-2

Lundberg, S., & Lee, S. (2017). A Unified Approach to Interpreting Model Predictions. Retrieved from https://arxiv.org/pdf/1705.07874.pdf

MacPherson, Y. and Pham, K., (2020). Ethics in Health Data Science. In Leveraging Data Science for Global Health (pp. 365-372). Springer Nature.

McLennan, S., Fiske, A., Celi, L.A., Müller, R., Harder, J., Ritt, K., Haddadin, S. and Buyx, A., (2020). An embedded ethics approach for AI development. Nature Machine Intelligence, pp.1-3.

Mosquera-Lopez, C., Dodier, R., Tyler, N. S., Wilson, L. M., El Youssef, J., Castle, J. R., & Jacobs, P. G. (2020). Predicting and Preventing Nocturnal Hypoglycemia in Type 1 Diabetes Using Big Data Analytics and Decision Theoretic Analysis. *Diabetes Technology & Therapeutics*. doi:10.1089/dia.2019.0458

Organisation for Economic Co-operation and Development. (2020). Trustworthy AI in health. Retrieved from http://www.oecd.org/health/trustworthy-artificial-intelligence-in-health.pdf

Panch, T., Szolovits, P., & Atun, R. (2018). Artificial intelligence, machine learning and health systems. *J Glob Health, 8*(2), 020303. doi:10.7189/jogh.08.020303

Price, N. (2019). Risks and remedies for artificial intelligence in healthcare. *BROOKINGS*. Retrieved from https://www.brookings.edu/research/risks-and-remedies-for-artificial-intelligence-in-health-care/#:~:text=While%20AI%20offers%20a%20number,health%2Dcare%20problems%20may%20result

Ribeiro, M., Singh, S., & Guestrin, C. (2016). "Why Should I Trust You?" Explaining the Predictions of Any Classifier. Retrieved from https://arxiv.org/pdf/1602.04938.pdf

Robbins, R., & Brodwin, E. (2020). An invisible hand: Patients aren't being told about the AI systems advising their care. *STAT*. Retrieved from https://www.statnews.com/2020/07/15/artificial-intelligence-patient-consent-hospitals/

Schiff, D., & Borenstein, J. (2019). How should clinicians communicate with patients about the roles of artificially intelligent team members? *AMA Journal of Ethics, 21*(2), E138-145.

U.S. Food and Drug Administration. (2017). *Digital health innovation action plan*. Retrieved from https://www.fda.gov/media/106331/download.

U.S. Food and Drug Administration. (2019). *Proposed Regulatory Framework for Modifications to Artificial Intelligence/Machine Learning (AI/ML)-Based Software as a Medical Device (SaMD): Discussion Paper and Request for Feedback*. Retrieved from https://www.fda.gov/files/medical%20devices/published/US-FDA-Artificial-Intelligence-and-Machine-Learning-Discussion-Paper.pdf.

Vickers, A. J., Cronin, A. M., Elkin, E. B., & Gonen, M. (2008). Extensions to decision curve analysis, a novel method for evaluating diagnostic tests, prediction models and molecular markers. *BMC Med Inform Decis Mak, 8*, 53. doi:10.1186/1472-6947-8-53

Wachter, S., Mittelstadt, B., & Russell, C. (2017). Counterfactual Explanations without Opening the Black Box: Automated Decisions and the GDPR. Retrieved from https://arxiv.org/ftp/arxiv/papers/1711/1711.00399.pdf