CLIAS

CENTRO DE INTELIGENCIA ARTIFICIAL Y SALUD PARA AMÉRICA LATINA Y EL CARIBE

Inteligencia Artificial (IA) Responsable: Claves para aplicar los principios éticos en las soluciones de IA en el campo de la salud.

DOCUMENTO TECNICO 3

Septiembre 2023

CONTENIDO

Equipo de Trabajo

Santiago Esteban: Médico de familia (Universidad Austral). Master en Salud Pública con orientación en datos en la Escuela de Salud Pública T.H. Chan de Harvard. Master en Administración de Negocios de la Universidad de San Andrés. Investigador staff en IA en la intersección de epidemiología, salud pública, aprendizaje automático, inferencia causal, ciencia de datos y sistemas de información en el CIIPS-IECS.

Rosa Angelina Pace: Medica Cirujana (UNNE) y Magister en Bioética por la Universidad Complutense de Madrid (UCM). Coordinadora del Centro de Bioética del Hospital Italiano de Buenos Aires (HIBA), y Directora del Departamento de Ciencias Humanas y Sociales del Instituto Universitario Hospital Italiano (IUHIBA). Miembro del Consejo de Ética en Medicina Academia Nacional de Medicina. Recibió premios en Bioética, Velasco Suarez OPS. Consultora CIIPS-IECS

Velén Pennini: Licenciada en Antropología (UNLP), especialista en Epidemiología de Campo por el programa de formación en servicio del Ministerio de Salud de la Nación, y especialista en Estadística aplicada a la Salud (FCEN-UBA). Investigadora del CIIPS-IECS.

Adrián Santoro: Licenciado en Sociología (UBA) y Magister en Generación y Análisis de Información Estadística (UNTREF). Se desempeña en el campo de la investigación en epidemiología, demografía y estadísticas de salud. Experto en programación y desarrollo de modelos matemáticos del CIIPS-IECS.

Adolfo Rubinstein: Médico de Familia (UBA). Máster en Epidemiología Clínica por la Harvard TH Chan School of Public Health, Diplomado en Economía de la Salud por la Universidad de York. Doctor en Salud Pública (UBA). Profesor Regular Titular de Salud Pública (UBA). Certificado de implementación de políticas públicas por la Harvard Kennedy School. Ministro de Salud de la Nación de Argentina (2017-2019). Director del Centro de Implementación e Innovación en Políticas de Salud (CIIPS-IECS).

Cintia Cejas: Lic. en Ciencias Políticas (UCA) y Magister en Ciencias Sociales y de la Salud (FLACSO-CEDES). Especialista en gestión de proyectos de salud. Coordinadora del Centro de Implementación e Innovación en Políticas de Salud (CIIPS-IECS) y del Centro de Inteligencia Artificial en Salud para Latinoamérica y el Caribe (CLIAS).

1. Presentación

El presente documento, elaborado por el Centro de Implementación e Innovación en Políticas de Salud (CIIPS) del Instituto de Efectividad Clínica y Sanitaria (IECS), se enmarca en una Serie de Documentos Técnicos sobre Inteligencia Artificial y Salud (https://clias.iecs.org.ar/publicaciones/).

Estos documentos tienen por objetivo aportar al conocimiento de la región, abordando distintos ejes y perspectivas relevantes en el análisis de esta temática.

Destinados a equipos de salud, formuladores de programas y políticas de salud y decisores en todos los niveles, y público en general, con especial interés en la transformación digital del sector salud y su vinculación a la salud sexual, reproductiva y materna (SSRM), esta serie de documentos sobre IA que estamos elaborando se complementan con las actividades llevadas a cabo por el CLIAS (Centro de Inteligencia Artificial en Salud para Latinoamérica y el Caribe) que se desarrolla en el CIIPS, con el apoyo del International Development Research Centre (IDRC). Para mayor información sobre el CLIAS, visitar http://clias.iecs.org.ar

Este documento en particular aborda el uso de la inteligencia artificial (IA) en salud, desde la perspectiva de la responsabilidad. La inteligencia artificial responsable se refiere a la práctica de desarrollar, implementar y utilizar sistemas de inteligencia artificial (IA) de manera ética para minimizar los riesgos y consecuencias negativas asociadas con su aplicación. Esto implica considerar una serie de principios y prácticas para garantizar que la IA beneficie a la sociedad en su conjunto y no cause daño.

2. Mensajes claves del documento

- En el contexto de la salud y la atención médica, la IA se ha convertido en una herramienta prometedora con el potencial de mejorar el diagnóstico, el tratamiento y la gestión de enfermedades, así como el análisis de datos médicos y de salud a gran escala.
- Sin embargo, es fundamental considerar los riesgos asociados, que principalmente se relacionan con el manejo y protección de los datos, así como los sesgos que pueden producirse o agravarse, colocando en una posición desfavorable a grupos vulnerables y acentuando las disparidades ya existentes, como las de género o la exclusión de las minorías, entre otras.
- Los daños derivados de la aplicación de la IA pueden ser tanto de índole material, incluyendo daños a la seguridad (filtración de datos personales) y salud de las personas (errores diagnósticos), como inmateriales, tales como la pérdida de privacidad, limitaciones a la libertad de expresión, dignidad y discriminación en el acceso a oportunidades laborales, entre otros aspectos.
- En consecuencia, se plantea la necesidad de tomar con extrema seriedad la incorporación de perspectivas complementarias en la producción y evaluación de las aplicaciones de IA en el ámbito de la atención de salud que incluya enfoques desde áreas que no se limiten únicamente al campo técnico de desarrollo. Estos enfoques adicionales, especialmente provenientes de las ciencias humanas y sociales, y en particular de expertos en ética, deben participar desde las fases iniciales de los proyectos para intentar mitigar los sesgos en los algoritmos y programas concebidos únicamente por tecnólogos, pero que pueden soslayar algunos aspectos que contribuyen a la ampliación de desigualdades y a la negligencia de otros valores humanos (ética integrada).
- Los principios de “seguridad”, “autodeterminación”, “benevolencia” y “universalismo”, no solo intentan garantizar el diseño responsable de tecnologías de IA, sino que también abren un camino hacia soluciones más equitativas, inclusivas y beneficiosas para la sociedad en su conjunto.
- Los sesgos son errores o desvíos sistemáticos, inclinación en las decisiones o predicciones de un modelo de IA que pueden llevar a resultados injustos o inequitativos. Uno de ellos es cuando los datos utilizados para el entrenamiento del modelo no representan adecuadamente la diversidad o variabilidad de la población objetivo. También, puede ocurrir cuando una base de datos tiene problemas en relación a su estructura como, por ejemplo, codificando el género de forma binaria borrando otras identidades de género en categorías agrupadas.
- Por ello, se debe trabajar en lograr que el set de datos sea la mejor representación posible de la población objetivo. En este sentido, es necesario remarcar la importancia de contar con desarrollos atravesados, desde sus orígenes, por la pluralidad, el contexto y la intersectorialidad.

- Abordar estos aspectos éticos de manera efectiva requiere la colaboración entre profesionales de la salud, científicos de datos, ingenieros informáticos, legisladores, decisores y expertos en ética.
- Por eso, es imperativo establecer marcos normativos sólidos que guíen el desarrollo y la implementación de la IA, asegurando que los beneficios se maximicen a la vez que se minimizan los posibles daños.
- La búsqueda de soluciones éticas en el ámbito de la IA no solo garantiza la integridad y confiabilidad de las tecnologías emergentes, sino que también promueve un enfoque responsable y sostenible hacia la innovación tecnológica, en beneficio de la sociedad en su conjunto.

3. Introducción

La relación entre la inteligencia artificial (IA) en el ámbito de la salud y la ética es un tema de creciente interés y debate. La IA se define como el campo de estudio y desarrollo de sistemas y tecnologías capaces de simular la inteligencia humana para llevar a cabo tareas complejas de manera autónoma[1]. **En el contexto de la salud, la IA se ha convertido en una herramienta prometedora con el potencial de mejorar el diagnóstico, el tratamiento y la gestión de enfermedades, así como el análisis de datos médicos y de salud a gran escala**. Sin embargo, la aplicación de la IA en la salud plantea una serie de desafíos éticos que deben abordarse de manera cuidadosa y reflexiva. **Los principales temas a considerar son los riesgos asociados, que principalmente se relacionan con el manejo y protección de los datos, así como los sesgos que podrían producirse o agravarse, colocando en una posición desfavorable a minorías de distintos orígenes y acentuando las disparidades ya existentes, como las de género y otras**.

A lo largo de este documento, se define el concepto de ética y sus principios asociados; cuál es el rol de la ética en las soluciones de IA; qué son los sesgos y por qué son tan importantes en el desarrollo de modelos de IA, especialmente en el campo de la salud y finalmente el documento aborda, con ejemplos, la aplicación de los principios éticos a todo el ciclo de vida de las soluciones basadas en IA: selección y definición del problema a resolver, planificación y diseño, desarrollo y validación, despliegue e implementación y operación y monitoreo.

4. ¿Qué es la ética?

La ética es una disciplina o forma de conocimiento que proporciona orientación a la acción, lo que la convierte en una forma de conocimiento práctico. **Se trata de un conocimiento destinado a guiar una conducta racional**. Dentro de esta categoría de conocimientos prácticos, que se centran en dirigir la acción para lograr un resultado tangible, como ocurre en el ámbito de la técnica o el arte, la ética persigue un objetivo más amplio al intentar reflexionar y orientar los esfuerzos hacia una actuación correcta de manera racional[2], sin soslayar las circunstancias que siempre nos condicionan, apelando a elegir prudentemente, racionalmente."

Diferentes autores[3] argumentan que la moral está estrechamente ligada a la naturaleza humana y es esencial para ella. También sugieren que la inteligencia y, por ende, la moralidad, tienen raíces en la biología, y su propósito es asegurar la supervivencia de nuestra especie. La inteligencia funciona como una forma de adaptación, donde las decisiones humanas no están determinadas únicamente por la selección natural, sino que incluyen elecciones conscientes y responsables. En la especie humana, debemos justificar las decisiones que tomamos.

Esta estructura moral nos impide ser amorales. Es decir, todos debemos darle contenido a nuestra moralidad. Podemos actuar de manera inmoral, pero no podemos carecer por completo de moral. La calidad de nuestra vida, acciones y proyectos dependerá de cómo llenemos de contenido esa estructura moral.

Así es que cualquier actividad humana debería ser mirada a la luz de valores positivos, tendiendo a mejorar la calidad de vida y minimizar posibles daños que se provoquen por su actividad. La inteligencia artificial, en cualquiera de sus aplicaciones, no escapa a este análisis.

Pensar la inteligencia artificial en clave ética es ser extremadamente cuidadosos en que sea beneficiosa para los seres humanos y el ambiente, que sea capaz de potenciarnos en todo sentido, especialmente en perfeccionar las estrategias para una vida saludable, en evitar cualquier acción dañina tal como daño directo o por errores, discriminación o injusticia en sus resultados.

Para ello y tal como ya está demostrado, las estrategias para resultados aceptables desde el punto de vista ético deben plantearse desde el momento mismo de la concepción de los proyectos e incluyendo miradas complementarias a las específicamente tecnológicas para intentar prevenir y mitigar funcionamientos no deseados.

5. ¿Por qué hablar de ética en IA?

De acuerdo con Klaus Schwab, el mundo ha dado inicio a su cuarta revolución industrial, y los cambios han adquirido una velocidad inimaginable[4]. Schwab plantea que, si bien la revolución digital tuvo sus comienzos a mediados del siglo XX, diluyendo las fronteras entre lo físico, lo biológico y lo computacional mediante una fusión de tecnologías, la aceleración notable que ha dado lugar a esta cuarta revolución ha visto surgir la inteligencia artificial, la robótica, el blockchain, la nanotecnología, la biotecnología, entre otras, dando lugar a sistemas ciberfísicos. Estos sistemas ciberfísicos se caracterizan por ser una representación virtual del universo físico, operando de forma digital y descentralizada e interactuando mediante el “Internet de las cosas”, una red de dispositivos interconectados que pueden recopilar y compartir datos a través de internet, permitiendo la comunicación y la automatización entre objetos físicos y sistemas digitales.

Las características distintivas que definen la presente revolución radican en su acelerado ritmo de avance, su alcance abarcador y su impacto tangible en el ámbito físico.

En este contexto, resulta imperativo reconocer la relevancia de la responsabilidad ética, la cual debe ser observada y evaluada de manera crítica durante el desarrollo de estos procesos. Tal responsabilidad ética implica una genuina reflexión y un diálogo profundo que conducen a un entendimiento cabal de las obligaciones que incumben a la humanidad en el marco de esta revolución tecnológica, la cual difícilmente puede o debe ser detenida.

Estas consideraciones introspectivas deben traducirse en acciones concretas y efectivas, en las cuales **las decisiones concernientes a la creación de herramientas de inteligencia artificial (IA) se**

forjen desde las etapas iniciales del diseño hasta su completa implementación y posterior seguimiento. La cautela y previsión en el proceso de construcción de estas herramientas se tornan fundamentales para garantizar un enfoque responsable y ético que salvaguarde los valores inherentes a la dignidad humana y el bienestar colectivo. En esta línea vienen trabajando diferentes grupos, incluyendo organizaciones internacionales como la OMS[5] o UNESCO[6].

La IA tiene el potencial de transformar significativamente la sociedad siendo un medio prometedor para favorecer la prosperidad humana y, de ese modo, mejorar el bienestar individual y social y el bien común, además de traer consigo progreso e innovación[7]. No obstante, es crucial reconocer que la implementación de la IA también conlleva ciertos riesgos y desafíos que deben ser abordados de manera adecuada y proporcional. **Entre estos riesgos, se destacan la opacidad en el funcionamiento de los sistemas, el aumento de las brechas de género, la exclusión de las minorías, la intromisión en la esfera privada de los individuos, la especulación financiera y el uso indebido para actividades delictivas o guerras.**

Los daños derivados de la aplicación de la IA pueden ser tanto de índole material, incluyendo daños a la seguridad (filtración de datos personales) y salud de las personas (errores diagnósticos), como inmateriales, como la pérdida de privacidad, limitaciones a la libertad de expresión, dignidad y discriminación en el acceso a oportunidades laborales, entre otros aspectos [6,7].

Se ha alcanzado cierto consenso en cuanto a los principios que deben regir el desarrollo e implementación de sistemas basados en IA. Sin embargo, existen voces que expresan la insuficiencia de estos principios para guiar adecuadamente las acciones, argumentando que son demasiado genéricos frente a los daños reales y potenciales [8]. En consecuencia, se plantea la **necesidad de tomar con extrema seriedad la incorporación de perspectivas complementarias en la producción y evaluación de las aplicaciones de IA en el ámbito de la atención de salud** que incluya enfoques desde áreas que no se limiten únicamente al campo técnico de desarrollo. **Estos enfoques adicionales, especialmente provenientes de las ciencias humanas, y en particular de expertos en ética, deben participar desde las fases iniciales de los proyectos para intentar mitigar los sesgos en los algoritmos y programas concebidos únicamente por tecnólogos, que pueden soslayar algunos aspectos que contribuyen a la ampliación de desigualdades y a la negligencia de otros valores humanos.**

En esta dirección, es clave hablar de la **ética integrada** o “embedded ethics”[9] que busca dar respuesta a la necesidad de que los grandes lineamientos éticos para la IA sean atendidos y respetados, de modo que puedan anticipar, identificar y mitigar estas problemáticas durante el desarrollo de las soluciones basadas en IA. Así, McLennan y cols. [9], proponen un **modelo de desarrollo que integre a la ética desde el inicio de los proyectos**, sobre todo para aquellos en el ámbito de la salud. Este modelo promueve principalmente el **trabajo integrado entre equipos de desarrollo informático, equipos con conocimiento temático y equipos de eticistas desde el inicio del proyecto, velando por la transparencia en la medida en que no comprometa la confidencialidad y la propiedad intelectual**. Esto implica establecer en forma coordinada los

objetivos, el impacto buscado y los métodos utilizados con intercambios regulares y marcos teóricos claros y explícitos.

Solanki y cols. [8] proponen una serie de valores humanos que se mapean con principios éticos que sirven de guía para equipos que desarrollen herramientas basadas en IA. Estos lineamientos propuestos enfatizan la importancia de la ética en el desarrollo de herramientas de IA, destacando aspectos cruciales como la **seguridad**, la **autodeterminación**, la **benevolencia** y el **universalismo**. Estos principios no solo garantizan el diseño responsable de tecnologías de IA, sino que también abren un camino hacia soluciones más equitativas, inclusivas y beneficiosas para la sociedad en su conjunto. A continuación, se resumen y comentan estos conceptos:

SEGURIDAD	Sistemas de IA seguros, o sea, que no generen daños, peligros, riesgos o amenazas como consecuencia de su uso. Esto abarca tanto la esfera psicológica como la salvaguardia contra perjuicios físicos y sociales, cuestiones como la preservación de la privacidad, la integridad y la seguridad. Se concede primacía al principio de **no maleficencia**, el cual implica la firme obligación de prevenir daños, prevaleciendo sobre la intención de promover el bien.
AUTODETERMINACIÓN	Reviste una significación primordial la consideración hacia la dignidad intrínseca de los individuos, la preservación de su autonomía, la salvaguarda de sus libertades fundamentales y el respeto irrestricto de su **derecho al consentimiento informado**, el cual engloba tanto la aquiescencia inicial como la posibilidad de retractarse, en lo que respecta a la participación o sujeción a cualquier sistema de IA en el ámbito de la atención médica. El consentimiento informado, entraña la comprensión completa de las implicancias inherentes a la propuesta en cuestión. Si bien en el caso de ciertas herramientas de IA, caracterizadas por su opacidad asemejándose a “cajas negras” en donde no se puede ver el interior, la plena comprensión puede verse obstaculizada, es indudable que cada vez más individuos conciben la inherente incertidumbre que subyace en el ámbito de la atención médica. Concomitantemente, el derecho a la privacidad, reconocido como prerrogativa fundamental del ser humano, impone responsabilidades a las personas y, de manera especial, al personal médico, entre las cuales destaca la **salvaguarda de la confidencialidad** [10].

BENEVOLENCIA	Importancia de sopesar, en primera instancia, el imperativo de la beneficencia, que se refiere a la obligación de los profesionales de la salud de actuar en **beneficio de los pacientes y de buscar su bienestar**, por lo tanto, cualquier herramienta basada en IA que se utilice debe estar aplicada con esto como uno de sus objetivos centrales. A su vez, resultan fundamentales la **transparencia** y la "**explicabilidad"**, que hace referencia a hacer inteligibles los resultados de los sistemas de IA y como se obtuvieron, para lograr confianza de los usuarios en estas herramientas. Estos principios se refieren a la habilidad de entender cómo funcionan las herramientas, interpretar lo que producen, comprender sus tendencias y desviaciones, y entender cómo interactuar con estas. De todos modos, la discusión sobre las implicancias de la explicabilidad y su necesidad continúan vigentes[11]. Finalmente, los autores enfatizan la importancia de incluir ideas de **solidaridad**. En concreto, sugieren que es fundamental tener siempre en cuenta, durante la creación de estas herramientas, a los grupos de personas que son más vulnerables a los efectos de sesgos sociales embebidos en los datos que eventualmente resulten en una performance marcadamente inferior. Esto es debido a que existen posibles consecuencias graves que podrían perjudicar a estos grupos, las cuales difieren significativamente de los efectos previstos para la población general.
UNIVERSALISMO	La pregunta que conlleva este concepto es: **¿Es plausible que dicha solución pueda conferir una ventaja a todos los individuos de manera equitativa?** (principio de justicia). En ese contexto, surge la necesidad de dilucidar qué medidas y dispositivos deben ser elaborados para contrarrestar las desigualdades manifestadas en esferas tales como la brecha digital. Asimismo, se plantean cuestionamientos acerca de cómo facilitar el acceso a colectivos vulnerables y cómo atenuar el posible impacto que esta solución podría ejercer sobre grupos subrepresentados en el ámbito digital (punto focal en la equidad). Todas estas consideraciones alcanzan una particular relevancia cuando se aborda la cuestión de género, ya que los **sesgos inherentes en los ámbitos tecnológicos, que frecuentemente tienen una composición predominantemente caucásica y masculina, terminan por permear en los algoritmos desarrollados**. Es difícil anticipar que estos algoritmos posean, desde su génesis, una perspectiva acorde con las perspectivas modernas e inclusivas, lo cual aboga por la conformación de equipos

de desarrollo heterogéneos que logren reflejar la diversidad de enfoques necesaria.

6. ¿Qué son los sesgos en modelos de IA?

Los sistemas de IA basados en aprendizaje automático buscan aumentar el rendimiento de las predicciones optimizando una función de pérdida[1], o, dicho de otro modo, **minimizando el error de esas predicciones**. Sin embargo, el error se presenta en diferentes tipos. Existe un error aleatorio que es intrínseco a los sistemas de IA y no es posible eliminarlo por completo, sino que buscamos minimizarlo. Este suele provenir de diversas fuentes como puede ser el tamaño muestral y la variabilidad de los datos y variaciones en los procesos de entrenamiento. Por otro lado, **existe un error no aleatorio, conocido como sesgo**.

Los sesgos son errores sistemáticos o inclinación en las decisiones o predicciones de un modelo de IA que pueden llevar a resultados injustos o inequitativos, o simplemente erróneos.

Al trabajar en el desarrollo de herramientas de IA, los sesgos pueden producirse de diferentes modos o en diferentes etapas.

Uno de ellos se presenta **cuando los datos utilizados para el entrenamiento no representan adecuadamente la diversidad o variabilidad de la población objetivo**. Cuando esto sucede, el modelo puede tener dificultades para generalizar, es decir funcionar con una performance similar ante datos nuevos o nunca vistos. Un conjunto de datos puede estar sesgado desde su diseño, subrepresentado o no representando adecuadamente a una población particular lo que puede conducirnos a conductas discriminatorias. Incluso estando bien representada, **una base de datos puede tener problemas en relación a su estructura** como, por ejemplo, codificando el género de forma binaria (femenino-masculino) invisibilizando otras identidades de género en categorías agrupadas.

Un objetivo importante que debemos tener en cuenta al trabajar con estos desarrollos es el de lograr que el conjunto de datos sea la mejor representación posible de la población objetivo.

Esto se puede alcanzar a partir de estudios de validación. Sin embargo, el proceso de entrenamiento de estos datos también puede incorporar sesgos en nuestros resultados. Sobre esto, ha sido señalado en numerosa bibliografía que, a gran escala, el problema de los sesgos en IA proviene de las universidades y empresas que desarrollan estas tecnologías, compuestas mayoritariamente por hombres blancos de alto nivel socioeconómico y con orientaciones muy técnicas. Ante esta situación, se propone avanzar hacia el desarrollo de proyectos de IA colaborativos con disciplinas

[1] Una función de pérdida es una medida matemática que evalúa qué tan bien un modelo basado en aprendizaje automatizado se ajusta a los datos de entrenamiento al cuantificar la discrepancia entre las predicciones del modelo y los valores reales observados.

sociales y que impliquen a comunidades y a organizaciones de la sociedad civil. En este sentido, se remarca nuevamente la importancia de **contar con desarrollos atravesados, desde sus orígenes, por la pluralidad, el contexto y la intersectorialidad**.

Como se explica, la noción de sesgo es compleja y los humanos también tienen sesgos en su propia práctica. Sin embargo, es posible, y por lo tanto éticamente necesario, diseñar sistemas de IA que ayuden a compensar los sesgos cognitivos y así conducir a resultados más justos y equitativos [12]. En situaciones donde se produce un desacuerdo entre el profesional experto y la herramienta de IA diseñada para respaldar la toma de decisiones, puede surgir un escenario interesante. Si la herramienta de IA ha sido desarrollada de manera sólida desde el punto de vista ético y ha sido sometida a pruebas exhaustivas en diferentes escenarios, podría tener la capacidad de identificar posibles sesgos presentes en los propios profesionales. Esta capacidad de la IA para señalar sesgos en los expertos es especialmente relevante, ya que estos sesgos pueden ser difíciles de abordar mediante estrategias tradicionales, como la supervisión de casos o el intercambio de experiencias con otros profesionales ya que los profesionales de un área similar comparten sesgos.

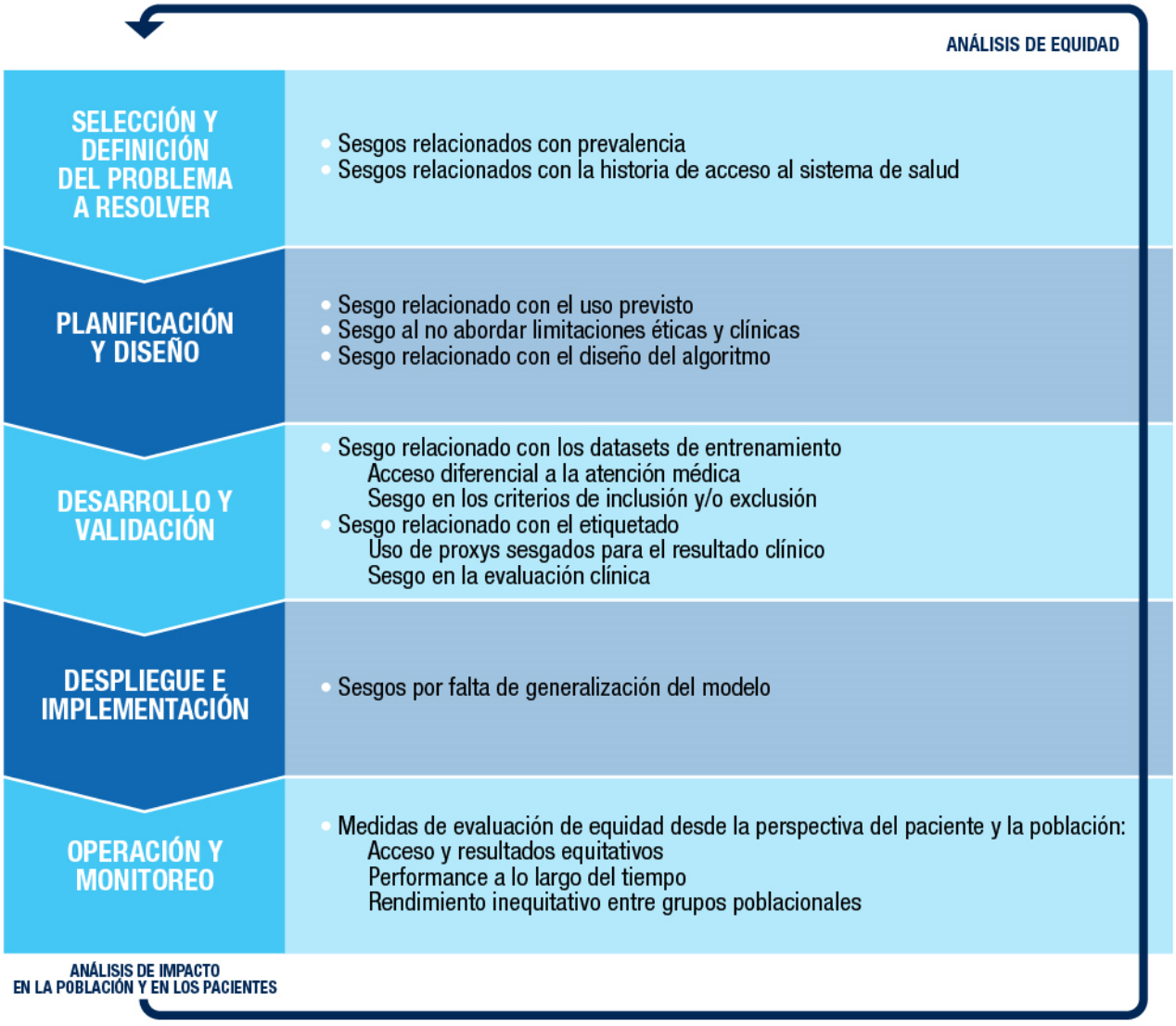
Los ejemplos más relevantes sobre el uso de la IA para abordar el problema del sesgo en la toma de decisiones vienen del campo de los recursos humanos. En la industria tecnológica, donde las mujeres y diversidades se encuentran subrepresentadas, el uso de la IA en decisiones de contratación puede conducir a la toma de decisiones menos sesgadas y aumentar la promoción de mujeres y diversidades en estas posiciones laborales. Al explorar el potencial de la IA en este sentido, se puede ampliar la comprensión de los sesgos subyacentes en las decisiones humanas y trabajar en su mitigación.

7. Aplicación de los principios éticos al ciclo de vida de las soluciones basadas en IA

Según la Organización para la Cooperación y el Desarrollo Económicos (OCDE)[13], el ciclo de vida de los proyectos que incluyen inteligencia artificial podría desarrollarse en las siguientes etapas:

1. Selección y definición del problema a resolver
2. Planificación y diseño
3. Desarrollo y validación
4. Despliegue e implementación
5. Operación y monitoreo

Potenciales fuentes de sesgo durante el ciclo de vida de las soluciones basadas en IA



Adaptado de: Abràmoff MD, Tarver ME, Loyo-Berrios N, Trujillo S, Char D, Obermeyer Z, Eydelman MB; Foundational Principles of Ophthalmic Imaging and Algorithmic Interpretation Working Group of the Collaborative Community for Ophthalmic Imaging Foundation, Washington, D.C.; Maisel WH. Considerations for addressing bias in artificial intelligence for health equity. NPJ Digit Med. 2023 Sep 12;6(1):170. doi: 10.1038/s41746-023-00913-9. PMID: 37700029; PMCID: PMC10497548.

De acuerdo a lo descripto previamente, es fundamental embeber los aspectos éticos de modo que queden integrados en todas las fases del proceso de desarrollo de una herramienta de IA. A continuación, se detallan estas fases y desafíos éticos principales involucrados.

7.1. Selección y definición del problema a resolver

El proceso de toma de decisiones sobre qué desarrollar en IA sanitaria es complejo y suele implicar a diversas partes interesadas. Sin embargo, es importante garantizar que este proceso sea transparente e inclusivo, de modo que se satisfagan las necesidades de todos los actores, teniendo en cuenta:

- *Beneficios y riesgos potenciales*: Los beneficios de la IA en la asistencia sanitaria pueden ser significativos, pero también existen riesgos potenciales. Nuevamente, estos riesgos incluyen la posibilidad de que la IA tenga sesgos importantes, de que se utilice con fines malintencionados o que provoque la pérdida de puestos de trabajo en algunos oficios o profesiones.
- *Implicaciones éticas y jurídicas*: El uso de la IA en la asistencia sanitaria plantea una serie de cuestiones éticas y jurídicas. Entre ellas, el derecho a la intimidad, el derecho al consentimiento informado y la posibilidad de que se reproduzcan ciertos sesgos sociales embebidos en los datos.
- *Disponibilidad de datos*: El desarrollo de este tipo de soluciones requiere el acceso a una gran cantidad de datos. Estos datos pueden ser difíciles de obtener, y es importante garantizar que sean precisos y representativos de la población objetivo a la vez que sean tratados de forma confidencial. El desafío radica en encontrar el punto medio entre la disponibilidad de datos abiertos para promover el desarrollo de herramientas basadas en IA y la preservación rigurosa de la privacidad y la confidencialidad de la información personal, para asegurar que estos desarrollos se lleven a cabo de manera éticamente correcta y responsable.
- *Costos de desarrollo y aplicación*: Dados los costos implicados en el desarrollo y despliegue de soluciones sanitarias basadas en IA, es importante garantizar que estas soluciones sean asequibles y accesibles para todos los usuarios y pacientes.

La inclusión de la mirada de todos los actores clave desde el inicio del proyecto, resulta fundamental para garantizar productos que agreguen valor en la cadena productiva de salud y que a la vez no generen daño. Esto incluye a:

- *Agencias gubernamentales*: Los gobiernos desempeñan un papel en la regulación del desarrollo y el uso de la IA en la asistencia sanitaria. Por ejemplo, las agencias que regulan la administración de medicamentos y tecnología sanitaria desempeñan un papel clave en la aprobación de dispositivos médicos basados en inteligencia artificial.
- *Organizaciones de salud*: Las instituciones prestadoras de servicios, públicas y privadas, son responsables de implantar soluciones basadas en IA en sus entornos clínicos. También desempeñan un papel importante en la recopilación y gestión de los datos que se utilizan para entrenar los modelos.

- *Equipos de investigadores y desarrolladores de IA, incluyendo a los eticistas*: Estos equipos son responsables de crear y mantener soluciones sanitarias impulsadas por inteligencia artificial. Trabajan con las organizaciones sanitarias para comprender las necesidades de los médicos y los pacientes, y para desarrollar soluciones que satisfagan esas necesidades.
- *Pacientes y usuarios*: Para la mayoría de los casos de uso los pacientes son los usuarios finales o los receptores finales de decisiones basadas en herramientas de IA. Tienen un rol que desempeñar para garantizar que estas soluciones sean seguras, eficaces y accesibles.
- *Instituciones de investigación*, *universidades e instituciones académicas:* Estas instituciones investigan el uso de la IA en la asistencia sanitaria. Desarrollan nuevas soluciones basadas en IA y trabajan para comprender las implicaciones éticas y jurídicas de la IA en la asistencia sanitaria.
- *Asociaciones de la sociedad civil (Organizaciones No Gubernamentales, Sociedades Científicas)*: Las asociaciones del sector trabajan para promover el desarrollo y uso responsables de la IA en salud. Elaboran normas y directrices para el uso de la IA, y también ofrecen educación y formación a los profesionales sanitarios.
- *Industria de tecnología en salud*: La industria de tecnologías en salud juega un papel crucial en el desarrollo responsable de la inteligencia artificial. No solo son desarrolladores y financiadores de tecnologías, sino que su experiencia en la regulación y cumplimiento de estándares éticos les otorga una posición ventajosa para promover buenas prácticas y hasta influir en agendas. No obstante, es importante tener en cuenta los claros conflictos de interés que un actor como este presenta.

7.2. Planificación y diseño

Ética embebida

Como mencionamos previamente, es durante el proceso de diseño de sistemas basados en IA, que la consideración de aspectos éticos se erige como un imperativo fundamental. Tanto desde la constitución del equipo de diseño, como desde la selección de fuentes de datos y los casos de uso contemplados.

Una vez definidos los aspectos éticos a considerar, es fundamental **identificar a los responsables de embeber estos aspectos en el desarrollo**. En este sentido, Miller[14] identifica dos cuestiones cruciales a la hora de definir responsables para la incorporación de aspectos éticos y control de sesgos en los procesos de desarrollo de aplicaciones basadas en inteligencia artificial: en primer lugar, **la multiplicidad de perfiles enrolados bajo la categoría de desarrolladores** (a los que considera actores responsables en términos éticos de los desarrollos[14–18]). Aquí pueden confluir técnicos, diseñadores, financiadores, etc. En segundo lugar, destaca que, en el marco de los proyectos dentro de los cuales se desarrollan las soluciones, suele haber cambios en los equipos de trabajo lo que determina que a lo largo de dichos proyectos entren y salgan profesionales con

perfiles diferentes. Esto requiere una mirada amplia acerca del rol de los actores o "*stakeholders*", tradicionalmente definido como los jugadores involucrados en un proyecto y quienes, a su vez, son impactados por sus resultados. En este sentido Miller[14] incorpora la figura del *stakeholder* pasivo que incluye a otros actores que, sin participar activamente del proceso, pueden llegar a verse afectados, esto implica generar un balance que incluya en la valoración a la comunidad (y todos los problemas étnicos, culturales y sociales que esto implica) y el medio ambiente [19].

Desarrollo propio o reutilización

Parte de la discusión durante la fase de diseño implica decidir respecto de la posibilidad de reutilizar herramientas ya desarrolladas. Esto es particularmente relevante en el caso de países en vías de desarrollo que buscan adoptar tecnologías desarrolladas e implementadas en los países desarrollados. Se han destacado principalmente los inconvenientes asociados a la adopción de soluciones basadas en otros contextos étnicos, culturales y sociales[20–22], desde la aplicación de modelos entrenados originalmente en otros lenguajes, hasta aplicaciones que se las asume "plug & play", es decir de uso directo sin necesidad de adaptación local, sin tener en cuenta la diferencia en los procesos que existe. En ese contexto, el papel de las partes involucradas en cada parte del proceso y sus habilidades son determinantes.

Gobernanza

Otro aspecto crucial es la gobernanza de datos, definida como la "**lógica organizadora de la gestión de datos: recolección, almacenamiento, procesamiento, uso, intercambio y destrucción**". Janssen y cols. proponen un marco para la gobernanza de datos que apunta a garantizar que los datos correctos se compartan de manera segura y confiable, y que el intercambio cumpla con regulaciones [23]. Este marco también promueve la apertura controlada de datos y algoritmos para permitir el escrutinio externo, el intercambio de información confiable dentro y entre organizaciones, la gobernanza basada en el riesgo, los controles a nivel del sistema y el control de datos a través de identidades auto-soberanas (self-sovereign identities)[2] y de propiedad compartida.

7.3. Desarrollo y validación

En la bibliografía se pueden distinguir **tres desafíos éticos** claves que enfrenta la implementación de la IA en la práctica médica y sanitaria. Ellos son: **los sesgos potenciales en los modelos de IA, la protección de la privacidad del paciente y la confianza de los médicos, usuarios y público en general** en la incorporación de la IA en la atención médica y de la salud [24,25].

Generación de datos

La construcción de bases de datos o "datasets" para modelos de aprendizaje automático (machine learning) constituye un proceso esencial en el desarrollo de sistemas inteligentes y automatizados

[2] Una "identidad auto-soberana" se refiere a un sistema de gestión de datos personales en el cual un individuo tiene control completo y autónomo sobre su propia información, permitiéndole compartir selectivamente detalles específicos de su identidad de manera segura y confiable en línea.

[26]. Sin embargo, **es crucial reconocer y abordar los sesgos inherentes que pueden surgir en dichos conjuntos de datos**.

Los sesgos en los datasets pueden manifestarse de diversas formas. Uno de los más comunes es el **sesgo de selección**, que surge cuando los datos recopilados no representan adecuadamente la diversidad y la variabilidad presentes en la población real [27]. Por ejemplo, en aplicaciones como el reconocimiento facial, la subrepresentación de etnias minoritarias puede dar lugar a un funcionamiento insatisfactorio para dichos grupos [28]. Lo mismo se ha descrito por ejemplo en aplicaciones de identificación de lesiones cutáneas [29,30]. Estos sesgos pueden ser amplificados durante el proceso de entrenamiento, ya que los algoritmos tienden a aprender patrones de los datos proporcionados, independientemente de si son adecuados o no.

Por otro lado, los **sesgos de etiquetado** pueden introducirse cuando las anotaciones son subjetivas o reflejan percepciones culturales y sociales. Esto surge dado que, los anotadores, es decir, quienes etiquetan o clasifican los datos de entrenamiento para un modelo, son personas insertas en una sociedad y llevan a esta tarea sus propios sesgos. Por ejemplo, al momento de clasificar mensajes en redes sociales de acuerdo a su polaridad, los anotadores pueden condicionar la clasificación dependiendo de cómo interpretan el género de quién escribió el mensaje. Esto más frecuente cuando la tarea de anotación es compleja (asignar polaridad a un texto, detectar sarcasmo o ironía, detección de discursos de odio, diagnósticos médicos). Esto en parte puede ser mitigado con manuales de anotación[3] detallados y probados en terreno que aseguren niveles adecuados de estandarización en la anotación.

Para abordar efectivamente estos desafíos, se requiere una combinación de enfoques.

1. En primer lugar, se debe llevar a cabo un análisis exhaustivo de los datos en busca de sesgos potenciales. Esto implica evaluar las distribuciones demográficas, las relaciones de género, las características étnicas y otras variables relevantes para asegurarse de que el dataset refleje la diversidad de la población [32].
2. Posteriormente, se deben implementar estrategias de pre-procesamiento[4] como por ejemplo la reponderación de muestras (que realiza un ajuste de los pesos de las clases) o la generación de datos sintéticos (que crea ejemplos artificiales similares a los existentes para grupos subrepresentados) [32]. Estas técnicas son necesarias cuando existe un desequilibrio en los conjuntos de datos en relación a alguna de las variables de interés. Balancear o equilibrar las clases mediante alguna de las técnicas mencionadas mejora notablemente la capacidad del modelo de generalizar cuando existen clases subrepresentadas o sesgos en nuestros conjuntos de datos. La incorporación de la retroalimentación y la revisión por parte de expertos en el dominio también es esencial en esta etapa para garantizar una identificación y mitigación efectivas de los sesgos [32,33]**.**

[3] Documento detallado que proporciona pautas y directrices específicas a los anotadores humanos sobre cómo etiquetar y anotar correctamente los datos de entrenamiento para un modelo de aprendizaje automático.
[4] Se refiere al conjunto de técnicas utilizados para preparar y limpiar los datos antes de entrenar un modelo, habitualmente con el objetivo de lograr una mejor performance y generalización del mismo.

Además, **la transparencia y la documentación detallada del proceso de construcción del dataset son fundamentales**. Los equipos de investigación deben registrar todas las decisiones tomadas, desde la selección de las fuentes de datos hasta los métodos de limpieza y etiquetado. Esto permite una evaluación crítica externa y facilita la detección temprana de posibles sesgos inadvertidos [33]. Complementariamente, se debe **fomentar la colaboración interdisciplinaria, involucrando a expertos en ética, diversidad y sociología, junto con los ingenieros de machine learning, para garantizar una perspectiva integral y una comprensión profunda de los posibles impactos sociales y éticos, desde el inicio del proyecto** [8].

Privacidad de datos

En el ámbito de la IA y la salud, es común que las herramientas desarrolladas se entrenen con o utilicen como input una cantidad considerable de datos personales y clínicos de los pacientes. Esto significa que existe un riesgo inherente de que esta información sensible pueda ser "hackeada" o comprometida de alguna manera. Por lo tanto, es crucial garantizar, desde nuestros diversos roles, que estos desarrollos vayan acompañados del respeto a la privacidad y la confidencialidad.

Afortunadamente, **se han establecido normas y guías para gestionar el uso de datos personales en el contexto de la IA**. Por ejemplo, la Red Iberoamericana de Protección de Datos, que incluye a 22 autoridades de protección de datos de países como Portugal, España, México y otros de Centroamérica, Sudamérica y el Caribe, ha publicado recomendaciones para el tratamiento de datos personales en inteligencia artificial [34]. De estas recomendaciones destacamos algunas: diseñar esquemas apropiados de Gobernanza sobre tratamiento de datos personales en las organizaciones que realizan desarrollos de IA; asegurar la calidad de los datos personales, utilizar herramientas de anonimización; incrementar la confianza y transparencia con los titulares de los datos personales. Este tipo de iniciativas son fundamentales para garantizar que los avances en IA no comprometan la privacidad y seguridad de los datos personales.

Pre-procesamiento de datos

La fase de pre-procesamiento de los datos, con la finalidad de rectificar inexactitudes desde sus fuentes originales, ostenta un grado de trascendencia importante en la salvaguarda de la precisión y equidad inherentes a un modelo dado. La labor en este ámbito demanda la **capacidad de discernir de manera clara los grupos vulnerables dentro de los datos en cuestión (tales como aquellos derivados de variables de género o identidad étnica), permitiéndonos así medir las disparidades en la integridad y completitud de los datos entre dichos grupos**. El trabajo de Larrazabal y cols.[37] explora este aspecto en modelos de aprendizaje automático aplicados a imágenes de rayos x para predecir diferentes diagnósticos . En este trabajo los investigadores entrenaron diferentes modelos utilizando conjuntos de datos de entrenamiento con variados niveles de desequilibrio de género, como 25% de hombres y 75% de mujeres o 0% y 100%, y luego evaluaron cómo estos modelos se desempeñaban en la detección de patologías en imágenes de personas de ambos sexos por separado. Evidenciaron que cuando el desequilibrio de género en los datos de entrenamiento era alto, el rendimiento del modelo disminuía significativamente en el grupo subrepresentado, e incluso en casos de desequilibrio intermedio, como 25% de hombres y 75% de mujeres, el rendimiento del

modelo en el grupo minoritario se veía afectado negativamente. Esto tiene importantes implicancias, ya que la falta de consideración de este desequilibrio podría dar lugar a la generación de falsos positivos o falsos negativos en las predicciones de enfermedades que afecten particularmente al grupo sub representado.

En situaciones más desafiantes, en las cuales los conjuntos de datos presentan problemas más severos, se proponen procedimientos complejos de limpieza de los datos. Un ejemplo de esto es la estrategia de preprocesamiento MLClean, propuesta por Tae y cols[36], E, que busca integrar en un solo pipeline de pre-procesamiento, la limpieza de datos (remoción de duplicados, corrección de valores erróneos), la mitigación de sesgos mediante la reponderación de los datos en base a variables relacionadas con estos sesgos y la sanitización de los datos, es decir la eliminación de información confidencial o sensible.

Entrenamiento y validación del modelo

Las métricas de rendimiento o performance en un modelo de inteligencia artificial deben seleccionarse cuidadosamente según su propósito. Por ejemplo, en problemas de clasificación, la evaluación puede enfocarse en determinar si es más relevante focalizarse en falsos positivos (por ejemplo, radiografías normales clasificadas erróneamente por el modelo como “con neumonía”) o en los falsos negativos (utilizando el mismo ejemplo previo, radiografías “con neumonía” que se clasifican como normales), dependiendo del contexto. No obstante, al evaluar un modelo de IA, también es crucial **utilizar métricas que reflejen de manera equitativa su rendimiento en diferentes grupos o segmentos de la población, especialmente aquellos que se consideren vulnerables**. Es así que la utilización de métricas globales puede resultar engañosas y sesgadas, ya que podría ocultar problemas de discriminación o subrepresentación en grupos minoritarios. Una alternativa más adecuada sería evaluar métricas para cada grupo específico, como diferentes grupos étnicos, géneros u otras características sensibles, por ejemplo, métricas habituales como el F1 score[5], o incluso la exactitud[6] (accuracy en inglés) pueden ser métricas adecuadas si se desagrega por género o raza. Esto puede mitigar especialmente problemas que aparecen al incluir variables en un modelo que mejoran la performance general, pero la empeoran dentro de determinados grupos.

En casos en que los modelos muestran un bajo rendimiento en las métricas de performance, sean de clasificación o regresión, en alguno de los subgrupos evaluados, es recomendable analizar los ejemplos dentro de ese subgrupo y considerar la posibilidad de reentrenar el modelo. Sin embargo, los acúmulos (clusters) de ejemplos similares aún pueden tener una alta variabilidad de características, lo que dificulta su resumen e interpretación. Por lo tanto, es **fundamental encontrar una técnica efectiva para detectar subpoblaciones en donde las métricas de rendimiento sean**

[5] Métrica que combina la sensibilidad (o también “recall” en inglés) y el valor predictivo positivo (o “precision” en inglés) de un modelo en una sola medida, proporcionando una evaluación balanceada del rendimiento al considerar tanto los falsos positivos como los falsos negativos.
[6] Métrica que mide la proporción de predicciones correctas realizadas por un modelo sobre el total de predicciones realizadas

deficientes y que al mismo tiempo permita identificar subconjuntos de datos fáciles de entender [8].

La identificación de subconjuntos problemáticos a través de herramientas como Slice Finder[41] ayuda a los usuarios permite mejorar la equidad del modelo y brindar resultados más confiables y responsables en la toma de decisiones a través de la identificación de subconjuntos interpretables de datos en los que el modelo tiene un rendimiento deficiente. Alternativamente, es posible realizar auditorías periódicas para identificar los subgrupos con bajo rendimiento, mejorando así iterativamente el rendimiento del modelo para los subgrupos de interés. Algunos sesgos pueden incluso corregirse con técnicas que impliquen un nuevo etiquetado, como el caso del Reinforcement Learning from Human Feedback, utilizadas en el difundido chatbot, chatGPT [42].

Las prácticas para abordar la equidad deben ir acompañadas del reconocimiento realista por parte de los equipos. **Siempre que sea posible, el algoritmo de IA debe probarse en múltiples instituciones de salud, grupos socioeconómicos y rangos de edad** [8,24].

La búsqueda continua de enfoques para abordar problemas de equidad en la inteligencia artificial es fundamental para avanzar hacia un desarrollo tecnológico más inclusivo y justo.

7.4. Despliegue e implementación

Explicabilidad del modelo

La noción de explicabilidad se refiere a la capacidad inherente de un sistema de inteligencia artificial para reconstruir el proceso subyacente mediante el cual arriba a determinadas predicciones o resultados específicos [43,44]. Este atributo ostenta una significación fundamental, tanto en el contexto de la adaptación de tales sistemas como en el marco de la imperativa evaluación ética que rige su operatividad.

Las distintas categorías de sistemas basados en IA engendran desafíos heterogéneos en lo que respecta a su nivel de explicabilidad. Por ejemplo, en el caso de los sistemas expertos o sistemas de inteligencia artificial simbólica, ampliamente empleados en el dominio de la salud, se procede a la codificación de los saberes clínicos o médicos, haciendo uso de algoritmos basados en reglas para la toma de decisiones. Estos sistemas se adecuan mediante la adaptación o modificación de reglas establecidas por la comunidad científica, aplicándolas a un conjunto de casos de referencia. Esto confiere a dichos sistemas un grado de explicabilidad sumamente elevado, dado que cada regla se encuentra codificada de manera explícita y, por ende, cada decisión del algoritmo puede ser seguida en su trazado hasta una regla o una combinación de las mismas.

En contraposición, **los algoritmos de aprendizaje automático se abocan a encontrar patrones intrínsecos en los datos, con el propósito de alcanzar un grado óptimo de generalización**. Estos algoritmos "aprenden" mediante el ajuste de sus parámetros con base en datos de entrenamiento,

optimizando una función de pérdida (es decir, una función que cuantifica la discrepancia entre las predicciones del modelo y los valores reales de entrenamiento) con miras a resolver tareas específicas. **En el contexto de las redes neuronales profundas, la multiplicidad de estratos de cómputos distribuidos, que median entre los inputs y outputs, oscurece el procedimiento subyacente y los asemeja a una "caja negra".** Por consiguiente, es difícil comprender sus predicciones ya que no suele ser fácil "descomponer" o desentrañar el flujo informativo, sea numéricamente o en una representación visual, a diferencia de modelos más sencillos, tales como la regresión logística o los árboles de decisión [45].

Esto implica que, **a medida que aumenta la complejidad inherente al algoritmo con fines de mejorar la performance de predicción, se intensifica la dificultad para dilucidar con precisión qué regla o conjunto de reglas ha sido instrumental en la generación de la predicción efectuada** [45].

La contraposición entre complejidad, interpretabilidad o nivel de explicabilidad se configura como uno de los desafíos preponderantes en el trayecto de la adopción de dichas herramientas en el ámbito de la salud.

La explicabilidad en IA puede ser intrínseca al algoritmo que se va a utilizar (por ejemplo, regresiones lineales, árboles de decisión) [46] o puede ser una aproximación que se realiza por otros métodos (por ejemplo, LIME[47] o SHAP[48]) extrínsecos al modelo. Esta diferenciación puede ayudar a comprender la común denominación de "cajas negras" a algunos métodos de la IA. Es importante destacar que la explicabilidad inherente a un algoritmo resultará normalmente más precisa que otros métodos de explicabilidad aproximada, aunque normalmente también tiene una performance menor, como es el caso de una regresión lineal o logística comparado con las redes neuronales convolucionales o multicapa.

De este modo, podemos decir que **existe una compensación o "trade-off" entre la explicabilidad y la performance del modelo a la cual debemos considerar al momento del desarrollo y validación de nuestra herramienta y, sobre todo, cuando se utiliza para apoyar las decisiones clínicas.** Considerando la creciente preferencia por métodos de alto rendimiento y la necesidad de explicabilidad que existe en el ámbito de la salud, **se debe priorizar un enfoque donde se evalúen críticamente los métodos de explicabilidad y, además, se aborden desde las múltiples partes interesadas priorizando la pluralidad, intersectorialidad e interdisciplina**.

Visto desde el lado del desarrollo, la explicabilidad resulta importante para validar modelos a partir de la coherencia y no solamente a partir de su rendimiento. Un ejemplo de aplicación médica que refleja este problema lo describen Zech y cols[49]. En su trabajo describen el impacto de variables confundidoras (por ejemplo, metadatos de las imágenes embebidos en las placas radiográficas) sobre la performance del modelo, lo cual claramente es un problema para la generalización. En este sentido, es necesario trabajar en conjunto con los científicos de datos y los ingenieros en comprender el tipo de explicabilidad o interpretabilidad que se necesita en particular para la herramienta que se quiere desarrollar.

El desarrollo de los algoritmos será diferente si se busca una explicabilidad local (donde se quiere explicar una predicción en particular, qué variables tuvieron peso en esa decisión o qué reglas se

aplicaron para llegar a ella) o general (donde se quiere comprender el modelo, como por ejemplo los pesos de las variables incluidas o aproximar el modelo a un conjunto de reglas o árbol de decisiones) [50]. De este modo, **desarrollar modelos que busquen satisfacer la necesidad de explicabilidad que tienen los usuarios, así como utilizar enfoques centrados en humanos que busque mantener a las personas informadas** [50], **permitirá abordar los problemas éticos y sociales asociados con el uso de la IA en salud**.

7.5. Operación y monitoreo

Una vez que estos modelos hayan sido desplegados en producción, es decir, que ya se utilizan dentro de procesos en el mundo real, resulta esencial llevar a cabo una **evaluación continua a lo largo del tiempo**, con el propósito de identificar de manera temprana cualquier indicio de deterioro en su rendimiento. Esto adquiere relevancia debido a la **posibilidad de que las modificaciones en la composición de la población a lo largo del tiempo difieran significativamente de las condiciones originalmente consideradas durante la fase de despliegue inicial del modelo**. Feng y colaboradores

[51] hacen una descripción detallada del proceso de monitoreo y sugieren la **conformación de áreas de mejora de la calidad de este tipo de sistemas dentro de las instituciones de salud**. La realización de este proceso de monitoreo se erige como un pilar fundamental, debiendo abordarse con una consideración integral de los elementos expuestos previamente. Ejemplificando, resulta imperativo llevar a cabo una vigilancia rigurosa del rendimiento del modelo en contextos que involucren grupos vulnerables, así como los cambios en la distribución de sus variables.

8. Conclusiones

La rápida evolución de la inteligencia artificial en los últimos años ha planteado una serie de desafíos éticos que demandan una atención rigurosa y una reflexión profunda. Los avances en IA han demostrado un potencial significativo para transformar diversas industrias, pero al mismo tiempo han generado preocupaciones en torno a la privacidad, el sesgo algorítmico, la responsabilidad y la sustitución de tareas humanas. Abordar estos aspectos éticos de manera efectiva requiere la colaboración entre profesionales de la salud, científicos de datos, ingenieros, legisladores y expertos en bioética.

Es imperativo establecer marcos normativos sólidos que guíen el desarrollo y la implementación de la IA, asegurando que los beneficios se maximicen mientras se minimizan los posibles daños.

La búsqueda de soluciones éticas en el ámbito de la IA no solo garantiza la integridad y confiabilidad de las tecnologías emergentes, sino que también promueve un enfoque responsable y sostenible hacia la innovación tecnológica en beneficio de la sociedad en su conjunto.

Referencias

1. European Group on Ethics in Science and New technologies. RTD:Directorate-General for Research, Innovation, corporate-body. ETHI:European Group on Ethics, New Technologies. Statement on artificial intelligence, robotics and “autonomous” systems : Brussels, 9 March 2018. Publications Office of the European Union; 2018. Recuperado: https://data.europa.eu/doi/10.2777/531856

2. Cortina A, Orts AC. Etica de la empresa: claves para una nueva cultura empresarial. Trotta; 1994. Recuperado: https://play.google.com/store/books/details?id=IUIQSgAACAAJ

3. Shermer M. Morality is real, objective, and natural. Ann N Y Acad Sci. 2016;1384: 57–62. doi:10.1111/nyas.13077

4. Schwab K, World Economic Forum. The Fourth Industrial Revolution: what it means and how to respond. En: World Economic Forum [Internet]. 14 de enero de 2016 [citado 11 de agosto de 2023]. Recuperado: https://www.weforum.org/agenda/2016/01/the-fourth-industrial-revolution-what-it-means-and-how-to-respond/

5. Ethics and governance of artificial intelligence for health. World Health Organization; 28 de junio de 2021 [citado 11 de agosto de 2023]. Recuperado: https://www.who.int/publications/i/item/9789240029200

6. Recommendation on the Ethics of Artificial Intelligence. [citado 2 de agosto de 2023]. Recuperado: https://unesdoc.unesco.org/ark:/48223/pf0000380455

7. White Paper on Artificial Intelligence: a European approach to excellence and trust. En: European Commission [Internet]. [citado 11 de agosto de 2023]. Recuperado: https://commission.europa.eu/publications/white-paper-artificial-intelligence-european-approach-excellence-and-trust_en

8. Solanki P, Grundy J, Hussain W. Operationalising ethics in artificial intelligence for healthcare: a framework for AI developers. AI and Ethics. 2023;3: 223–240. doi:10.1007/s43681-022-00195-z

9. McLennan S, Fiske A, Tigard D, Müller R, Haddadin S, Buyx A. Embedded ethics: a proposal for integrating ethics into the development of medical AI. BMC Med Ethics. 2022;23: 6. doi:10.1186/s12910-022-00746-3

10. Davis SLM. The Trojan Horse: Digital Health, Human Rights, and Global Health Governance. Health Hum Rights. 2020;22: 41–47. Recuperado: https://www.ncbi.nlm.nih.gov/pubmed/33390691

11. Ghassemi M, Oakden-Rayner L, Beam AL. The false hope of current approaches to explainable artificial intelligence in health care. Lancet Digit Health. 2021;3: e745–e750. doi:10.1016/S2589-7500(21)00208-9

12. Char DS, Shah NH, Magnus D. Implementing Machine Learning in Health Care - Addressing Ethical Challenges. N Engl J Med. 2018;378: 981–983. doi:10.1056/NEJMp1714229

13. Mulya MODP, Ali M. Artificial Intelligence crime within the concept of society 5.0: Challenges and opportunities for acknowledgment of Artificial Intelligence in Indonesian Criminal Legal System. International Journal of Law and Politics Studies. 2023;5: 07–15. doi:10.32996/ijlps.2023.5.1.2

14. Miller GJ. Stakeholder roles in artificial intelligence projects. Project Leadership and Society. 2022;3: 100068. doi:10.1016/j.plas.2022.100068

15. Wieringa M. What to account for when accounting for algorithms: a systematic literature review on algorithmic accountability. Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency. New York, NY, USA: Association for Computing Machinery; 2020. pp. 1–18. doi:10.1145/3351095.3372833

16. Touretzky DS, Cune CG-M, Martin F, Seehorn D. K-12 guidelines for artificial intelligence: What students should know. [citado 25 de agosto de 2023]. Recuperado: https://upload01.uocslive.com/ISTE/ISTE2019/PROGRAM_SESSION_MODEL/HANDOUTS/112142285/ISTE2019Presentation_final.pdf

17. Manders-Huits N. Moral responsibility and IT for human enhancement. Proceedings of the 2006 ACM symposium on Applied computing. New York, NY, USA: Association for Computing Machinery; 2006. pp. 267–271. doi:10.1145/1141277.1141340

18. Martin KE. Ethical Implications and Accountability of Algorithms. SSRN; 2018. Recuperado: https://play.google.com/store/books/details?id=6kb8zgEACAAJ

19. Derry R. Reclaiming Marginalized Stakeholders. J Bus Ethics. 2012;111: 253–264. doi:10.1007/s10551-012-1205-x

20. Mancilla-Caceres JF, Estrada-Villalta S. The Ethical Considerations of AI in Latin America. Digital Society. 2022;1: 16. doi:10.1007/s44206-022-00018-y

21. Buolamwini J, Gebru T. Gender Shades: Intersectional Accuracy Disparities in Commercial Gender Classification. En: Friedler SA, Wilson C, editores. Proceedings of the 1st Conference on Fairness, Accountability and Transparency. PMLR; 23--24 Feb 2018. pp. 77–91. Recuperado: https://proceedings.mlr.press/v81/buolamwini18a.html

22. Stray V, Hoda R, Paasivaara M, Kruchten P, van der Aalst W, Mylopoulos J, et al. Agile processes in software engineering and extreme programming: 21st international conference on agile software development, XP 2020, Copenhagen, Denmark, June 8-12, 2020, proceedings. 1ª ed. Stray V, Hoda R, Paasivaara M, Kruchten P, editores. Cham, Switzerland: Springer Nature; 2020. doi:10.1007/978-3-030-49392-9

23. Janssen M, Brous P, Estevez E, Barbosa LS, Janowski T. Data governance: Organizing data for trustworthy Artificial Intelligence. Gov Inf Q. 2020;37: 101493. doi:10.1016/j.giq.2020.101493

24. Wiens J, Saria S, Sendak M, Ghassemi M, Liu VX, Doshi-Velez F, et al. Do no harm: a roadmap for responsible machine learning for health care. Nat Med. 2019;25: 1337–1340. doi:10.1038/s41591-019-0548-6

25. Reddy S, Allan S, Coghlan S, Cooper P. A governance model for the application of AI in health care. J Am Med Inform Assoc. 2020;27: 491–497. doi:10.1093/jamia/ocz192

26. Ghassemi M, Naumann T, Schulam P, Beam AL, Chen IY, Ranganath R. A Review of Challenges and Opportunities in Machine Learning for Health. AMIA Jt Summits Transl Sci Proc. 2020;2020: 191–200. doi:10.1001/jama.2017.18391

27. Obermeyer Z, Powers B, Vogeli C, Mullainathan S. Dissecting racial bias in an algorithm used to manage the health of populations. Science. 2019;366: 447–453. doi:10.1126/science.aax2342

28. Buolamwini JA. Gender shades : intersectional phenotypic and demographic evaluation of face datasets and gender classifiers. Massachusetts Institute of Technology. 2017. Recuperado: https://dspace.mit.edu/handle/1721.1/114068?show=full?show=full

29. Daneshjou R, Smith MP, Sun MD, Rotemberg V, Zou J. Lack of Transparency and Potential Bias in Artificial Intelligence Data Sets and Algorithms: A Scoping Review. JAMA Dermatol. 2021;157: 1362–1369. doi:10.1001/jamadermatol.2021.3129

30. Adamson AS, Smith A. Machine Learning and Health Care Disparities in Dermatology. JAMA Dermatol. 2018;154: 1247–1248. doi:10.1001/jamadermatol.2018.2348

31. Wiens J, Price WN 2nd, Sjoding MW. Diagnosing bias in data-driven algorithms for healthcare. Nat Med. 2020;26: 25–26. doi:10.1038/s41591-019-0726-6

32. Vokinger KN, Feuerriegel S, Kesselheim AS. Mitigating bias in machine learning for medicine. Commun Med. 2021;1: 25. doi:10.1038/s43856-021-00028-w

33. Norori N, Hu Q, Aellen FM, Faraci FD, Tzovara A. Addressing bias in big data and AI for health care: A call for open science. Patterns (N Y). 2021;2: 100347. doi:10.1016/j.patter.2021.100347

34. Red Iberoamericana de Protección de Datos. [citado 11 de agosto de 2023]. Recuperado: https://www.redipd.org/es

35. Rajkomar A, Hardt M, Howell MD, Corrado G, Chin MH. Ensuring Fairness in Machine Learning to Advance Health Equity. Ann Intern Med. 2018;169: 866–872. doi:10.7326/M18-1990

36. Tae KH, Roh Y, Oh YH, Kim H, Whang SE. Data Cleaning for Accurate, Fair, and Robust Models: A Big Data - AI Integration Approach. Proceedings of the 3rd International Workshop on Data Management for End-to-End Machine Learning. New York, NY, USA: Association for Computing Machinery; 2019. pp. 1–4. doi:10.1145/3329486.3329493

37. Larrazabal AJ, Nieto N, Peterson V, Milone DH, Ferrante E. Gender imbalance in medical imaging datasets produces biased classifiers for computer-aided diagnosis. Proc Natl Acad Sci U S A. 2020;117: 12592–12594. doi:10.1073/pnas.1919012117

38. Sramka M, Safavi-Naini R, Denzinger J, Askari M. A practice-oriented framework for measuring privacy and utility in data sanitization systems. Proceedings of the 2010 EDBT/ICDT Workshops. New York, NY, USA: Association for Computing Machinery; 2010. pp. 1–10. doi:10.1145/1754239.1754270

39. Gehrke J, Hay M, Lui E, Pass R. Crowd-Blending Privacy. Advances in Cryptology – CRYPTO 2012. Springer Berlin Heidelberg; 2012. pp. 479–496. doi:10.1007/978-3-642-32009-5_28

40. Pessach D, Shmueli E. A Review on Fairness in Machine Learning. ACM Comput Surv. 2022;55: 1–44. doi:10.1145/3494672

41. Chung Y, Kraska T, Polyzotis N, Tae KH, Whang SE. Slice Finder: Automated Data Slicing for Model Validation. 2019 IEEE 35th International Conference on Data Engineering (ICDE). 2019. pp. 1550–1553. doi:10.1109/ICDE.2019.00139

42. Stiennon N, Ouyang L, Wu J, Ziegler D, Lowe R, Voss C, et al. Learning to summarize with human feedback. Adv Neural Inf Process Syst. 2020;33: 3008–3021. Recuperado: https://proceedings.neurips.cc/paper_files/paper/2020/hash/1f89885d556929e98d3ef9b86448f951-Abstract.html

43. Vilone G, Longo L. Notions of explainability and evaluation approaches for explainable artificial intelligence. Inf Fusion. 2021;76: 89–106. doi:10.1016/j.inffus.2021.05.009

44. Combi C, Amico B, Bellazzi R, Holzinger A, Moore JH, Zitnik M, et al. A manifesto on explainability for artificial intelligence in medicine. Artif Intell Med. 2022;133: 102423. doi:10.1016/j.artmed.2022.102423

45. Barredo Arrieta A, Díaz-Rodríguez N, Del Ser J, Bennetot A, Tabik S, Barbado A, et al. Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI. Inf Fusion. 2020;58: 82–115. doi:10.1016/j.inffus.2019.12.012

46. Payrovnaziri SN, Chen Z, Rengifo-Moreno P, Miller T, Bian J, Chen JH, et al. Explainable artificial intelligence models using real-world electronic health record data: a systematic scoping review. J Am Med Inform Assoc. 2020;27: 1173–1185. doi:10.1093/jamia/ocaa053

47. Ribeiro MT, Singh S, Guestrin C. "Why Should I Trust You?": Explaining the Predictions of Any Classifier. Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. New York, NY, USA: Association for Computing Machinery; 2016. pp. 1135–1144. doi:10.1145/2939672.2939778

48. Lundberg SM, Lee S-I. A unified approach to interpreting model predictions. Adv Neural Inf Process Syst. 2017;30. Recuperado: https://proceedings.neurips.cc/paper_files/paper/2017/hash/8a20a8621978632d76c43dfd28b67767-Abstract.html

49. Zech JR, Badgeley MA, Liu M, Costa AB, Titano JJ, Oermann EK. Variable generalization performance of a deep learning model to detect pneumonia in chest radiographs: A cross-sectional study. PLoS Med. 2018;15: e1002683. doi:10.1371/journal.pmed.1002683

50. Liao QV, Gruen D, Miller S. Questioning the AI: Informing Design Practices for Explainable AI User Experiences. Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems. New York, NY, USA: Association for Computing Machinery; 2020. pp. 1–15. doi:10.1145/3313831.3376590

51. Feng J, Phillips RV, Malenica I, Bishara A, Hubbard AE, Celi LA, et al. Clinical artificial intelligence quality improvement: towards continual monitoring and updating of AI algorithms in healthcare. NPJ Digit Med. 2022;5: 66. doi:10.1038/s41746-022-00611-y

CLIAS

CENTRO DE INTELIGENCIA
ARTIFICIAL Y SALUD
PARA AMÉRICA LATINA
Y EL CARIBE

CiPS IMPLEMENTACIÓN E INNOVACIÓN EN POLÍTICAS DE SALUD

IECS INSTITUTO DE EFECTIVIDAD CLÍNICA Y SANITARIA