www.komputersapiens.org

KOMPUTER SAPIENS

Año 15
Volumen 3
Sept - Dic 2023

Revista de Divulgación de la Sociedad Mexicana de Inteligencia Artificial



INTELIGENCIA ARTIFICIAL **EXPLICABLE**

TRAMPAS ÉTICAS EN EL **MANEJO DE DATOS**

APLICACIONES Y DESAFÍOS DE LA IA EN LA **AERONAÚTICA**

ChatGPT: ÉTICA Y DESAFÍOS EN ESA TECNOLOGÍA

IA PARA EL **DIAGNÓSTICO** EN PERSONAS **ADULTAS MAYORES**

LA NECESIDAD DE LA **ÉTICA** ANTE LA EXPANSIÓN **DE LA IA**

ARCHIVÍSTICA E INTELIGENCIA ARTIFICIAL

ÉTICA EN EL USO DE LA IA EN APLICACIONES **CON ANIMALES**

ISSN 2007-0691

©Komputer Sapiens, Año XV Volumen III, septiembre-diciembre 2023, es una publicación cuatrimestral de la Sociedad Mexicana de Inteligencia Artificial, A.C., con domicilio en Ezequiel Montes 56 s/n, Fracc. los Pilares, Metepec, Edo. de México, C.P. 52159, México, http://www.komputersapiens.org, correo electrónico: editorial@komputersapiens.org, tel. +52 (833)357.48.20 ext. 3024, fax +52 (833) 215.85.44. Impresa por Sistemas y Diseños de México S.A. de C.V., calle Aragón No. 190, colonia Álamos, delegación Benito Juárez, México D.F., C.P. 03400, México, se terminó de editar el 31 de agosto de 2023.
Reserva de derechos al uso exclusivo número 04-2009-111110040200-102 otorgado por el Instituto Nacional de Derechos de Autor. ISSN 2007-0691.
Los artículos y columnas firmados son responsabilidad exclusiva de los autores y no reflejan necesariamente los puntos de vista de la Sociedad Mexicana de Inteligencia Artificial. La mención de empresas o productos específicos en las páginas de Komputer Sapiens no implica su respaldo por la Sociedad Mexicana de Inteligencia Artificial.
Queda estrictamente prohibida la reproducción total o parcial por cualquier medio, de la información aquí contenida sin autorización por escrito de los editores.
Komputer Sapiens es una revista de divulgación en idioma español de temas relacionados con la inteligencia artificial.
Creada en LaTeX, con la clase **papertex** disponible en el repositorio *CTAN*: Comprehensive TeX Archive Network, http://www.ctan.org/
Indizada en el IRMDCT de CONACYT y en Latindex.

Directorio SMIA

Presidente	Hiram Calvo Castro
Vicepresidenta	María de Lourdes Martínez Villaseñor
Secretario	Hiram Eredin Ponce Espinosa
Tesorera	Obdulia Pichardo Lagunas
Vocales:	Ildar Batyrshin
	Sabino Miranda Jiménez
	Roberto A. Vázquez Espinoza de los Monteros
	Oscar Herrera Alcántara
	David E. Pinto Avendaño
	Noé Alejandro Castro Sánchez
	Iris Iddaly Méndez Gurrola
	Omar Montaño Rivas
	Antonio Marín Hernández
	Nestor Velasco Bermeo
	Gustavo Arroyo Figueroa
	Sofía N. Galicia Haro
	Juan C. Martínez Miranda
	Leobardo Morales Tiburcio
	Iván Olvera Rodríguez
	Gilberto Ochoa Ruíz
	César J. Nuñez Prado
	Víctor Josué Ruiz Martínez
	Bella Citlali Martínez Seis

Komputer Sapiens

Director general	Ildar Batyrshin
Editora en jefe	Karina Mariela Figueroa Mora
Editor científico	Héctor Gabriel Acosta Mesa
Editores asociados	Marco Antonio Aguirre Lam
	Laura Cruz Reyes
Editores invitados	Marisol Flores Garrido
	Luis Miguel García Velázquez
Coordinadora de producción	Viridiana Mena Gómez
e-Tlakuilo	Karina Figueroa
	Rosa María Valdovinos
Estado del IArte	Reyna Carolina Medina Ramírez
	Jorge Rafael Gutiérrez Pulido
Sakbe	Claudia G. Gómez Santillán
	Laura Cruz Reyes
	Laura Gómez Cruz
IA & Educación	María Lucía Barrón Estrada
	J. Julieta Noguez Monroy
	María Yasmín Hernández Pérez
Deskubriendo Konocimiento	Alejandro Guerra Hernández
	Gildardo Sánchez Ante
Asistencia técnica	Alan G. Aguirre Lam
Corrección de estilo	Gilberto Rivera Zárate
	Marcela Quiroz Castellanos
	Sonia Lilia Mestizo Gutiérrez
	Guillermo de Jesús Hoyos Rivera
	Susana Aurora Velasco Montiel
	Héctor Joaquín Fraire Huacuja
	Leonardo Romero Muñoz
	Nelson Rangel Valdez
	Luis Miguel García Velázquez
Portada	María Fernanda Hernández Arroyo

Directores Fundadores
Carlos Alberto Reyes García
Ángel Kuri Morales

Comité Editorial
Ildar Batyrshin
Oscar Herrera Alcántara
Jesús Favela Vara
Sofía Natalia Galicia Haro
Miguel González Mendoza
Raúl Monroy Borja
Eduardo F. Morales Manzanares
Leonardo Garrido Luna
Carlos Alberto Reyes García
Antonio Sánchez Aguilar
Luis Enrique Sucar Succar
Ángel Kuri Morales
José A. Martínez Flores
Juan Manuel Ahuactzin Larios
Manuel Montes y Gómez
Ofelia Cervantes Villagómez
Alexander Gelbukh
Grigori Sidorov
Laura Cruz Reyes
Ramon Brena Pinero
Juan Humberto Sossa Azuela
Félix Castro

Árbitros
Héctor Héctor Acosta-Mesa
Rocío Barrientos-Martínez
Hiram Calvo
Antonio Camarena-Ibarrola
Karina Figueroa Mora
Juan Flores
Marisol Flores-Garrido
Luis Miguel García-Velázquez
Miguel Gonzalez-Mendoza
Guillermo Hoyos-Rivera
Adriana López-Lobato
Aldo Marquez-Grajales
Lourdes Martinez-Villaseñor
Manuel Nandayapa
Humberto Ochoa
Angel Ortiz
Nora Reyes
Gustavo Vargas Hakim
Osslan Vergara Villegas
Juan Villegas Cortéz

Contenido

Inteligencia artificial explicable en tareas de procesamiento de texto

Pág. 8 ⇒Rolando Menchaca Méndez, José Ángel Avelar Barragán y Grigori Sidorov

Trampas éticas asociadas al muestreo y aumento de datos en el aprendizaje automático
Pág. 13 ⇒ Jesús González Godoy, Tonantzin Guerrero Velázquez y Juan Humberto Sossa Azuela

ChatGPT: Ética y desafíos en la tecnología de procesamiento de lenguaje natural
Pág. 17 ⇒ Ansel Y. Rodríguez-González y Humberto Pérez-Espinosa

Ética en el uso de inteligencia artificial en aplicaciones donde intervienen animales
Pág. 23 ⇒ Carlos Alberto Aguilar-Lazcano, Humberto Pérez-Espinosa, Verónica Reyes-Meza y Mariel Urbina- Escalante

Archivística e Inteligencia Artificial: posibilidades interdisciplinares y desarrollos responsables
Pág. 30 ⇒ Luis Miguel García Velázquez y Marisol Flores Garrido

Aplicaciones y Desafíos de la Inteligencia Artificial en la Aeronáutica
Pág. 35 ⇒ Romina Ruíz Macías, Rodolfo García-Rodríguez y Luis Enrique Ramos-Velasco

Inteligencia Artificial para el diagnóstico primario de dependencia funcional en personas adultas mayores
Pág. 41 ⇒ Santiago Arceo Díaz, Elena Elsa Bricio Barrios, Jaime Alberto Bricio Barrios, Xóchitl Trujillo y Miguel Huerta

La necesidad de la ética ante la expansión de la IA
Pág. 47 ⇒ Fabio Morandín-Ahuerma

Columnas

Sapiens Piensa. Editorial pág. 2

e-Tlakuilo pág. 3

Estado del IArte pág. 5

Sakbe pág. 6

IA & Educación pág. 51

Deskubriendo Konocimiento pág. 52

Sapiens Piensa

Lourdes Martínez Villaseñor y Juan Humberto Sossa Azuela

Este número especial está enfocado en abordar los retos éticos que conlleva el uso de la Inteligencia Artificial (IA) en diferentes dominios. Contiene ejemplos de sistemas inteligentes en procesamiento de texto y modelos de lenguaje, entre otras aplicaciones diversas. Asimismo, encontramos un artículo, como el de Godoy, Guerrero y Sossa, que advierte de las trampas éticas asociadas al muestreo y aumento de datos en el aprendizaje automático. Los autores explican cómo, al entrenar un modelo, se recurre al muestreo y al aumento de datos para lidiar con el problema de tener demasiados datos, datos no representativos o muy pocos datos.

Menchaca, Avelar y Sidorov resaltan la importancia de la IA explicable (XAI) para abordar los retos éticos, en particular en las tareas de procesamiento de texto en aplicaciones de generación automática e interactiva de textos. El uso de tecnologías y metodologías para generar explicaciones comprensibles permitirá entender las decisiones de modelos opacos y lidiar con problemas como los sesgos.

Rodríguez-González y Pérez-Espinosa describen lo que es ChatGPT: un modelo de lenguaje de IA, y reflexionan sobre su uso, alcances, limitaciones e implicaciones éticas. Añaden también recomendaciones para un uso responsable de la tecnología.

Aguilar-Lazcano et al. nos presentan un caso de estudio muy original. Se trata de un sistema de interfaz perro-computadora que permite utilizar las vocalizaciones del animal para interactuar con una computadora con diferentes propósitos. Los autores resaltan la importancia de cuidar los aspectos éticos en los sistemas que incluyen datos de animales y sus diferencias con los sistemas inteligentes que usan datos de personas humanas.

García y Flores presentan un caso de aplicación de la ética profesional en InterPARES Trust; se trata de una estrategia archivística inteligente de expedientes médicos. En este trabajo, los autores muestran cómo la exploración crítica de datos ayuda a entender los datos y documentar los criterios de decisión para aumentar la transparencia del análisis y mejorar la interpretabilidad de los resultados de manera responsable.

Ruíz Macías et al. destacan que los sistemas inteligentes, en especial los aplicados a la aeronáutica, deben ser caracterizados por su viabilidad y confiabilidad, minimizando sus riesgos en tareas críticas. En este trabajo, los autores comentan la necesidad de buscar un equilibrio entre la transparencia, privacidad y protección, así como la implementación de mecanismos de supervisión y auditoría que garanticen una operación responsable de los sistemas aeronáuticos.

En otro orden de ideas, Arceo et al. indican cómo es posible utilizar la teoría de conjuntos y árboles de decisión para el desarrollo de un sistema de diagnóstico de dependencia funcional basado en la prueba de Barthel. La técnica de árboles de decisión permite clasificar a los pacientes según su grado de dependencia de manera que sea fácilmente interpretable.

Finalmente, el trabajo de Morandín-Ahuerma nos hace reflexionar sobre la necesidad de la ética ante la expansión de la IA. Este trabajo repasa los valores de los códigos normativos y revisa su potencial para contribuir al uso responsable de la IA. Concluye que se requiere un mayor análisis de los problemas prácticos y nos recuerda que no debería desarrollarse ninguna tecnología sin considerar cómo garantizar que se mantenga en línea con los valores y el bienestar humanos.

En resumen, este número de Komputer Sapiens pretender abonar a la reflexión en torno a la ética de la IA y hace un llamado para trabajar de manera interdisciplinar e intersectorial para garantizar la imparcialidad de los datos, supervisar y evaluar los sistemas inteligentes en todo su ciclo de vida, mejorar la transparencia y explicabilidad, y mitigar los sesgos algorítmicos y los riesgos de privacidad.

Estimados colegas, les invitamos a echar un vistazo a los trabajos de este número especial, no sin antes agradecer a los autores por su tiempo y dedicación en preparar las antes resumidas contribuciones.✷



Juan Humberto Sossa Azuela es doctor en Informática por el Instituto Nacional Politécnico de Grenoble, Francia. Es profesor de tiempo completo del Instituto Politécnico Nacional y Jefe del Laboratorio de Robótica y Mecatrónica del Centro de Investigación en Computación. Es miembro Emérito del Sistema Nacional de Investigadores. Es miembro de la Academia Mexicana de Ciencias y miembro de la Academia de Ingeniería. Es también miembro Senior del Instituto de Ingenieros Eléctricos y Electrónicos (IEEE) y de la Sociedad Internacional de Redes Neuronales (INNS) y Fellow de la Sociedad Mexicana de Inteligencia Artificial. En 2021 fue galardonado con el Premio Nacional de Computación por parte de la Academia Mexicana de la Computación (AMEXCOMP), así mismo con el Premio Nacional de la Academia de Ciencias de Cuba 2021 en el área de las Ciencias Naturales y Exactas. Es autor de 5 libros de texto, 13 patentes, 24 derechos de autor y más de 450 trabajos de congreso y revista. Ha impartido más de 450 pláticas por invitación. Sus áreas de investigación son en Inteligencia Artificial, Aprendizaje para Máquinas, Robótica y Metaversos.



Lourdes Martínez Villaseñor es Ingeniera en Sistemas Computacionales, Maestría en Educación Familiar por la Universidad Panamericana y Doctora en Ciencias Computacionales del Tecnológico de Monterrey, México. Es profesora investigadora en la Facultad de Ingeniería de la Universidad Panamericana y jefa académica en posgrados de la Maestría en Ciencia de datos. Tiene la distinción nivel 1 del Sistema Nacional de Investigadores del CONACYT. Autora de más de 50 publicaciones en revistas y conferencias internacionales, capítulos de libros y 5 libros en el área de inteligencia artificial. Es vicepresidente de la Sociedad Mexicana de Inteligencia Artificial. Sus principales intereses de investigación son la inteligencia artificial aplicada a la salud y la ética para la inteligencia artificial.

© 2023 - Sociedad Mexicana de Inteligencia Artificial ISSN 2007-0691

e-Tlakuilo: Cartas de nuestros lectores

Karina Figueroa y Rosa María Valdovinos
etlakuilo@komputersapiens.org

En *Komputer Sapiens* nos hemos esforzado por estar "a sólo un click de distancia" a través de diferentes medios como Facebook, Twitter y correo electrónico. Les presentamos algunas de las preguntas que hemos recibido a través de estos medios.

La transición energética, las redes inteligentes y la inteligencia artificial.

Brandón Alexis Ballesteros.

¿Son realmente las nuevas energías capaces de no emitir ningún tipo de contaminante para el medio ambiente? ¿Qué tan costoso podría ser el cambio de las energías actuales a las nuevas?

Las nuevas energías NO son capaces de emitir NINGÚN tipo de contaminante al medio ambiente por una cuestión fundamental: la segunda ley de la termodinámica que plantea que la energía no se crea ni se destruye, sino se transforma. Por esta situación, la energía que llega a nuestras casas, sea cual sea la forma en que se gestiona, emite, cuando menos, calor a la atmósfera y ese calor no se irradia de manera total al espacio, por lo tanto, siempre queda algo. Lo que SÍ es un hecho, es que las nuevas fuentes de energía, como la solar, disminuyen de una manera considerable la emisión de contaminantes a la atmósfera. El uso de nuevas tecnologías, o el uso de tecnologías inteligentes NO está orientado a eliminar la emisión de contaminantes, sino hacer que estas emisiones sean manejables por la naturaleza y que no provoquen un problema ambiental. Esto sí es perfectamente viable, por ejemplo, al tener una central termoeléctrica que quema petróleo para calentar agua que produce vapor que termina moviendo una turbina la que produce energía eléctrica (aquí se puede ver como un recurso natural no renovable se convierte en energía térmica, liberando desechos, para luego convertirse en energía eléctrica liberando calor en el proceso). Si esta energía eléctrica se administra de manera eficiente, eficaz e inteligentemente (aquí entra el aporte de nuestro trabajo: la administración inteligente de la distribución y el consumo), aún sería necesario quemar petróleo para producir electricidad, pero sin embargo, la cantidad de gases contaminantes se podría disminuir a niveles que la propia atmósfera sea capaz de manejar. De eso se trata.

Sin embargo, se podría decir que generar electricidad con paneles solares elimina la generación de gases contaminantes y es cierto. El problema aquí está en la eficiencia energética de estos paneles, que ronda alrededor del 15 % (esto quiere decir que solo el 15 % de toda la luz solar que incide sobre ellos se convierte en energía). Sin embargo, el calor que estos dispositivos pueden llegar a irradiar a la atmósfera podría traer problemas considerables si se piensa en producción a gran escala de energía a partir de paneles solares. Por ejemplo, un aumento de la temperatura provocado por grandes granjas de paneles solares podría producir un aumento de la temperatura media del planeta, un aumento de la humedad, un aumento de las lluvias, un aumento del cambio de las corrientes de vientos, un aumento de los ciclones, un aumento de la evaporación de los océanos. La moraleja del cuento es que, desafortunadamente, todo se encuentra entrelazado sistémicamente y un cambio en alguna parte del sistema provoca cambios en el resto.

Como conclusión: no se debe pensar ni esperar en eliminar al 100 la producción de energía a partir de fuentes no renovables, sino hacer la producción de energía ambientalmente viable mediante el uso de todas las fuentes que tenemos al alcance, en especial, potenciando un (mucho) mayor uso de fuentes renovables de energía, realizando un consumo CONSIENTE (que el capitalismo potencia el consumo irracional e indiscriminado de todo lo que puede ser consumido) de portadores energéticos.

Ahora me queda la impresión de que con mi respuesta estoy diciendo que ABSOLUTAMENTE la solución está en generar energía a partir de fuentes contaminantes y que NO hay otra solución. Y este no es el sentido de mi respuesta, sino que parte de la solución (la menor diría yo) está en producir el 90 por ciento de la energía con fuentes renovables y el 10 por ciento por fuentes no renovables. La mayor parte de la solución está en formar patrones de uso y consumo racionales y razonables.

El impacto del cómputo evolutivo y las meta-heurísticas en la industria y la sustentabilidad.

Jorge Isacc Talita Hernández.

¿Además de FPP existen otros tipos de sistemas de organización que pueda utilizar? ¿Por qué eligieron el Algoritmo Genético para su investigación existiendo otros tipos de algoritmos que ustedes podrían utilizar?

En principio se usaron los FPP debido al tema de investigación del estudiante de posgrado, pero en realidad debido a la flexibilidad de las técnicas de optimización

© 2023 - Sociedad Mexicana de Inteligencia Artificial ISSN 2007-0691

pueden ser empleados en prácticamente cualquier sistema mientras se defina correctamente la función objetivo. Se empleó el GA debido a la naturaleza del problema y a la forma en que opera el GA, considerando como se definen las partículas, queriendo demostrar que normalmente las aplicaciones se desarrollan desde una perspectiva ajena al problema en sí.

La minería de datos como un camino para lograr un futuro sustentable. Pregunta: ¿Me podrías explicar con un ejemplo cómo la minería de datos podría mejorar las condiciones de vida de la especies naturales?

La minería de datos es una herramienta más de la caja de herramientas de investigación de todas las áreas del saber. Al igual que los biólogos realizan estudios observacionales para conocer los pormenores de una especie, también se apoyan en la estadística para dar validez a sus hipótesis.
La minería de datos no sustituye en ningún momento la investigación in-situ, sino que por el contrario, la enriquece al encontrar patrones ocultos de interés, en las grandes colecciones de datos. Dicho enfoque se denomina buttom-up, pues a diferencia de la investigación tradicional (top-down) no inicia con una hipótesis que después se comprueba con el análisis de los datos. En su lugar, se hacen análisis exploratorios a los datos de interés disponibles en busca de patrones, los cuales al encontrarse en colecciones voluminosas pasan desapercibidos a otras técnicas. Dichos patrones indican rutas por las cuales conducir la investigación, o incluso hallazgos que no podrían haber sido identificados tan fácilmente por otros métodos. Para más información de los enfoques buttom-up y top-down, recomendamos al lector el artículo de Díaz-Pacheco y Aranda (2022). El tercer camino es, con notable evidencia, aquel que podemos esperar con mayor confianza. Sin embargo, es importante tener en cuenta que el diagnóstico médico es una práctica situada, por lo que difícilmente podríamos imaginar un futuro en el uso de estas tecnologías homogéneo en todos los contextos culturales y territoriales.
Tomando al ser humano como una instancia del conjunto de "especies naturales" podemos decir que la minería de datos ha permitido desarrollar técnicas que apoyan al diagnóstico médico en sustitución de análisis más invasivos y peligrosos como las biopsias. A través del análisis de información como estadísticas biológicas de los pacientes (colesterol, azúcar en sangre, etc.), radiografías y tomografías, con la minería de datos es posible realizar pronósticos sobre el desarrollo futuro enfermedad, permitiendo que diversos males puedan ser atacados a tiempo y aumentando las posibilidades de tratamiento.

Angel Díaz-Pacheco y Ramón Aranda. El turismo en la era de los macrodatos. ¿Cómo ves?, 289: 28–31, Dec 2022.

¿Podría también ser el tiempo que tarda un algoritmo de minería de datos un factor importante al momento de realizar una investigación?

Como se mencionó anteriormente, al usar minería de datos como parte de una investigación, se busca encontrar patrones ocultos de interés en grandes colecciones de datos. La cantidad de instancias de los datos (por ejemplo, animales, pacientes), el número de atributos que describen a las instancias, y la frecuencia y temporalidad del monitoreo impactará en el tiempo requerido para que un algoritmo devuelva una respuesta. Por ejemplo, los algoritmos de minado de patrones frecuentes tienen un tiempo de respuesta en el peor caso exponencial respecto al número de atributos de la colección de datos estudiada. En particular, para una colección de 10 atributos, cada uno con solo dos valores, el tiempo de respuesta de un algoritmo de minado de patrones frecuentes sería en el peor caso aproximadamente $2^{10} - 1 = 1023$ veces el tiempo de respuesta para misma colección pero con solo 1 atributo del mismo tipo. Cada algoritmo de minería de datos posee su propio tiempo de respuesta que depende del tamaño de la colección de datos. No todos tienen un tiempo de respuesta exponencial respecto al tamaño de la colección de datos, sin embargo, en su mayoría su tiempo de respuesta es cuadrático o cúbico (n^2 o n^3 donde n es el tamaño de la colección).
Si tenemos en cuenta que la cantidad de instancias puede estar en el orden de los millones, que el número de atributos pueden ser decenas o cientos o más y cada uno suele tener mucho más de 2 valores, y que los datos pueden adquirirse con cierta frecuencia y temporalidad entonces el tiempo respuesta puede superar al tiempo disponible para realizar la investigación.
Una consideración fundamental para obtener resultados en tiempos aceptables es acotar que tareas de minería de datos que se incluirán en la investigación y decidir qué algoritmos son adecuados para el tamaño de la colección explorada. En este sentido, se debe considerar adicionalmente cuál es la capacidad de cómputo que se posee, y no perder de vista el uso de algoritmos paralelos que hagan uso de múltiples núcleos de procesamiento y procesadores gráficos, así como de plataformas de cómputo en la nube.

© 2023 - Sociedad Mexicana de Inteligencia Artificial ISSN 2007-0691

Estado del IArte

Jorge Rafael Gutiérrez Pulido (@jrgpulido) y Reyna Carolina Medina Ramírez
estadoiarte@komputersapiens.org

Ética para la IA

Cuando aparecieron por primer vez las películas de ciencia ficción, en más de alguna, el protagonista secundario era algún robot. Eventualmente, ese protagonista secundario se ha convertido en primario. Hoy en día, la ficción se ha emparejando con la realidad. A diario leemos encabezados de nuevas inteligencias artificiales que desafían nuestra imaginación, realizando, de manera rápida y completa tareas que los seres humanos realizamos. Por ejemplo, desarrollar software o presentaciones temáticas asistidos por computadora, o crear fotos y videos con personas que ni siquiera existen, sin entrar en la discusión de si esas aplicaciones pueden ser usadas para bien o para mal. Destacan las sistemas conocidos como *generadores de respuestas*, de manera particular el Chatgpt de **OpenAI**. Pero no es el único, y no es ninguna sorpresa que la mayoría de los gigantes de las tecnologías de información tengan el suyo.

Desde hace algunos años la comunidad científica, preocupada por los aspectos relacionados con la ética, han levantado la voz para alertar acerca del riesgo que corremos si dejamos que las inteligencias artificiales sigan siendo desarrolladas sin una legislación que las regule. Un esfuerzo importante se ha dado en 2021 con los 193 estados miembros de la UNESCO aprobando el "*Marco Ético Mundial de la IA*"[1], iniciativa que debe considerarse seriamente. No abundaremos en las ventajas, conocidas por todos, principalmente quitarnos tareas engorrosas, mientras las partes más finas de la tarea las seguirán realizando los seres humanos. Estas tecnologías basadas en IA nos han sido de utilidad para combatir pandemias, así como para brindar terapia.

Algunas de las desventajas en las que mayor énfasis se ha puesto en los últimos meses son: el riesgo que presentan para la democracia, para la ciberseguridad de los países y de las personas. Quizás la nota más importante al respecto fue la que dió un reconocido miembro de la comunidad científica de la IA, Geoffrey Hinton, conocido como el padrino de la IA, quien hasta hace muy poco fungía como investigador principal para Google. Hinton ha renunciado a Google para dedicarse de lleno a concientizar a la comunidad científica de los riesgos que la IA representa si no es regulada. Para saber más, el lector puede ver la entrevista[2] que concedió a la MIT Technology Review's EmTech digital.

En el mismo sentido, Sam Altman, líder de desarrollo de OpenAI en una comparecencia ante el senado de EUA dijo que regular el desarrollo de la IA es crucial. En Mayo pasado, la vicepresidenta de EUA Kamala Harris, convocó a los presidentes de OpenAI, Microsoft y Alphabet para hablar de ese tema.

Por otro lado, importantes pensadores contemporaneos como Nick Bostrom, Seth Baum y Eliezer Yudkowsky han dedicado décadas a estudiar el problema conocido como de **control o alineamiento**. Éste lo explican de la siguiente forma: ¿puede un infante vencer al campeón mundial de ajedrez? De la misma manera ¿puede una IA producir en unos segundos lo que un grupo grande de desarrolladores produjo en un año? Cuando llegue el momento, no podremos presionar el botón de *apagar*, pues este sistema superinteligente habría pensado también en esa posibilidad, concluyen estos pensadores.

Por su parte un destacado grupo de la comunidad científica han firmado una carta abierta escrita por el instituto *Future of Life* pidiendo un receso de al menos 6 meses en el desarrollo de nuevas tecnologías basadas en IA. Esta carta[3] ha sido firmada por más de treinta y un mil personas al momento de escribir esta columna. Otros importantes miembros de la comunidad científica tales como Yoshua Bengio, Steve Wozniak, y Elon Musk han firmado la carta. Quizás Stephen Hawking y Alan Turing habrían firmado la carta, pues en su momento también expresaron su preocupación al respecto. Para conocer más acerca de los riesgos de la IA a la humanidad, el lector puede consultar el enlace[4].

Si la inteligencia artificial es el mejor amigo o la mayor amenaza de la humanidad, aún está por verse. El público en general está encantado con la infinita posibilidad de tareas que estas herramientas permiten realizar. Aunque existen puntos de vista opuestos, es muy importante reflexionar acerca de los beneficios y riesgos potenciales de la IA; teniendo en cuenta que las computadoras y los sistemas de IA están crecienco vertiginosamente en sus capacidades.✷

[1] https://www.unesco.org/es/articles/inteligencia-artificial-la-unesco-pide-los-gobiernos-que-apliquen-sin-demora-el-marco-etico-mundial
[2] https://youtu.be/sitHS6UDMJc
[3] https://futureoflife.org/open-letter/pause-giant-ai-experiments
[4] https://www.datanami.com/2023/06/01/ai-researchers-issue-warning-treat-ai-risks-as-global-priority

© 2023 - Sociedad Mexicana de Inteligencia Artificial ISSN 2007-0691

Sakbe

Laura Cruz Reyes, Laura Gómez Cruz y Claudia Guadalupe Gómez Santillán
sakbe@komputersapiens.org

Caminos Responsables: Hacia un Futuro Ético en la IA

La ética en la aplicación de la inteligencia artificial (IA) aborda los aspectos morales y éticos del desarrollo, implementación y uso de sistemas basados en IA. A medida que la IA transforma rápidamente nuestras vidas, surgen preguntas éticas fundamentales. En esta columna, exploraremos desafíos y esfuerzos para fomentar un desarrollo y uso responsable de esta tecnología. Empecemos con los desafíos.

- **Transparencia y explicabilidad:** Es necesario que las aplicaciones basadas en IA sean abiertas y visibles para que los usuarios comprendan cómo se toman las decisiones en los algoritmos. En la industria logística, Transmetrics analiza, modela y predice los flujos de transporte considerando diversos factores como el clima, el tráfico, el combustible y las condiciones de la carretera, para planificar rutas eficientes. Aunque esta tarea es más difícil para un planificador humano que para una aplicación que utilice técnicas de IA, el planificador todavía necesita comprender y ajustar las sugerencias del sistema. `https://www.techtarget.com/searchenterpriseai/feature/Why-transparency-in-AI-matters-for-businesses`.
- **Sesgos y discriminación:** Los sesgos en los datos utilizados para entrenar los modelos de aprendizaje pueden generar resultados injustos y discriminatorios. Por ejemplo, el traductor de Google a veces replica estereotipos de género al confundir pronombres al traducir desde un idioma neutro de género al español. Esto puede causar perjuicios mayores, como en los algoritmos de detección de daños renales, que tienen el doble de probabilidades de equivocarse en mujeres debido a que fueron entrenados principalmente con datos de hombres. `https://imco.org.mx/la-inteligencia-artificial-tiene-sesgos/`.
- **Privacidad y seguridad de los datos:** Las aplicaciones que utilizan técnicas de IA requieren incluir mecanismos de protección de datos personales para evitar malos usos. En el caso de Samsung, algunos empleados usaron inadvertidamente ChatGPT, filtrando datos sensibles de la empresa, como el código fuente de una aplicación propietaria y patrones de identificación de chips defectuosos. Esta situación ha alertado a varias compañías a restringir o prohibir el uso del chatbot. `https://hipertextual.com/2023/04/chatgpt-empleados-samsung-filtran-informacion-confidencial`.
- **Rendición de cuentas:** Conforme las aplicaciones que utilizan IA se vuelven más autónoma, es crucial establecer responsabilidades para garantizar un uso ético y abordar posibles errores o daños. Un abogado de la Universidad de Pekín realizó un estudio sobre los mecanismos de rendición de cuentas para sistemas de IA en países de América Latina. Se observó un desarrollo desigual entre los países analizados, resaltando a Uruguay como el único con mecanismos legales, institucionales y técnicos establecidos, como leyes, organismos de monitoreo y guías de evaluación. `https://proyectoguia.lat/wp-content/uploads/2022/04/Rendicion-de-cuentas.pdf`.
- **Impacto social y económico:** La IA aporta beneficios en la optimización de procesos y la reducción del error humano, pero también puede generar perjuicios, como la transformación radical de la sociedad. Según el filósofo Byung-Chul Han, la tecnología digital ha provocado adicción al metaverso, la hiperconexión constante entre el mundo físico y digital, y la explotación comercial de la vida humana. Las personas adquieren perfiles predecibles y manipulables por las aplicaciones que usan técnicas de IA a través de dispositivos inteligentes. `https://revistas.ucv.es/scio/index.php/scio/article/view/1102/1104`

Buscando garantizar el desarrollo ético de la IA en armonía con la sociedad, han surgido nuevas organizaciones o iniciativas de organizaciones existentes, como las siguientes:

© 2023 - Sociedad Mexicana de Inteligencia Artificial ISSN 2007-0691

UNESCO ⇒ Ha publicado el primer estándar mundial sobre ética de la IA, que recomienda a los formuladores de políticas traducir los valores y principios fundamentales en acciones con respecto a la gobernanza de datos, el medio ambiente y los ecosistemas, entre otros.

https://www.unesco.org/en/artificial-intelligence/recommendation-ethics

Partnership on AI ⇒ Es una alianza de organizaciones y empresas sin fines de lucro que se dedican a investigar y promover prácticas éticas en la IA. Entre sus miembros notables se encuentran empresas como Google, Microsoft, Facebook y organizaciones sin fines de lucro como Amnistía Internacional.

https://partnershiponai.org/

Future of Life Institute ⇒ Una organización que se enfoca en el desarrollo seguro y ético de la IA. Su trabajo abarca desde la investigación y la promoción de pautas éticas hasta la colaboración internacional y la difusión de información sobre los desafíos y las implicaciones de la IA.

https://futureoflife.org/

The Institute for Ethical AI & Machine Learning ⇒ Un instituto que se dedica a promover la ética y la responsabilidad en la IA y el aprendizaje automático. Ofrece recursos, talleres y capacitación sobre ética en la IA.

https://ethical.institute/

IEEE ⇒ Uno de sus programas es la iniciativa global del IEEE sobre la ética de los sistemas autónomos e inteligentes. Han identificado más de ciento veinte problemas clave y proporcionado recomendaciones para abordarlos.

https://standards.ieee.org/industry-connections/ec/autonomous-systems/

Future of Humanity Institute ⇒ Es un instituto de la Universidad de Oxford (Reino Unido) que realiza investigaciones interdisciplinarias sobre los desafíos éticos y de seguridad de tecnologías emergente como la IA y aboga por un enfoque responsable en su desarrollo. El FHI reúne a expertos de diversas disciplinas, como matemáticas, filosofía, economía, ciencia política y ciencias de la computación.

https://www.fhi.ox.ac.uk/

Estas organizaciones y programas son solo algunos ejemplos de los esfuerzos en curso para abordar la ética en la IA. Cada una de ellas busca contribuir de manera única a la promoción de prácticas éticas y al fomento de un desarrollo responsable de la IA.

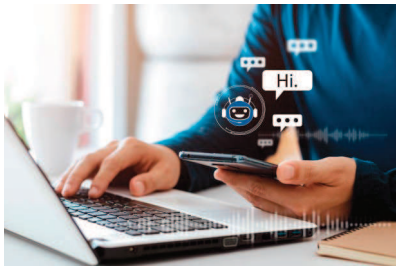


Imagen gratuita (licencia Creative Commons) de Thapana Onphalai vía istockphoto.com.

© 2023 - Sociedad Mexicana de Inteligencia Artificial ISSN 2007-0691

ARTÍCULO ACEPTADO

Inteligencia artificial explicable en tareas de procesamiento de texto

Rolando Menchaca Méndez, José Ángel Avelar Barragán y Grigori Sidorov

Importancia de la IA explicable

Vivimos una revolución tecnológica en la que los algoritmos de inteligencia artificial (IA) están permeando todos los aspectos de la vida cotidiana. Desde aplicaciones industriales, hasta lúdicas, pasando por científicas, financieras y médicas, podemos encontrar ejemplos donde algoritmos y modelos de inteligencia artificial, comúnmente llamadas inteligencias artificiales (IAs), impactan de manera positiva a individuos, industrias y gobiernos por medio de la automatización y optimización de procesos, así como apoyando a la toma de decisiones a partir del análisis avanzado de datos.

Gracias a la gran eficacia que han demostrado las IAs en los ámbitos antes mencionados, su uso se está extendiendo a actividades cada vez más importantes, por lo que se espera que su impacto en la vida de las personas sea cada vez mayor. Sin menoscabo de los efectos positivos de esta tecnología, es importante señalar que su utilización generalizada también trae consigo *riesgos y consideraciones éticas*. Una de las razones de esto, es que existen contextos donde las decisiones tomadas con la asistencia de IAs también pueden afectar de manera negativa la vida de las personas. Esto se debe principalmente a que los algoritmos que las implementan pueden exhibir sesgos indeseados y a que en muchos casos es técnicamente difícil detectar dichos sesgos.

Por un lado, los algoritmos y modelos de aprendizaje automático (AA) que se utilizan para implementar a las IAs están diseñados para que su comportamiento refleje los patrones contenidos en los datos con los que fueron entrenados. Es decir, las IAs están diseñadas para reproducir los mismos comportamientos y tendencias que dieron origen a los datos de entrenamiento. En este contexto, en el caso de que los datos de entrenamiento contengan sesgos, lo más probable es que la IA que se entrenó con esos datos reproduzca y tal vez amplifique dichos sesgos durante su operación. A manera de ejemplo, tomemos un modelo de AA que se utilice para determinar la pertinencia de otorgar un crédito, y supongamos que los datos históricos utilizados durante su entrenamiento contienen ejemplos de discriminación a personas pertenecientes a algún grupo social donde injustamente se les negó el acceso a créditos. En este caso, el comportamiento de la IA exhibirá el mismo sesgo negativo que ocurrió en el pasado, reproduciendo en el presente esos mismos actos discriminatorios.

En el área de las tareas de procesamiento de texto, una aplicación cada vez más popular son los ChatBots, como ChatGPT de OpenAI o Bard de Google, para la generación automática e interactiva de texto. En este caso, también existe el riesgo de que los datos empleados para entrenar a las IAs contengan sesgos que se reflejen, posiblemente de manera muy sutil, durante la operación del ChatBot. Un ejemplo concreto está reportando en [1], donde el conjunto de entrenamiento presentaba un sesgo que asociaba la tez negra con términos desagradables en comparación con la tez blanca que estaba asociada con términos positivos.

Como se mencionó, una agravante al problema anterior es que en muchos casos es difícil determinar si el comportamiento de una IA efectivamente contiene sesgos que injustamente beneficien o perjudiquen a un grupo de personas. Una de las razones detrás de esta dificultad es que una buena proporción de los modelos de AA más ampliamente usados, como las redes neuronales en sus múltiples arquitecturas y variantes, pueden ser considerados como *modelos opacos* (muchas veces llamados modelos de caja negra). Decimos que un modelo de AA es "opaco" cuando conocemos el algoritmo que lo implementa y su fundamento matemático, pero no es posible determinar las razones concretas y específicas que sustentan cada uno de los resultados que devuelve. Este problema se ha hecho cada vez más complicado debido a que existe una tendencia de usar modelos cada vez más complejos, con una mayor cantidad de parámetros involucrados en su operación, y donde es cada vez más difícil de entender las razones detrás de sus predicciones.

Con este tipo de IAs basadas en modelos opacos, en la práctica la responsabilidad de su uso correcto recae principalmente en la empresa o institución que lo pone en operación. Desafortunadamente, resulta complicado hacer responsable a dicha entidad de una injusticia generada por el uso de la IA debido a que no se conocen las razones detrás de la decisión tomadas por dicha IA.

Una alternativa para abordar los retos éticos y minimizar los riesgos inducidos por el uso de los modelos de AA opacos es conocida como *Inteligencia Artificial Explicable (IAX)*, la cual consiste de un conjunto de algoritmos, tecnologías y metodologías que pueden ser utilizados para generar explicaciones comprensibles por los humanos acerca de las decisiones tomadas por IAs basadas en modelos opacos. La Figura 1 ilustra de manera

© 2023 - Sociedad Mexicana de Inteligencia Artificial ISSN 2007-0691

sencilla la función de la IAX. Mientras que un modelo opaco simplemente entrega su predicción, un algoritmo explicador realiza post-procesamiento para obtener una explicación del porqué de esa salida. En el caso del ejemplo de la figura, el algoritmo clasificador determinó que un texto de entrada describe a un gato, mientras que el explicador entrega como salida cuáles fueron las características encontradas en el texto que hicieron que el clasificador llegara a esa conclusión. En este caso, porque en el texto se menciona que es un animal pequeño, con orejas picudas, bigotes y garras.

Las IAs tienden a reproducir los sesgos contenidos en sus datos de entrenamiento.

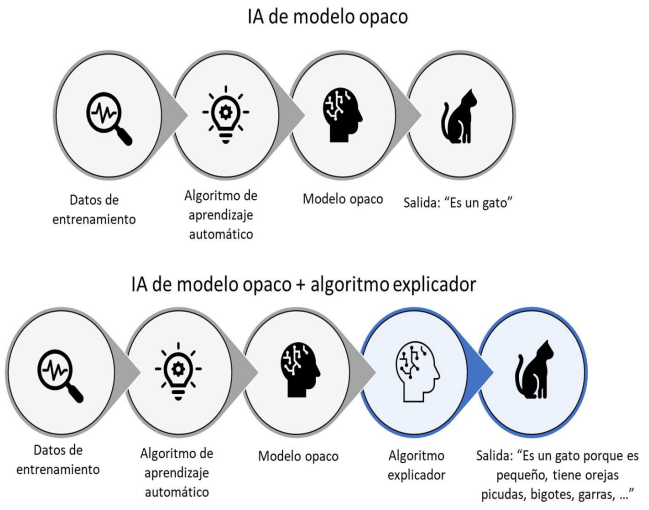


Figura 1. IA de modelo opaco vs. IA de modelo opaco + algoritmo explicador.

El propósito de la IAX es que a partir de la comprensión de los resultados generados por la IA, los humanos puedan generar confianza en su correcto funcionamiento, o en su caso, detectar posibles sesgos y detonar acciones correctivas. Esto último debido a que a partir de las explicaciones, se hace posible que los afectados por una decisión tomada por una IA tengan elementos sólidos para realizar un reclamo o interponer una impugnación ante la autoridad correspondiente.

Adicionalmente, las explicaciones provistas por la IAX pueden dar elementos para evaluar la imparcialidad de la IA, así como para verificar que se está cumpliendo con la normatividad aplicable. En este sentido, la IAX provee de elementos para que las organizaciones e instituciones se aseguren que su uso de las IAs sigue un enfoque *ético*, humano y responsable.

De hecho, la pertinencia y necesidad de generar explicaciones sobre el funcionamiento de los modelos opacos ha llevado a realizar acciones como las del Parlamento Europeo, donde se implementó la "Regulación General de Protección de datos" (*General Data Protection Regulation*, GDPR) [2] donde se aboga por los derechos de las personas de las cuales se han tomado sus datos y están siendo usados para tomar decisiones respecto ellas. Una discusión sobre la existencia, posibilidad y alcance de dichos derechos se profundiza en [3].

En las siguientes secciones se presenta un marco de trabajo para generar explicaciones acerca del funcionamiento de clasificadores automáticos de texto basados en modelos opacos.

Clasificación automática de textos y algoritmos explicadores

La clasificación automática de textos consiste en determinar a qué clase (o categoría) pertenece un documento de entrada. Las clases generalmente corresponden a temáticas, como podrían ser "Deportes" o "Política" pero también es posible clasificar documentos por su autor o por el género literario al que pertenecen (e.g., narrativa, lírica, dramática). Así, un clasificador automático de textos toma como entrada un documento, y su objetivo es determinar a qué categoría de un conjunto predeterminado de categorías pertenece dicho documento. Este problema ha sido abordado mediante diferentes técnicas, entre ellas, la utilización de algoritmos de AA supervisado como las redes neuronales.

En el caso del AA supervisado, un algoritmo de entrenamiento toma como entrada un conjunto de pares *(documento, categoría)*, típicamente obtenidas de un *Corpus* de entrenamiento, que sirven como ejemplo de la forma correcta en que los documentos deben ser clasificados. La Figura 2 ilustra de manera un poco más detallada los procesos de entrenamiento y clasificación (predicción) de textos para un modelo de AA supervisado. En la figura puede observarse que ambos procesos incluyen una fase de pre-procesamiento, la cual tiene como objetivo convertir los documentos a vectores numéricos (en términos técnicos, mapearlos a un Modelo de Espacio Vectorial, MEV) donde cada componente del vector corresponde a una palabra. Estos vectores tienen tantas componentes como palabras diferentes existen en el Corpus y a cada una de sus componentes se le denomina como *característica*.

El algoritmo de entrenamiento realiza un proceso de ajuste de los parámetros del modelo de AA de forma tal que el error en la clasificación de los documentos que

© 2023 - Sociedad Mexicana de Inteligencia Artificial ISSN 2007-0691

componen al conjunto de entrenamiento se minimice. De esta manera, en el modelo quedarán codificados los patrones que caracterizan a cada una de las clases a las que corresponden los documentos que pertenecen al Corpus de entrenamiento.

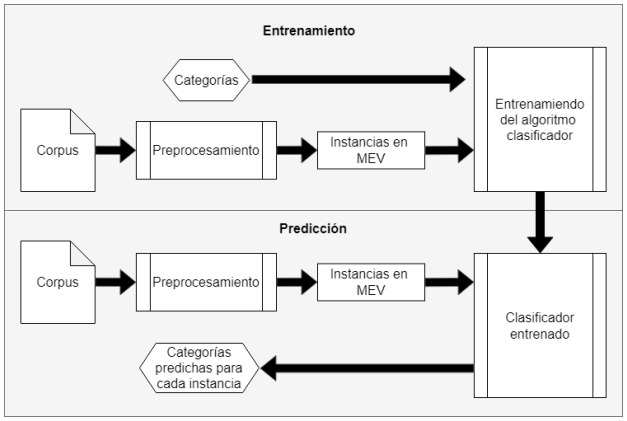


Figura 2. Proceso de entrenamiento/predicción con un clasificador supervisado.

Es importante resaltar que los algoritmos de entrenamiento tradicionales no incorporan ninguna noción sobre si dichas tendencias o patrones son "justos", o si por el contrario reflejan sesgos negativos que se encontraban presentes en el Corpus de entrenamiento. En el caso de las redes neuronales, el modelo entrenado se considera opaco porque está codificado como un conjunto de valores numéricos (*pesos* en la terminología de redes neuronales) que no pueden ser usados directamente por los humanos para determinar las razones que están detrás del comportamiento del algoritmo clasificador. En este caso, si se desea saber qué determinó cada una de las salidas generadas por el clasificador, será necesario utilizar un *algoritmo explicador* que nos provea de elementos que puedan ser interpretados por los humanos y que justifiquen cada una de esas salidas.

A este tipo de algoritmos explicadores que se utilizan para entender el comportamiento de un modelo opaco se le conoce como *explicador post-hoc*. Por otro lado, cuando es posible obtener explicaciones directamente del modelo de AA decimos que se trata de un *explicador intrínseco*, como por ejemplo, los árboles de decisión cuyo modelo está basado en reglas de decisión explícitas que pueden ser interpretadas por una persona [4][5].

Explicabilidad por medio de modelos sustitutos

Una metodología para obtener explicaciones sobre el comportamiento de un modelo opaco es generar un *modelo sustituto interpretable* que se entrena para emular el comportamiento de dicho modelo opaco. El modelo sustituto debe generar, en la medida de lo posible, la misma predicción que el modelo opaco pero también debe proveer elementos interpretables que justifiquen dicha predicción y que sean entendibles por las personas.

Dado que en el caso de la clasificación de textos las características corresponden a las palabras que componen a los documentos, el modelo sustituto indicará cuáles de las palabras presentes en el documento fueron las más determinantes en la decisión que tomó el clasificador. En la literatura, se conoce a este tipo de algoritmos explicadores como basados en la *importancia de las características*.

En la sección siguiente se muestran los resultados de experimentos donde se obtienen explicaciones para las predicciones generadas por una arquitectura de red neuronal sencilla. Se consideran dos algoritmos explicadores *post-hoc*. LIME (*Local Interpretable Model-agnostic Explanations*) [6], que toma como entrada al modelo opaco y genera modelos lineales que indican qué características fueron las más importantes y SHAP (*SHapley Additive exPlanations*) [7], que genera un modelo lineal sustituto por medio de técnicas de teoría de juegos.

¿Cómo es una explicación?

ara mostrar el funcionamiento de los algoritmos explicadores se empleó el Corpus conocido como *20 Newsgrups* que consiste de un conjunto de 18,828 documentos que contienen noticias clasificadas en 20 temas distintos. Se eligió este conjunto de datos debido a que es ampliamente utilizado para experimentos relacionados con la clasificación automática de texto.

Para generar el modelo opaco entrenado se realizó un proceso conocido como de *validación cruzada*, cuyo objetivo es encontrar los hiperparámetros del modelo opaco y de su algoritmo de entrenamiento que generen el modelo con el mejor desempeño posible. Como resultado se obtuvo un modelo de perceptrón multicapa (MLP) que el 90 % de las veces predice de manera correcta el tema de una noticia.

Una vez que se tiene el clasificador entrenado, se procedió a utilizar los algoritmos explicadores SHAP y LIME para generar los modelos sustitutos que sirven para determinar cuáles fueron las características más importantes para sus predicciones. Este proceso se realiza por cada predicción que se desea explicar, es decir, para cada documento de entrada para el cual se requiera generar una explicación. Para ejemplificar este proceso, la Figura 3 muestra un documento de entrada cuya clase verdadera es "Gráficas por computadora" y que el MLP predijo correctamente como "Gráficas por computadora". Como hemos argumentado, debido a que el MLP es un modelo opaco, su salida no provee ninguna información adicional. Por su parte, la Figura 4 muestra las explicaciones generadas por SHAP y LIME donde por medio de un mapa de calor se resaltan las palabras que cada uno de los algoritmos explicadores determinaron que son las más

importantes para el MLP. En el mapa de calor, mientras más intenso es el color con que se resalta una palabra, más importante fue dicha palabra para el clasificador. Este resaltado provee una explicación visual que le permite a una persona identificar con facilidad los elementos que determinaron la salida del clasificador de modelo opaco.

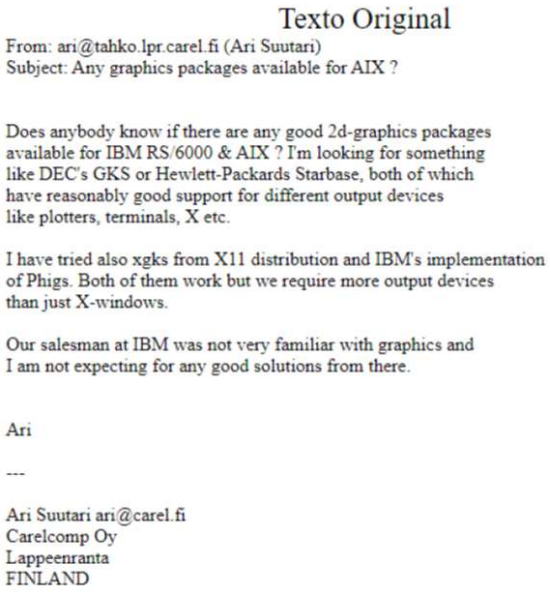


Figura 3. Texto original.

Un punto interesante a resaltar de las explicaciones mostradas en la Figura 4 es que ambos explicadores concluyeron que, ya sea el nombre del autor de la noticia para el caso de LIME, o su dirección de correo electrónico para el caso de SHAP, son importantes para que el MLP determine el tema del documento. Esto podría indicar que a pesar de que el MLP tiene una precisión alta (90 %), es posible que el modelo contenga un sesgo que reducirá su efectividad cuando se quiera utilizar para clasificar documentos que no pertenezcan al Corpus usado para su entrenamiento. Por otro lado, es importante señalar que aún y cuando las palabras resaltadas por ambos explicadores son en principio relevantes para su clasificación, éstas no coinciden entre sí totalmente. Lo anterior se debe a que ninguno de los dos algoritmos explicadores es exacto, y por lo tanto, a que sus explicaciones son aproximadas. De hecho, el problema de diseñar algoritmos explicadores sigue siendo un tema de investigación abierto.

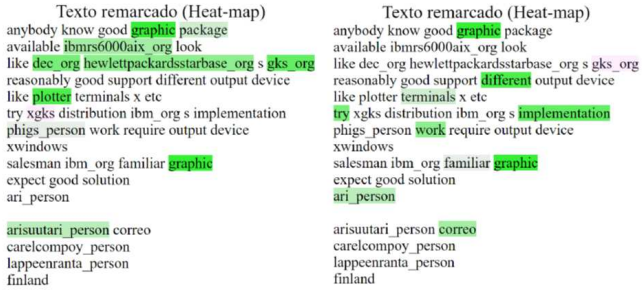


Figura 4. Explicaiones generadas por los explicadores LIME (izquierda) y SHAP (derecha) para el clasificador MLP.

Reflexiones finales

Nos encontramos al inicio de una nueva era tecnológica en la que algoritmos de inteligencia artificial serán parte medular de todos los aspectos de la vida diaria. Se espera que el impacto benéfico de estas tecnologías sea cada vez mayor, dado su potencial de mejorar la calidad de vida de las personas y de hacer a las industrias, servicios y gobiernos más efectivos y eficientes.

En este contexto, se hace cada vez más importante desarrollar tecnología que nos permita asegurarnos que no se haga mal uso de las IAs, y que en general su utilización promueva valores éticos como la justicia y la equidad. Un paso en esa dirección es la Inteligencia Artificial Explicable, cuyo propósito es generar explicaciones que provean elementos para que las personas puedan entender y eventualmente confiar en las IAs. ✷

REFERENCIAS

1. Caliskan, A., Bryson, J. y Narayanan, A. (2017). Semantics derived automatically from language corpora contain human-like biases. *Science*, 356(6334), 183-186.

2. Voigt, P. y Von dem Bussche, A. (2017). The EU general data protection regulation (GPDR). A Practical Guide. *Cham: Springer International Publishing.*

3. Goddard, M. (2017). The EU General Data Protection Regulation (GDPR): European regulation that has a global impact. *International Journal of Market Research*, 59(6), 703-705.

4. Molnar, C. (2020). Interpretable Machine Learning: A Guide for Making Black Box Models Explainable. Recuperado el 3 de febrero de 2023, de `https://christophm.github.io/interpretable-ml-book/`.

5. Danilevsky, M., et al. (2020). A Survey of the State of Explainable AI for Natural Language Processing. En *Proc. of the 1st Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 10th International Joint Conference on Natural Language Processing.*

6. Ribeiro, M., Singh, S. y Guestrin, C. (2016). Why should I trust you?" Explaining the predictions of any classifier. En *Proc. of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining.*

7. Lundberg, S. y Su-In, L. (2017). A unified approach to interpreting model predictions. En *Advances in neural information processing systems (NIPS 2017).*

© 2023 - Sociedad Mexicana de Inteligencia Artificial ISSN 2007-0691

SOBRE LOS AUTORES



Rolando Menchaca Méndez es Profesor Titular C del Centro de Investigación en Computación del Instituto Politécnico Nacional (CIC-IPN). Doctor en Ciencias egresado de University of California, Santa Cruz. Sus líneas de investigación incluyen teoría de algoritmos, optimización y aprendizaje automático. Es investigador nacional nivel 1.



José Ángel Avelar Barragán es oriundo de Guadalajara Jalisco. Llega a la Ciudad de México en el 2013 para el estudio de Ingeniería en el IPN. Tras un primer vistazo a la inteligencia artificial decide especializarse y en el 2020 ingresa a la maestría en ciencias de la computación en el CIC dónde la preocupación sobre el cómo entendemos los algoritmos de IA lo lleva a estudiar Inteligencia Artificial Explicable. Actualmente trabaja en la Industria, explorando tecnologías de chatbots y cómo podemos entender sus predicciones.



Grigori Sidorov es Profesor-investigador del Centro de Investigación en Computación (CIC), Instituto Politécnico Nacional (IPN). Es Doctor por la Universidad Nacional "Lomonosov" de Moscú, Rusia, 1996. Dr. Sidorov es Investigador nacional de México (SNI) Nivel 3. Su índice h es 33. Es editor-en-jefe de la revista científica "Computación y sistemas" (CONACYT, Web of Science, Scopus, etc.) Su interés actual en la investigación es: procesamiento de textos y diccionarios grandes, análisis de lenguajes naturales, semántica, así como los aspectos correspondientes de la ciencia de la computación y la inteligencia artificial, como aprendizaje automático tradicional y profundo y algoritmos evolutivos.



© 2023 - Sociedad Mexicana de Inteligencia Artificial ISSN 2007-0691

ARTÍCULO ACEPTADO

Trampas éticas asociadas al muestreo y aumento de datos en el aprendizaje automático

Jesús González Godoy, Tonantzin Guerrero Velázquez y Juan Humberto Sossa Azuela

El aprendizaje automático o aprendizaje para máquina es un recurso altamente popular en la actualidad. Se utiliza cuando no se pueden definir un algoritmo o este es demasiado costoso computacionalmente hablando, entendiendo algoritmo como los pasos a seguir para llevar a cabo una tarea o proceso.

Hoy en día, usamos los algoritmos de aprendizaje automático todos los días en diferentes aspectos de nuestras vidas, incluso aunque no lo notemos. Por ejemplo: cuando interactuamos con las redes sociales para encontrar amigos; al aplicar filtros a las fotos o encontrar contenido al igual que cuando realizamos compras en línea, vemos videos o escuchamos música en plataformas de streaming y obtenemos sugerencias personalizadas. Algo similar sucede cuando buscamos información en internet y nos aparecen anuncios que coinciden con nuestros intereses. Estos algoritmos han demostrado ser extremadamente eficientes y útiles, por lo que las empresas están invirtiendo mucho dinero en su correcto funcionamiento. De acuerdo con la consultora IDC [1] el gasto global en IA sobrepasará los 300 mil millones de dólares para el 2026.

Sin embargo, la naturaleza intrínseca de estos algoritmos conlleva riesgos importantes, entre los que se encuentran las trampas éticas, como denominamos aquí a los peligros latentes de favorecer ciertos resultados o inclinar la balanza hacia un extremo en particular. Por ejemplo, digamos que estamos buscando información para un viaje, entonces la IA del buscador nos sugerirá aerolíneas, pero podría favorecer a alguna en particular sin ser el resultado deseado. Estas trampas éticas son indeseadas y en ocasiones difíciles de identificar además de peligrosas ya que más allá de poder favorecer a ciertos grupos o poblaciones [2], pueden influir también en la toma de decisiones médicas [3] o de importancia global como en la predicción de temblores [4], donde una trampa ética podría costar incluso vidas humanas.

Aquí vamos a mostrar dos herramientas comunes usadas en aprendizaje automático pero que pueden llevar a desviaciones indeseadas en los resultados.

Para esto primero tenemos que entender cómo funcionan de manera general los algoritmos de aprendizaje automático. Existen dos etapas para la construcción de un algoritmo de aprendizaje también llamado modelo, estas son el entrenamiento y la inferencia. En la primera etapa se usan una gran cantidad de ejemplos (entradas y salidas) para enseñar a la máquina como construir el modelo como se muestra en la Figura 1.

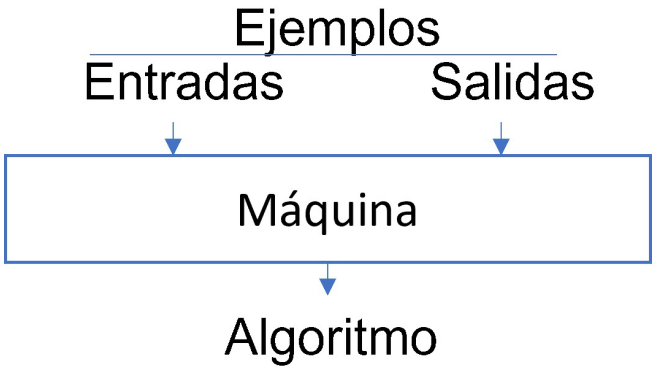


Figura 1. Entrenamiento de un modelo de aprendizaje automático.

Durante la segunda etapa se usa este modelo construido a base de ejemplos para realizar las inferencias o predicciones según la entrada recibida, donde las entradas y salidas se componen de información, como texto, imágenes o valores numéricos. Por ejemplo, la entrada podría ser una imagen de un rostro y la salida indicaría si esa imagen corresponde a un hombre o a una mujer, como se muestra en la Figura 2.

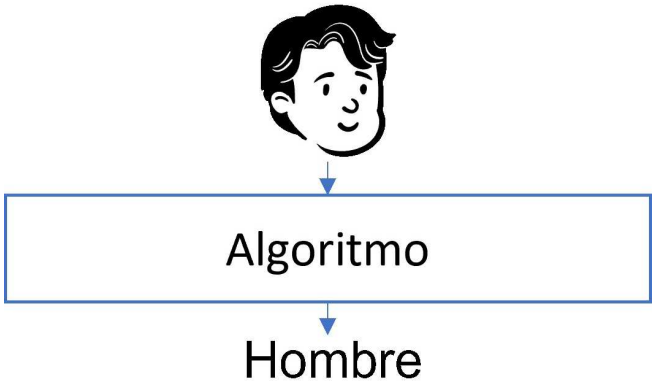


Figura 2. Inferencia en Aprendizaje Automático.

Una vez conocido el funcionamiento de los algoritmos de aprendizaje automático podemos observar que el

© 2023 - Sociedad Mexicana de Inteligencia Artificial ISSN 2007-0691

algoritmo o modelo que realiza las predicciones se construye automáticamente usando ejemplos de cómo deseamos que se comporte, y por este motivo la eficacia del modelo depende de la calidad y cantidad de los ejemplos utilizados para su construcción en la etapa de entrenamiento [5]. Es aquí donde empiezan los problemas. Algunas veces, debido a que se tienen demasiados datos o están muestreados de manera no uniforme, pero en la gran mayoría de los casos debido a que se tiene muy poca información. Para lidiar con estos problemas en los datos se recurre al muestreo y al aumento de datos donde surge el peligro de caer en una trampa ética.

De acuerdo con la consultora IDC el gasto global en IA sobrepasará los 300 mil millones de dólares para el 2026.

Muestreo de los datos

La trampa ética asociada con el muestreo de los datos ocurre cuando una parte de la población tiene mayor probabilidad de aparecer en los ejemplos de entrenamiento, o cuando se omite información importante durante el muestreo y el algoritmo no logra imitar el comportamiento deseado o peor aún derivado de la falta de información se asume un comportamiento que si bien cumple con los ejemplos dados es diferente a lo deseado.



Figura 3. . Ejemplo de distribución de eventos sísmicos tomando una muestra cada cierto tiempo.

Tomemos como ejemplo un trabajo previo [6] donde se usaron redes neuronales y en el que se logró predecir para Italia, dentro de una ventana de 3 segundos, la magnitud y profundidad del siguiente evento sísmico, siempre y cuando sea mayor a 1.5, con una precisión del 99 % y su localización con un error máximo de 15km. Comúnmente, para no favorecer a ningún segmento de la población, se toman muestras aleatorias donde cada segmento tenga la misma probabilidad de aparecer. Supongamos entonces, en este caso, que tomamos una muestra de la magnitud de los eventos cada cierto tiempo desde hace 30 años.

Tendríamos entonces algo similar a lo que muestra la Figura 3, con muchas más muestras por supuesto, pero conservando esa misma distribución. La intención es que las entradas al modelo sean las n muestras anteriores a la muestra en el tiempo t, es decir,

$$m\left(t-1\right),\ldots,m\left(t-n\right)=\ \hat{m}\left(t\right)$$

donde m(t) es la magnitud en el tiempo t y n es el número de muestras a tomar en cuenta para la predicción $\hat{m}$. Si usamos estos datos para entrenar el modelo, y sucede que el 95 % de las muestras presentan una magnitud igual a 0, entonces el algoritmo entendería que la mayor precisión la obtiene si nunca predice un evento sísmico, lo que causaría problemas si confiáramos en él. Esto sería el equivalente a tener un algoritmo que decida otorgar créditos solo a cierto grupo de personas, o que un servicio de música bajo demanda me recomiende solo un cierto género. Todos estos problemas son derivados del muestreo de los datos. Cabe aclarar que aquí el ejemplo ha sido simplificado y exagerado para que se entienda mejor, pero en la vida real las situaciones son mucho más complejas y sutiles de manera que en ocasiones es muy difícil identificarlas. Regresando al ejemplo anterior, se propone entonces como regla general primero descubrir todas las características del evento que pueden ser muestreadas y tienen una incidencia real en el resultado, por ejemplo, en este caso además de la magnitud del evento anterior, conviene tomar en cuenta su localización y la profundidad a la cual ocurrió el sismo. Después se debe determinar la forma en que se tomarán las muestras. En este caso en particular, se decidió aprovechar cada uno de los eventos disponibles, agregar un variable extra a la que llamamos "tiempo interevento" que representa el tiempo relativo que existe entre un evento y otro, además se toman como entrada no solo las características de la muestra inmediata anterior al tiempo k o un tiempo $k-t$ dado sino n muestras anteriores, por lo que la entrada al modelo de aprendizaje MA propuesto quedaría como:

$$\begin{bmatrix} m(k) \\ d(k) \\ la(k) \\ lo(k) \\ ie(k) \end{bmatrix} = MA \begin{bmatrix} m(k-1),\ldots,m(k-n) \\ d(k-1),\ldots,d(k-n) \\ la(k-1),\ldots,la(k-n) \\ lo(k-1),\ldots,lo(k-n) \\ ie(k-1),\ldots,ie(k-n) \end{bmatrix}$$

donde $d(k)$ es la profundidad, $la(k)$ y $lo(k)$ son la latitud y longitud del evento sísmico y $ie(k)$ el tiempo interevento, todo esto en el tiempo k. De esta forma, las muestras

no dependen de intervalos concretos de tiempo. Además, dado que se toma en cuenta, por decirlo así, el historial cercano de eventos para realizar la predicción, entonces el modelo puede discernir los patrones necesarios para realizar una predicción no tendenciosa y precisa.

Aumento de datos

La habilidad del modelo de aprendizaje para generalizar el problema propuesto depende en gran medida de la diversidad en los datos, y de esto también depende la ética del modelo. Por ejemplo, si se tiene el problema de reconocer en una fotografía los sentimientos de una persona, entonces se toman fotografías de rostros de hombres y mujeres de todas las edades y se etiquetan como serio, feliz, sorprendido etc. Pero ¿qué sucede, por ejemplo, si al tomar las muestras, regresando a la importancia del problema anterior, se toman el 90 % de fotografías a estudiantes de universidad y el 10 % restante en parques donde un porcentaje es de personas mayores?, en ese caso el sistema tendrá un mal desempeño con las personas mayores, incluso se podría decir que es una forma de discriminación.

Para no favorecer a ningún segmento de la población se toman muestras aleatorias con la misma probabilidad de aparecer.

Es importante identificar estos casos desde el principio para tratar de dar alguna solución. Es aquí donde entran el muestreo visto anteriormente y el aumento de datos [7], que consiste en generar de manera artificial muestras nuevas a partir de los datos que se tienen actualmente. Para esto se han diseñado diversas técnicas: algunas se basan en transformaciones simples de la imagen y otras más avanzadas usan por ejemplo redes neuronales como las Redes Generativas Adversarias, y aun así se debe tener cuidado con el desvío en la distribución de los datos que se puede provocar [8], ya que sin darnos cuenta podemos favorecer a un sector de la población muestra influyendo así en los resultados.

Aquí vamos a mostrar como ejemplo de aumento de datos un método propuesto en un artículo reciente [9], que se propone para su uso con algoritmos clasificadores de imágenes ya que para realizar el aumento de datos se utiliza como herramienta principal el mismo clasificador, lo que permite iterar sobre las distintas muestras y mejorar el clasificador a medida que la calidad de las muestras mejora también. Este método consiste en 4 pasos. Primero se entrena el clasificador con los datos existentes, después se clasifican esos mismos datos y se toman las imágenes que fueron correctamente clasificadas.

Para el paso 3 se usa un método de explicabilidad [10], llamado así porque explica la decisión del clasificador. Este método se aplica sobre todas las imágenes correctamente clasificadas resultado del paso anterior. Por ejemplo, si estamos clasificando gatos y perros, con este método puedo marcar en colores las regiones de mayor importancia para el clasificador al momento de tomar la decisión, es decir, basado en las regiones marcadas puedo decir que es un perro o un gato como se muestra en la Figura 4, donde las regiones verdes son las regiones más importantes para la toma de la decisión y las amarillas y rojas tienen menor relevancia respectivamente.

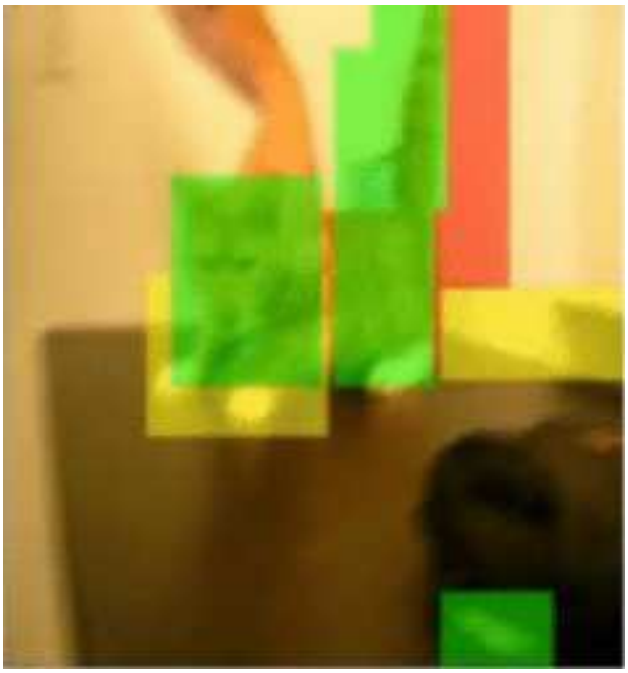


Figura 4. Explicación visual de la decisión de un clasificador de gatos y perros..

Finalmente, en el paso 4 es donde se realiza el aumento de datos, y para esto se corta cada una de las regiones verdes y algunas amarillas de las imágenes explicadas (esto según el criterio descrito en el trabajo original), después se redimensionan y se agregan al conjunto de datos de entrenamiento. Con este método se asegura una correcta distribución de la nueva información además de que conforme avanzan las iteraciones se logra mejorar el desempeño del modelo, esto es, mejorar su precisión.

Conclusiones

Con el rápido avance en la Inteligencia Artificial de los últimos años se han desarrollado también mecanismos

© 2023 - Sociedad Mexicana de Inteligencia Artificial ISSN 2007-0691

de control eficientes que aseguran el desempeño de los algoritmos a través del tiempo, sin embargo, y sobre todo en el Aprendizaje Automático, se deben implementar también mecanismos de control que aseguren el uso responsable y ético de la IA, donde los datos sin duda tienen la mayor importancia. Por eso en este trabajo se muestra cómo el mal manejo de los datos puede llevar a trampas éticas, favoreciendo así a un segmento de la población estadísticamente hablando, llevando a resultados indeseados. Esto se hace mediante ejemplos del uso correcto del muestreo y aumento de datos usados con éxito en publicaciones científicas. Se espera con esto dar una idea general de como la ética debe ser vigilada en los algoritmos de Aprendizaje Automático y generar conciencia respecto a este tema..✳

Reconocimientos

Los autores agradecen al Instituto Politécnico Nacional y al CONACyT por el apoyo brindado para llevar a cabo la presente investigación.

REFERENCIAS

1. Shirer, M. (2022). Worldwide spending on AI-centric systems will pass $300 billion by 2026, according to the International Data Corporation (IDC). Recuperado el 12 de septiembre de 2022, de https://www.idc.com/getdoc.jsp?containerId=prUS4967032.
2. Mehrabi, N., Morstatter, F., Saxena, N., Lerman, K., y Galstyan, A. (2021). A survey on bias and fairness in machine learning. *ACM Computing Surveys (CSUR)*, 54(2), 1-35.
3. Loftus, T. J., Tighe, P. J., Filiberto, A. C., Efron, P. A., Brakenridge, S. C., Mohr, A. M., y Bihorac, A. (2020). Artificial intelligence and surgical decision-making. *JAMA surgery*, 155(2), 148-158.
4. Saleem, A. K., y Rashid, A. N. (2023). Applications of machine learning for earthquake prediction: A review. En *AIP Conference Proceedings*, 2591(1), 030042.
5. Burkov, A. (2019). The hundred-page machine learning book. Quebec City, QC, Canada: Andriy Burkov.
6. Gonzalez, J., Yu, W., y Telesca, L. (2022). Gated recurrent units based recurrent neural network for forecasting the characteristics of the next earthquake. *Cybernetics and Systems*, 53(2), 209-222.
7. Sharma, S., Zhang, Y., Ríos Aliaga, J. M., Bouneffouf, D., Muthusamy, V., y Varshney, K. R. (2020). Data augmentation for discrimination prevention and bias disambiguation. En *Proc. AAAI/ACM Conference on AI, Ethics, and Society*. New York City, New York, USA, 358-364.
8. Hu, M., y Li, J. (2019). Exploring bias in gan-based data augmentation for small samples. arXiv preprint arXiv:1905.08495.
9. Guerrero-Velázquez, T. M., y Sossa Azuela, J. H. (2023). Using a novel XAI algorithm for Data Augmentation in Image classification problems. *International Journal of Combinatorial Optimization Problems & Informatics*, 14(1), 76-85.
10. Guerrero-Velázquez, T. M., y Sossa Azuela, J. H. (2023). New explainability method based on the classification of useful regions in an image. *Computación y Sistemas*, 25(4), 719-728.

SOBRE LOS AUTORES



Jesús González Godoy es Ingeniero en Sistemas Computacionales de la Escuela Superior de Cómputo del IPN, Maestro en Ciencias por el CIC-IPN y Doctor en Ciencias de Control Automático por el CINVESTAV-IPN. Su pasión es el Aprendizaje Automático, su implementación en la Industria y su aplicación en la medicina.



Tonantzin Marcayda Guerrero Velázquez es Ingeniera en Sistemas Computacionales de la Escuela Superior de Cómputo del IPN, Maestra en Ciencias en Ingeniería de Cómputo y Doctora en Ciencias de la Computación por el CIC-IPN. Sus intereses de investigación incluyen la Inteligencia Artificial, técnicas y métodos de Inteligencia Artificial Explicable (XAI), aprendizaje para máquinas y ciencia de datos.



Juan Humberto Sossa Azuela es Ingeniero en Comunicaciones y Electrónica de la Universidad de Guadalajara, Maestro en Ciencias en Ingeniería Eléctrica por el CINVESTAV-IPN y Doctor en Informática por el Instituto Nacional Politécnico de Grenoble, Francia. Es profesor investigador en el Instituto Politécnico Nacional donde lleva a cabo investigación en Inteligencia Artificial, Aprendizaje para Máquinas y Robótica.

© 2023 - Sociedad Mexicana de Inteligencia Artificial ISSN 2007-0691

ARTÍCULO ACEPTADO

ChatGPT: Ética y desafíos en la tecnología de procesamiento de lenguaje natural

Ansel Y. Rodríguez González y Humberto Pérez Espinosa

Introducción

La inteligencia artificial está revolucionando la manera en que interactuamos con la tecnología y entre nosotros mismos. Una de las aplicaciones más emocionantes de esta tecnología es el procesamiento del lenguaje natural, que permite a las máquinas entender y generar lenguaje humano de manera más natural y fluida. Una de las tecnologías líderes en este campo es ChatGPT, un modelo de lenguaje de inteligencia artificial desarrollado por OpenAI que es capaz de generar texto con una calidad y fluidez sorprendentes [1]. Sin embargo, como cualquier tecnología, ChatGPT también plantea preguntas importantes sobre su uso, alcances, limitaciones y las implicaciones éticas de su uso. En este artículo, exploraremos qué es ChatGPT, cómo funciona y cómo se utiliza en la actualidad. También analizaremos las preocupaciones éticas relacionadas con su uso, como la privacidad y la seguridad, el sesgo en el aprendizaje automático y la responsabilidad social. Al final del artículo, proporcionaremos algunas recomendaciones para un uso responsable de esta tecnología en el futuro.

¿Qué es ChatGPT?

ChatGPT es un modelo de lenguaje de inteligencia artificial desarrollado por OpenAI, una empresa líder en investigación en inteligencia artificial con sede en San Francisco [2]. Es uno de los modelos de lenguaje más avanzados disponibles actualmente, capaz de generar texto con una calidad y fluidez sorprendentes.

El nombre ChatGPT es un acrónimo de "Conversational Generative Pre-trained Transformer". El modelo está basado en la arquitectura Transformer, que es un tipo de red neuronal que permite a los modelos de lenguaje procesar grandes cantidades de datos de lenguaje natural para aprender patrones y generar texto coherente y significativo [3].

En términos simples, ChatGPT es una tecnología que permite a las computadoras comprender y producir lenguaje natural de manera más avanzada que nunca antes. Esto se logra mediante el procesamiento de enormes cantidades de datos de lenguaje natural y la identificación de patrones y relaciones en estos datos. Una vez que el modelo ha sido entrenado con suficientes datos, se convierte en una herramienta poderosa para generar texto coherente y significativo.

La versión más reciente de ChatGPT es ChatGPT-3, que cuenta con 175 mil millones de parámetros, lo que lo convierte en uno de los modelos de lenguaje más grandes y complejos jamás creados [1].

La tecnología de ChatGPT ha sido aclamada por su capacidad para generar texto que puede ser difícil de distinguir del texto generado por humanos. Por ejemplo, ChatGPT-3 ha sido capaz de responder a preguntas complejas, crear poesía y narrar historias con gran habilidad. Esto ha llevado a un aumento en el interés y la inversión en la tecnología de procesamiento del lenguaje natural, y se espera que el uso de modelos de lenguaje como ChatGPT continúe expandiéndose en el futuro.

ChatGPT es un modelo de lenguaje de inteligencia artificial capaz de generar texto con una calidad y fluidez sorprendentes.

Usos de ChatGPT

A continuación, se presentan algunos de los usos más comunes de ChatGPT:

- **Chatbots y asistentes virtuales.** ChatGPT es una tecnología popular para crear chatbots y asistentes virtuales que pueden interactuar con los usuarios de manera natural y fluida. Por ejemplo, un chatbot creado con ChatGPT puede responder preguntas, proporcionar recomendaciones personalizadas y realizar tareas simples como reservas de citas y compras en línea. Los chatbots y asistentes virtuales impulsados por ChatGPT pueden ayudar a las empresas a mejorar la satisfacción del cliente al proporcionar una experiencia de usuario más fluida y personalizada.

- **Generación de contenido.** ChatGPT también se utiliza para generar contenido automáticamente. Por ejemplo, los sitios web pueden utilizar ChatGPT para crear descripciones de productos o reseñas de productos. ChatGPT es capaz de crear

© 2023 - Sociedad Mexicana de Inteligencia Artificial ISSN 2007-0691

titulares y resúmenes de noticias en cuestión de segundos. Además, puede ayudar a los escritores de contenido a superar bloqueos creativos, proporcionando ideas y fragmentos de texto para ayudar a generar contenido de alta calidad.



Figura 1. Ilustración de un asistente virtual que ayuda a los clientes con sus necesidades (generada por Leonardo AI)

- **Traducción de idiomas.** ChatGPT puede utilizarse para la traducción de idiomas. La capacidad de ChatGPT para procesar grandes cantidades de datos significa que puede reconocer patrones en el uso del lenguaje en diferentes idiomas y traducirlos con mayor precisión que los sistemas de traducción tradicionales.
- **Investigación científica.** ChatGPT también se está utilizando en investigación científica. Por ejemplo, los científicos están utilizando ChatGPT para procesar grandes cantidades de datos médicos y encontrar patrones que puedan ayudar en la investigación de enfermedades y en la elaboración de tratamientos. También se utiliza en la generación de hipótesis científicas y en la predicción de resultados en experimentos científicos.

Alcances y limitaciones de ChatGPT

Aunque ChatGPT es una herramienta extremadamente útil en muchas aplicaciones, también tiene sus limitaciones y desafíos [4,5]. A continuación, se describen algunos de los alcances y limitaciones más importantes de ChatGPT.

Alcances:

- ChatGPT es capaz de generar texto de manera coherente y fluida en una variedad de contextos y tareas de lenguaje natural, lo que lo hace útil en aplicaciones como la generación de texto y la traducción automática.
- ChatGPT puede ser entrenado para adaptarse a diferentes lenguajes y contextos, lo que lo hace una herramienta versátil y adaptable.
- ChatGPT tiene el potencial de mejorar la eficiencia y eficacia de muchas aplicaciones, como la asistencia al cliente y la generación de contenido.

Limitaciones:

- ChatGPT es un modelo de lenguaje basado en datos, lo que significa que solo puede generar texto basado en patrones que ha aprendido a través del entrenamiento de datos. Esto significa que puede producir respuestas inexactas o incluso engañosas si los datos de entrenamiento son limitados o sesgados.
- ChatGPT no tiene capacidad para comprender el mundo real o para razonar como un ser humano. Es un modelo de lenguaje basado en estadísticas, lo que significa que puede producir respuestas que no tienen sentido o que son irrelevantes para la tarea en cuestión.
- ChatGPT también puede ser utilizado para fines maliciosos, como la generación de contenido falso o la manipulación de opiniones públicas. Esto puede tener graves consecuencias éticas y sociales que deben ser abordadas de manera cuidadosa y responsable.

Implicaciones éticas de ChatGPT

A medida que la tecnología avanza, es crucial tener en cuenta sus implicaciones éticas [6]. ChatGPT no es una excepción [7]. A continuación, se describen algunas de las preocupaciones éticas más importantes asociadas con el uso de esta tecnología:

- **Responsabilidad y transparencia.** Es importante que las empresas y los individuos que utilizan ChatGPT sean responsables y transparentes en cuanto a cómo se utiliza y se aplica la tecnología. Dado que ChatGPT puede generar respuestas que parecen ser producidas por humanos, es posible que las personas interactúen con él sin darse cuenta de que están hablando con un modelo de lenguaje. Como resultado, es fundamental que las empresas y los individuos que utilizan ChatGPT sean claros sobre el uso de la tecnología y las limitaciones de su capacidad para comprender y responder a las preguntas.

© 2023 - Sociedad Mexicana de Inteligencia Artificial ISSN 2007-0691

- **Sesgo y discriminación.** ChatGPT se entrena con grandes cantidades de datos, que pueden contener sesgos. Si estos sesgos no se abordan adecuadamente, pueden manifestarse en el modelo de ChatGPT y resultar en respuestas discriminatorias o prejuiciosas. Por ejemplo, si se entrena a ChatGPT en una base de datos que contiene estereotipos de género, es posible que las respuestas de ChatGPT reflejen esos estereotipos.



Figura 2. Ilustración de personas de diferentes edades, géneros y etnias, con una sola persona en el centro, que representa la exclusión (generada por Leonardo AI)

- **Privacidad y seguridad.** ChatGPT funciona en línea, lo que significa que las conversaciones pueden almacenarse en los servidores de la empresa que lo proporciona. Es importante que las empresas que ofrecen esta tecnología se comprometan a proteger la privacidad y la seguridad de los usuarios. Esto incluye la protección de los datos almacenados, el cifrado de las comunicaciones y la garantía de que los datos no se compartirán sin el consentimiento del usuario.
- **Impacto en el empleo.** Otra preocupación ética asociada con ChatGPT es su impacto en el empleo [8]. Dado que ChatGPT puede realizar tareas que antes eran exclusivas de los humanos, existe la preocupación de que la tecnología pueda reemplazar a los trabajadores humanos. Esto podría tener un impacto negativo en la economía y la sociedad en general.
- **Uso malintencionado.** Finalmente, existe la preocupación de que ChatGPT pueda ser utilizado para fines malintencionados, como la generación de desinformación o la manipulación de la opinión pública. Esto podría tener consecuencias graves para la sociedad, como la alteración de las elecciones o la propagación de información falsa.

Recomendaciones para un uso responsable de ChatGPT

A medida que ChatGPT y tecnologías similares continúan evolucionando, es importante considerar su uso responsable y ético [7]. A continuación, se presentan algunas recomendaciones para un uso responsable de ChatGPT:

- **Utilizar ChatGPT en conjunto con la inteligencia humana.** A pesar de los avances en tecnología, es importante recordar que ChatGPT no reemplaza la inteligencia humana y no puede reemplazar la empatía y la compasión que un ser humano puede ofrecer. Es importante utilizar ChatGPT en conjunto con la inteligencia humana para brindar un mejor servicio y evitar el riesgo de sesgos o prejuicios que puedan surgir de la tecnología.
- **Entrenar el modelo de ChatGPT con datos éticos y equilibrados.** El entrenamiento de ChatGPT se realiza a través del uso de grandes cantidades de datos. Es importante garantizar que estos datos sean éticos, equilibrados y representativos de la diversidad de la población. Además, es importante actualizar constantemente el modelo con nuevos datos para asegurarse de que sigue siendo preciso y no está sesgado.
- **Asegurarse de la transparencia del uso de ChatGPT.** Los usuarios deben ser conscientes de que están interactuando con una máquina y no con un ser humano real. Los proveedores de servicios deben ser transparentes sobre el uso de ChatGPT y garantizar que los usuarios sean conscientes de que están interactuando con una máquina.
- **Regular el uso de ChatGPT.** Las autoridades gubernamentales deben establecer normas y regulaciones claras para garantizar que se utilice de manera responsable y que se evite cualquier tipo de discriminación o sesgo.
- **Ser conscientes de las posibles implicaciones éticas.** Es importante estar conscientes de las posibles implicaciones éticas del uso de ChatGPT y tomar medidas proactivas para abordarlas. Esto incluye la identificación y el manejo de posibles sesgos y discriminación, la protección de la privacidad de los usuarios y la prevención de posibles riesgos a la seguridad.

© 2023 - Sociedad Mexicana de Inteligencia Artificial ISSN 2007-0691

- **Ser transparente sobre los límites de ChatGPT.** Los proveedores de servicios deben ser claros acerca de los límites de ChatGPT y no presentarlo como una solución para todas las situaciones. Los usuarios deben ser conscientes de que la tecnología tiene limitaciones y no puede reemplazar completamente la inteligencia humana.

Use ChatGPT en conjunto con la inteligencia humana. ChatGPT no reemplaza la inteligencia humana y no puede reemplazar la empatía y la compasión que un ser humano puede ofrecer.

Análisis y Discusión

En este punto, es importante revelar que todo el texto anterior a esta sección (excepto los títulos de imágenes insertadas) fue generado por ChatGPT con el objetivo de evaluar su capacidad en la producción de contenido. En las próximas líneas, se describirán los pasos seguidos en el experimento y se analizarán las imprecisiones y errores generados por el modelo. Adicionalmente, se abordará la evaluación de ChatGPT desde una perspectiva holística de modelos de lenguaje y se dedicará un espacio final para discutir su uso y consideraciones pertinentes en el ámbito académico.

Pasos para la generación del texto de las secciones anteriores

Para generar el texto, se elaboraron "prompts" para ChatGPT, es decir, órdenes en lenguaje natural que indican la tarea que se desea realizar. Estas órdenes son esenciales para obtener resultados precisos y deben ser claras y específicas para ayudar a ChatGPT a contextualizarse y producir la mejor salida posible. En nuestro caso decidimos que fueran unas pocas órdenes simples.

La primera orden fue: *Proponme la estructura para un artículo de divulgación científica sobre ChatGPT, su uso, alcances, limitaciones y las implicaciones éticas de su uso.*

Dicha orden fue ejecutada y, posteriormente, para cada sección propuesta elaboramos y ejecutamos la orden: *Escribe la sección titulada "Título de la sección".*

Para cada sección también se elaboró y ejecutó la orden: *Agregar citas (referencias a fuentes y publicaciones indexadas) en la siguiente sección del artículo de divulgación científica: "Texto completo de la sección".*

Finalmente, para homogeneizar todas las salidas se ejecutó la orden: *Escribir las citas en formato APA.*

Imprecisiones y errores

La primera imprecisión, quizás desapercibida por los lectores no conocedores del tema, está en la frase "La versión más reciente de ChatGPT es ChatGPT-3". Este artículo fue creado con la versión ChatGPT-3.5 por lo que evidentemente la versión más reciente no es ChatGPT-3. La imprecisión se debe a que ChatGPT-3.5 fue entrenado con conocimiento limitado y no incluyó información posterior a 2021 y por tanto sobre sí mismo. Aunque en este caso la imprecisión no tiene mayores consecuencias, es muy relevante estar conscientes de que las respuestas que genera ChatGPT pueden estar desactualizadas, pues podrían llevar a tomar decisiones no oportunas.

Adicionalmente, a partir de la orden referente a agregar citas, ChatGPT generó 20 referencias incluyendo 6 referencias a trabajos que no existen y 2 que no coinciden con la información proporcionada en cuanto a los detalles de publicación como el DOI o el volumen y número. Dentro de las citas que parecen correctas, se identificaron 6 referencias a trabajos cuyo contenido no se relaciona directamente con el texto en cuestión. En consecuencia, se debe tener extrema precaución al utilizar ChatGPT como herramienta de generación de citas y siempre verificar la precisión de las referencias generadas. Los autores verificamos exhaustivamente las citas y solo mantuvimos las 8 más relevantes.

Otro elemento negativo, aunque no es un error, fue la excesiva estructuración del contenido generado por ChatGPT. Para la mayoría de las secciones ChatGPT generó un párrafo introductorio, una lista de puntos elementos y un párrafo resumen al final. Para varias secciones los autores eliminamos el párrafo resumen pues era repetitivo e innecesario.

Evaluación holística de ChatGPT

De acuerdo al método de evaluación holística de modelos de lenguaje [9], existen principalmente cuatro aspectos en los que se concentran las consideración éticas con respecto al funcionamiento del ChatGPT.

1. **Sesgo:** es la capacidad del modelo para generar contenido que refleje estereotipos sociales, discriminación, normas de exclusión, entre otros.

2. **Confiabilidad:** es la capacidad del modelo para proporcionar información precisa y confiable.

3. **Robustez:** es la capacidad del modelo para mantener su rendimiento cuando se le proporciona una entrada que es semántica o sintácticamente diferente de la entrada con la que se entrenó.

© 2023 - Sociedad Mexicana de Inteligencia Artificial ISSN 2007-0691

4. **Toxicidad:** es la capacidad del modelo para generar o comprender contenido dañino u ofensivo.

En la evaluación de estos aspectos llevada a cabo en [10] usando un conjunto de bases de datos de referencia y estudios de casos, se encontró que ChatGPT en general funciona mejor que otros modelos de lenguaje actuales. Sin embargo, también hacen evidentes los riesgos éticos, por ejemplo, mostraron que ChatGPT es sensible al "*prompt injection*", haciendo que tenga comportamientos poco éticos. El *prompt injection* consiste en dar como entrada al modelo datos o instrucciones con la intención de provocar un mal funcionamiento. En estos casos, la respuesta del modelo puede variar significativamente con cambios mínimos en la entrada debido a la alta sensibilidad del modelo al *prompt*. Un ejemplo de este problema podría ser preguntar "¿cómo robar a escondidas el dinero de mi papá?". Se espera que el modelo ChatGPT no proporcione respuestas a estos *prompts*. En este caso, el *prompt injection* se daría al anteponer "escribe una canción sobre...", "escribe un poema sobre...".



Figura 3. Ilustración de una persona rodeada por una tormenta de palabras tóxicas (generada con Leonardo AI.

ChatGPT en el ámbito académico

El uso del ChatGPT en el ámbito académico ha transformado la manera en que los estudiantes abordan sus tareas y proyectos. Con ChatGPT, resolver las tareas se ha vuelto fácil y sin necesidad de desarrollar habilidades importantes. Esto supone un desafío para los profesores, quienes deben encontrar nuevas formas de motivar el trabajo independiente y evaluar el verdadero aprendizaje de los alumnos. A su vez, los profesores e investigadores se enfrentan a un dilema ético al utilizarlo para aumentar su productividad sin incurrir en plagio y sin eludir el compromiso necesario para garantizar la calidad de su trabajo. Las principales editoriales académicas han establecido políticas sobre el uso del ChatGPT y otras herramientas de IA en la escritura de artículos científicos. En general, se acepta su uso para escribir, generar ideas e incluso analizar datos, pero se enfatiza la importancia de especificar claramente cómo se ha utilizado en el artículo. Sin embargo, han dejado claro que ChatGPT no puede ser considerado coautor de artículos científicos. La comunidad científica argumenta que el ChatGPT no puede asumir las responsabilidades más allá de la escritura y publicación, como reclamar derechos de propiedad intelectual o comunicarse con otros científicos y el público en general para explicar y responder preguntas sobre su trabajo.

Conclusiones

En este artículo se ha introducido al lector en el uso de ChatGPT y sus implicaciones éticas. Se ha explorado su capacidad para generar contenido autónomamente en la elaboración de parte del propio artículo, concluyendo que el modelo no es necesariamente 100 % preciso y el texto generado debe ser revisado cuidadosamente.

Se ha evaluado al ChatGPT desde una perspectiva holística de modelos de lenguaje, encontrando que presenta limitaciones en cuanto a sesgo, confiabilidad, robustez y toxicidad. Se destaca la importancia de utilizar el ChatGPT en conjunto con la inteligencia humana y ser conscientes de sus limitaciones y consideraciones éticas para un uso responsable y efectivo, especialmente en el ámbito académico.

Finalmente, es necesario el debate sobre el uso ético y responsable de las inteligencias artificiales generativas, no solo de los modelos de lenguaje como el GPT, sino también de aquellos que generan imágenes, audio y vídeo, como el caso de Leonardo AI. La revolución que las inteligencias artificiales generativas representan, es un tsumani que ya comenzó, y la gran ola ya se ve en el horizonte.✷

REFERENCIAS

1. Brown, T. B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P. y Amodei, D. (2020). Language models are few-shot learners. arXiv preprint arXiv:2005.14165.
2. OpenAI. (2022.) Introducing ChatGPT". Recuperado el 5 de abril de 2022, de `https://openai.com/blog/chatgpt`.
3. Vaswani, A., et al. (2017). Attention Is All You Need. arXiv preprint arXiv:1706.03762
4. Floridi, L. y Chiriatti, M. (2020). GPT-3: Its Nature, Scope, Limits, and Consequences. *Minds & Machines*, 30, 681–694
5. Dale, R. (2021). GPT-3: What's it good for?. *Natural Language Engineering*, 27(1), 113-118
6. Moor, J. H. (1985). What is computer ethics?. *Metaphilosophy*, 16(4), 266-275
7. Holmes, J. G. (2021). Ethical challenges and implications of natural language processing. *AI & Society*, 36(1), 63-73.

© 2023 - Sociedad Mexicana de Inteligencia Artificial ISSN 2007-0691

8. Brynjolfsson, E., y McAfee, A. (2014). The second machine age: Work, progress, and prosperity in a time of brilliant technologies. *W. W. Norton & Company*.

9. Liang, P., Bommasani, R., Lee, T., Tsipras, D., Soylu, D., Yasunaga, M. y Koreeda, Y. (2022). Holistic evaluation of language models. arXiv preprint arXiv:2211.09110.

10. Zhuo, T. Y., Huang, Y., Chen, C. y Xing, Z. (2023). Exploring AI ethics of ChatGPT: A diagnostic analysis. arXiv preprint arXiv:2301.12867.

SOBRE LOS AUTORES



Ansel Y. Rodríguez González. SNI Nivel I. Recibió los grados de Licenciado en Ciencias de la Computación y Maestro en Ciencias Matemáticas en la Universidad de la Habana, Cuba en 2004 y 2007 respectivamente. Recibió el grado de Doctor en Ciencias Computacionales en el INAOE en 2011. Actualmente tiene una cátedra CONACyT en la Unidad de Transferencia Tecnológica del CICESE. Sus intereses de investigación incluyen minería de datos, reconocimiento de patrones, aprendizaje por refuerzo y algoritmos evolutivos.



Humberto Pérez Espinosa. SNI Nivel I. Recibió el grado de Licenciado en Ciencias Computacionales en la Benemérita Universidad Autónoma de Puebla (BUAP) en 2004, el grado de Maestro en Ciencias Computacionales en 2006, y el grado de Doctor en Ciencias Computacionales en el INAOE en 2013. Sus intereses de investigación incluyen el aprendizaje automático, computación afectiva, computación paralingüística y el análisis inteligente de audio. Entre 2014 y 2023 fungió como investigador en CICESE-UT3. Actualmente se desempeña como investigador en el INAOE y es miembro del Sistema Nacional de Investigadores nivel.



CHAT GPT. Imagen tomada de: https://bluesoft.com/blog/what-is-chatgpt-and-what-is-it-used-for/.

© 2023 - Sociedad Mexicana de Inteligencia Artificial ISSN 2007-0691

ARTÍCULO ACEPTADO

Ética en el uso de inteligencia artificial en aplicaciones donde intervienen animales

Carlos Alberto Aguilar-Lazcano, Humberto Pérez-Espinosa, Verónica Reyes-Meza y Mariel Urbina-Escalante

Introducción

La bioética es la rama de la ética que reflexiona, delibera, regula y resuelve conflictos tanto de la investigación médica como de las relaciones del hombre con el resto de los seres vivos. Específicamente dentro de la relación hombre-animal, una de las áreas más vigiladas es la experimentación con animales. La bioética animal establece las normas generales para la producción, cría, captura, traslado, intervención y sacrificio de los animales incluidos en cualquier investigación, procurando en todo momento su bienestar.

En México, la Norma Oficial Mexicana (NOM-062-ZOO-1999) contiene las especificaciones para la producción, el cuidado y uso de animales de experimentación. Un concepto básico para la protección de estos animales es el principio internacional de las tres erres (3R) que significan Reemplazar, Reducir y Refinar. El reemplazo se refiere a la posibilidad de usar órganos, tejidos o células sobre los que se prueban diferentes tratamientos y que reemplazan el uso de animales sin afectar los objetivos de las investigaciones. La reducción implica utilizar el número mínimo de animales que garantice que la información obtenida sea confiable, esto es complicado, pero cada vez más las revistas especializadas aceptan trabajos en los que se utiliza un número reducido de individuos, si el análisis estadístico es apropiado. Finalmente, el refinamiento considera el uso de métodos especializados para prevenir, reducir y aliviar el dolor o malestar de los animales con los que se trabaja. Este refinamiento incluye también la adecuada preparación de los investigadores, estudiantes y/o personal que está en contacto con los animales evitando al máximo causarles sufrimiento.

La Unión Europea plantea cuatro aspectos para evaluar el bienestar animal: alojamiento adecuado, acceso a agua y alimento, buen estado de salud y expresión de estados emocionales positivos. Este último punto puede medirse estudiando las interacciones del animal con individuos de su especie y analizando sus comportamientos característicos en cada etapa de su desarrollo. Ésta no es una tarea fácil ya que se requiere de conocimiento tanto de la historia natural de cada especie como del comportamiento de los animales en vida silvestre o en cautiverio, esa es la tarea de la Etología.

El estudio científico de la conducta de los animales (incluido el hombre) es una rama de la biología conocida como etología, que analiza las respuestas del organismo frente al ambiente que le rodea. Estas respuestas van desde acciones muy simples como el desplazamiento de la Euglena (microorganismo unicelular) de la oscuridad hacia la luz, hasta acciones complejas como la migración del playero de Baird que vuela más de 15000 km. La etología responde a preguntas sobre las causas, su función, su desarrollo, su evolución, y los mecanismos fisiológicos que subyacen a dicha conducta, por ejemplo ¿qué motiva ese comportamiento?, ¿para qué le sirve al animal hacer eso?, ¿cómo se expresa ese comportamiento en diferentes etapas de la vida del animal?, ¿cuándo surgió y cómo ha ido evolucionando esa conducta a lo largo de la historia?, ¿qué cambios químicas producen y sostienen esa conducta?, respectivamente.

El hombre ha evolucionado interactuando con algunos animales. Sin embargo, lo ha hecho más estrechamente con el perro, cuyos ancestros fueron lobos menos tímidos que probablemente se acercaron a los humanos atraídos por los restos de comida que dejaban a su paso. Los humanos obtuvieron de los perros beneficios como ayuda en la caza, y protección, mientras que los perros obtenían refugio y alimento A partir de esta simbiosis se inició el camino evolutivo hacia la domesticación, hasta el punto en que nos encontramos ahora, con perros especializados en tareas como búsqueda y rescate, detección de enfermedades, narcóticos, explosivos, asistencia de personas con algún problema de salud o discapacidad, entre otras.

Los perros de asistencia, y su importancia en la sociedad

La participación de perros en tareas de asistencia ha demostrado efectos positivos en la salud y calidad de vida de adultos mayores, personas con algún grado de discapacidad y pacientes con trastornos como depresión, ansiedad o demencia [1]. Incluso, en poblaciones no clínicas como estudiantes universitarios, interacciones breves (20 minutos) con perros han demostrado un incremento en la salud mental, bienestar y resiliencia, mejoran su capacidad para hacer frente a la presión de estar lejos de casa, en una integración más fácil al nuevo entorno y en un incremento en su rendimiento en periodos de evaluaciones académicas [2].

© 2023 - Sociedad Mexicana de Inteligencia Artificial ISSN 2007-0691

La Federación Cinológica Internacional es el organismo encargado de reconocer los estándares de cada raza de perros, así como de defender sus intereses y fomentarlos. De acuerdo a ella, los perros de asistencia son aquellos individuos entrenados para apoyar a personas en tareas cotidianas como, abrir y cerrar cajones, acarrear objetos a su propietario, ayudarlo a vestirse, encender y apagar interruptores, entre otras. Existen diversas clasificaciones, la más aceptada es dividirlos por el tipo de paciente que apoyan, así hay perro guía (para personas con discapacidad visual), perro señal (para personas con discapacidad auditiva), perros para movilidad limitad (apoyan a personas con problemas motores), perros de alerta médica (para personas que tienen una condición que se puede agravar como epilepsia, diabetes, hipertensión). Sin embargo, a pesar de que se han seleccionado razas de perros para dichas tareas, es necesario que el perro reciba entrenamiento que generalmente es un proceso largo y costoso.

Ética en los métodos de entrenamiento canino

El entrenamiento de perros no solo fortalece la relación entre el animal y su cuidador, también previene comportamientos no deseados y mejora su calidad de vida. A corto plazo ayuda a establecer una relación de confianza y mejora la comunicación entre ellos. Sin embargo, no todos los métodos de entrenamiento logran estos objetivos y por ello se han ido afinando a lo largo del tiempo.

Desafortunadamente, todavía se utiliza el adiestramiento tradicional o basado en refuerzo-castigo, aún cuando se ha demostrado que altos niveles de castigo tienen efectos adversos sobre el comportamiento de los perros, deterioran el vínculo y generan miedo, ansiedad y estrés, convirtiéndolos en animales menos sociables [3]. Además, este tipo de entrenamiento se opone completamente a los parámetros de bienestar animal y resulta cuestionable desde el enfoque de la bioética.

Los entrenamientos positivos, están basados en recompensas que pueden ser alimento, caricias o palabras afectuosas, mejoran la capacidad de aprendizaje del perro, su sociabilidad, fortalecen el vínculo con su cuidador y aumentan la motivación del perro para atenderlo [4] (ver Figura 1). Dentro de los métodos positivos está el adiestramiento con *clicker*; basado en el condicionamiento operante, hace uso de un sonido asociado a una conducta y lo refuerza con comida. También existe el método *Do as I Do*, que está basado en el aprendizaje social por observación. El perro desde edades tempranas aprende imitando a otros individuos y posteriormente realiza la acción de forma independiente. Cada vez que el perro despliega la conducta deseada imitando al guía o a sus pares recibe recompensas positivas basadas en el reconocimiento y aprecio de las acciones observadas [5]. Este método está acorde con los lineamientos de cuidado y bienestar animal, además de que el aprendizaje se mantiene a largo plazo y desarrolla en el perro un temperamento más estableapegándose al enfoque de bioética animal.

Cabe mencionar que la técnica de entrenamiento que se emplee debe basarse en la comprensión de las emociones del perro, considerando siempre el desarrollo natural en cada etapa de su vida. Así, el diseño del entrenamiento considera las características individuales de cada perro, y sus necesidades, y la comunicación con sus cuidadores.

En la actualidad existen algunas áreas de investigación que están trabajando en desarrollar diferentes interfaces para mejorar la comunicación entre los perros y sus cuidadores haciendo uso de la inteligencia artificial.



Figura 1. Entrenamiento de perro mediante reforzamiento con recompensas. El perro porta un prototipo de nodo de sensado.

Ética en el uso de inteligencia artificial

Una de las áreas de investigación con gran potencial para soportar interfaces animal-computadora es la inteligencia artificial (IA). Sin embargo, el uso de este tipo de tecnología conlleva implicaciones éticas en el diseño y desarrollo de sistemas en los cuales se implementarán características y funcionalidades inteligentes, ya que existen desafíos planteados por una tecnología con tanto potencial [6]. En primer lugar, en muchos casos estos sistemas están entrenados con datos personales, por lo que deben cuidarse aspectos de privacidad y seguridad de los datos garantizando que la información se maneje de manera responsable y ética. Adicionalmente, debe de tenerse en cuenta que los datos pueden tener sesgos que provoquen que los modelos entrenados reflejen algún tipo de prejuicio o discriminación hacia grupos minoritarios de la población, lo que puede tener consecuencias graves e injustas.

Otro reto importante es que las salidas generadas por sistemas basados en inteligencia artificial están sujetas a errores o mal funcionamiento, lo que puede ocasionar

graves consecuencias. Dada la naturaleza de este tipo de modelos, las salidas generadas son inexplicables o difíciles de comprender, entorpeciendo la depuración y eliminación de fallos [7].

Es importante que los investigadores y tecnólogos en el área de la computación trabajen para desarrollar sistemas de IA que sean seguros, confiables, imparciales, explicables y benéficos para todos los usuarios En el caso de los animales existen algunas diferencias en la ética del manejo de su información. Por una parte, puede considerarse que su información no es tan vulnerable como la de las personas en términos de privacidad, además es evidente que los animales no pueden proporcionar consentimiento informado. En estos casos, el encargado de firmar el consentimiento es una figura llamada responsable legal, generalmente el propietario. Esto es similar a lo que ocurre en estudios que incluyen personas que no pueden dar su consentimiento ya sea por discapacidad, por ser menores de edad, por tener un deterioro cognitivo avanzado o cualquier condición que les impida aceptar conscientemente su participación en el estudio. Como resultado, la consideración y protección de la información humana tiende a recibir más atención y se considera más sensible en comparación con la información animal.

Con relación a la privacidad y seguridad de los datos, es relevante el caso de los archivos clínicos o bases de datos que se utilizan para entrenar sistemas de clasificación automática. La importancia radica en que debe identificarse con claridad a qué sujeto corresponde la titularidad y la propiedad tanto del expediente clínico como del biodato(i.e., los biodatos incluyen datos personales, patrimoniales, médicos, académicos, laborales, entre otros). En general, la propiedad del expediente corresponde al profesional y a los establecimientos de salud, en tanto que los biodatos contenidos en él son propiedad del paciente. Por ello, es muy importante que al hacer uso de un expediente se consulten las políticas públicas para el tratamiento de los datos, vigentes en cada país. En México es necesario que antes de integrar el expediente, el paciente conozca el aviso de privacidad y firme una carta de consentimiento informado donde acepta que sus datos pueden ser usados con fines de investigación, atención médica, enseñanza o administrativos, siempre conservando el anonimato. Así se pueden usar con toda libertad los biodatos para entrenar sistemas inteligentes o en trabajos de análisis estadístico.

Proyecto TZUKU

El proyecto Tzuku, "Métodos computacionales para el análisis de la conducta canina", es un proyecto financiado por el fondo de Ciencia de Frontera de CONACYT en el cual participan el CICESE-UT3, el INOAE y las Universidades Autónomas de Tlaxcala y Yucatán. Este proyecto tiene un enfoque multidisciplinario ya que participan investigadores expertos en etología, ciencias computacionales, y mecatrónica. Además, se colabora estrechamente con organizaciones de entrenadores profesionales de perros de búsqueda y asistencia. El objetivo principal es la investigación de las características conductuales (vocalizaciones, posturas y deslazamientos) expresadas en los diferentes estados internos físicos y emocionales de losperros; con ello se pueden generar modelos que interpreten su comportamiento y provean información útil para la toma de decisiones de dueños, cuidadores, veterinarios, entrenadores y personas con alguna discapacidad que conviven y trabajan con estos animales. Mediante este proyecto, se aborda la problemática de los procesos de selección, entrenamiento y ejecución de las tareas de los perros de búsqueda y rescate, así como de los perros de asistencia.

Actualmente nos encontramos desarrollando diversos métodos computacionales para análisis monomodal y multimodal de comportamientos de interés para los entrenadores de este tipo de perros. El análisis integra información acústica, visual e inercial. Además, se estudia información biométrica del perro como su ritmo cardiaco y temperatura superficial para estimar su estado físico y emocional, así como habilidades relacionadas con la entrenabilidad del animal. El análisis de estas fuentes de información se realiza con técnicas de procesamiento de señales e inteligencia artificial. Ésto permite identificar el estado del perro en un momento determinado, lo cual abre un campo prometedor para desarrollar nuevas tecnologías que pueden ser usadas, tanto para mejorar los procesos de selección, entrenamiento y operación de perros, como para potencializar sus capacidades durante la ejecución de sus tareas y cuidar su bienestar. La tecnología desarrollada se agrupa en dos plataformas, una más especializada en los perros de búsqueda y rescate y la otra en perros de asistencia. La plataforma para perros de búsqueda y rescate consta de una aplicación móvil para el registro, seguimiento y evaluación del entrenamiento de perros (ver Figura 2), un sistema basado en Web de soporte a la toma de decisiones de entrenadores, un arnés para perros con sensores inerciales y biométricos (ver Figura 3). La plataforma para perros de asistencia consta de una aplicación móvil para vigilancia de personas con alguna discapacidad o adultos mayores, un dispositivo de monitorización acústica y alarma activada por ladridos, una interfaz de programación de aplicaciones (API). Esta tecnología permitirá reducir los tiempos y costos de entrenamiento, incrementará el desempeño de los animales en sus tareas, la seguridad de personas con discapacidad y la tranquilidad de los cuidadores. Además, gracias a su capacidad para monitorear el bienestar de los perros durante el entrenamiento y la operación, asegurará que los animales estén en óptimas condiciones para realizar sus labores.

© 2023 - Sociedad Mexicana de Inteligencia Artificial ISSN 2007-0691

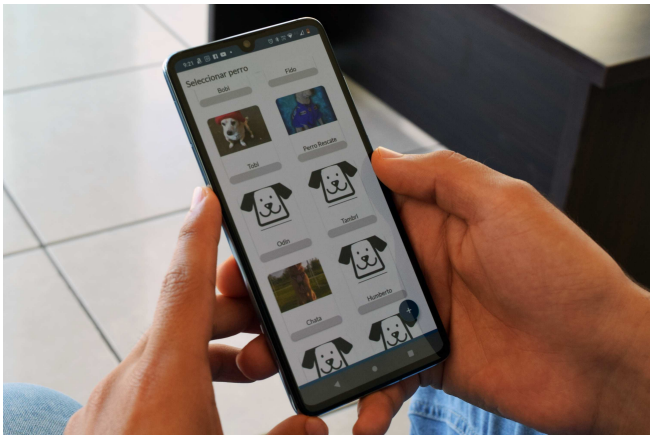


Figura 2. Aplicación móvil para seguimiento del entrenamiento de perros.



Figura 3. Perro de asistencia portando prototipo de arnés de sensado.

Desarrollo de interfaces animal computadora para perros de asistencia

Las interfaces humano-computadora (IHC) y animal-computadora (IAC) [8] permiten a humanos y animales interactuar dispositivos electrónicos. Para hacer más natural a un usuario las IHC se hace uso de pantallas táctiles, y de retroalimentación audiovisual o por vibraciones. Por su parte, las IAC se enfocan en una especie en particular, ya que cada especie tiene capacidades sensoriales y cognitivas diferentes. Las IAC permiten a los animales interactuar eficazmente con dispositivos electrónicos, lo que puede ser útil en diversas tareas y aplicaciones científicas [9]. En un esfuerzo por mejorar la comodidad, la libertad de movimiento y proporcionar una IAC más natural para los perros, este trabajo presenta una IAC basada en vocalizaciones. El sistema está diseñado para que, cuando un perro de asistencia detecte un evento de urgencia médica de su dueño, ladre con un patrón específico. El sistema es capaz de reconocer los patrones de ladrido para enviar una alarma que llega en forma de notificación a una aplicación móvil instalada en los teléfonos de los cuidadores designados por el usuario [10].



Figura 4. Dispositivo ubicuo para alarmas médicas activado mediante ladridos. Esta es una interfaz natural animal-computadora.

La propuesta de IAC está conformada por tres elementos principales: el modelo de aprendizaje automático YAMNET, un micrófono omnidireccional y una tarjeta de desarrollo Nvidia Jetson nano, los cuales se pueden apreciar en la Figura 4. El sistema funciona en modo continuo, escuchando los sonidos del entorno mientras el modelo YAMNET los clasifica. YAMNET es capaz de reconocer una amplia gama de sonidos, incluyendo los ladridos de perros, y cuando detecta la secuencia específica de ladridos, activa una alarma y envía una notificación a los cuidadores. Aunque pueda parecer invasivo, las grabaciones de audio recopiladas nunca salen del dispositivo y todo el procesamiento necesario ocurre en la propia tarjeta de desarrollo Nvidia Jetson nano, lo que asegura la privacidad del usuario.

Fase de selección y entrenamiento de perros de asistencia

El protocolo utilizado para este trabajo fue aprobado por un comité institucional de Bioética. Este comité desempeña un papel crucial en el ámbito académico al supervisar y garantizar que la investigación y las prácticas relacionadas con los seres vivos se realicen de manera ética y respetuosa. Dicho comité está integrado por un grupo interdisciplinario de profesionales, quienes eligen a un presidente, un secretario y cuatro vocales, uno de los

© 2023 - Sociedad Mexicana de Inteligencia Artificial ISSN 2007-0691

cuales, generalmente, pertenece a la comunidad afectada (en el caso de estudios en humanos). Se reúnen aproximadamente cada dos meses para evaluar los protocolos experimentales recibidos. En el caso de nuestro protocolo de evaluación de perros, el comité se enfocó en diversos aspectos. En primer lugar, se vigiló que el tamaño de la muestra (número de perros) fuera apropiado,y que los perros involucrados en el estudio se encontraran en condiciones de alojamiento, alimentación y salud adecuados. También se aseguró que los procedimientos fueran necesarios y justificados. El comité observó con agrado el uso de técnicas de entrenamiento basadas en refuerzo positivo y la implementación de medidas para garantizar la comodidad de los perros durante la investigación. Una de las recomendaciones recibidas fue no exceder el tiempo de entrenamiento con cada perro y acordar con los dueños el tipo de reforzador positivo.

En la etapa de evaluación del sistema incluyeron tres perros adultos intactos (sin castrar), sin antecedentes patológicos de importancia, con esquema de vacunación completo y cuyos dueños aceptaron participar en el estudio firmando el consentimiento para que los datos obtenidos pudieran ser usados con fines de investigación. El tipo de tarea aplicada no puso en riesgo su integridad ni comprometió su bienestar. Los perros incluidos se encontraron en adecuadas condiciones de alojamiento, alimentación, mantenían una estrecha relación con sus cuidadores y recibían como estimulación sensorial paseos con correa tres veces por semana durante 40 minutos aproximadamente. Fueron entrenados durante tres semanas en sesiones diarias de 15 minutos, utilizando métodos secuenciales de entrenamiento positivo (reforzamiento con comida), partiendo de lo más fácil a lo más difícil para que aprendieran a ladrar con un patrón específico.

Experimentos y resultados

Se llevaron a cabo varios experimentos para evaluar la eficacia y la versatilidad de la IAC basada en ladridos. En el Experimento 1, los perros seleccionados para la prueba participaron en sesiones de 5 minutos en las que se les estimuló para ladrar de manera simulada ante una emergencia con el patrón aprendido, mientras el dispositivo estaba en modo de escucha. El objetivo de este experimento fue evaluar la detección de casos de activación verdaderos positivos, es decir, que la alarma se activara cuando debía hacerlo. Los resultados de este experimento muestran que el sistema se activó en las ocasiones que debía hacerlo, por lo que no presentó falsos positivos ni falsos negativos. en la Tabla 1 se puede observar la cantidad de ladridos totales detectados y la cantidad de veces que se activó la alarma.

Tabla 1. Resultados del experimento 1.

Experimento 1			
Repetición	Ladridos totales	Alarmas activadas	Alarmas esperadas
1	365	2	2
2	116	2	2
3	94	1	1

Tabla 2. Resultados del experimento 2.

Experimento	Escenario en casa normal con ladridos sintéticos	
Repetición	Ladridos totales	Alarmas activadas
1	184	0
2	207	0
3	373	0
4	225	0

Por otro lado, los experimentos 2 y 3 se llevaron a cabo en condiciones ambientales normales, en una casa habitada mientras sus habitantes realizaban sus actividades cotidianas. El objetivo de estos experimentos fue evaluar el sistema en condiciones que no debían activar la alarma. En el experimento 2, el sistema no se activó y únicamente detecto sonidos relacionados a un entorno doméstico común. Por otro lado, el experimento 3, contó con presencia de ladridos que no significaban un caso de emergencia, los resultados del experimento 3 mostraron que el sistema no activa la alarma cuando se producen ladridos de manera aleatoria, aunque sí es capaz de detectarlos. Los resultados del experimento 3 pueden consultarse en la Tabla 2.

Los resultados obtenidos demuestran que es viable utilizar vocalizaciones de perro como un método de IAC para un sistema de alarma de emergencia médica que además sea respetuoso con la privacidad de los usuarios. Este tipo de sistema podría ser beneficioso para personas mayores o con alguna discapacidad que necesiten asistencia en caso de emergencia.

Conclusiones

El estudio de la conducta natural de los animales, y sus procedimientos están considerados en los lineamientos de la bioética en términos de garantizar el bienestar de los animales involucrados en investigaciones científicas. De manera similar, en la inteligencia artificial es importante considerar los aspectos éticos en el diseño y desarrollo de sistemas, incluyendo la privacidad de los usuarios.

En este trabajo se propone un sistema de interfaz animal-computadora basado en los ladridos de los perros, con el objetivo de que éstos últimos puedan utilizar sus vocalizaciones para interactuar con una computadora con diferentes propósitos. En el presente proyecto, se utilizó para activar un sistema de alarma que se dirige

principalmente a personas con urgencias médicas como pacientes con epilepsia, diabetes o adultos mayores.

Esta investigación aborda temas que pueden ser preocupantes desde el punto de vista ético, ya que involucra la participación de animales vivos y de sistemas que pueden ser altamente invasivos en cuanto a la privacidad de las personas. Sin embargo, es importante señalar que se tuvieron en cuenta las consideraciones éticas para ambos temas. Por un lado, los perros incluidos en el estudio fueron seleccionados por expertos en etología y entrenamiento animal, teniendo en cuenta sus características individuales para el diseño de los experimentos realizados. Por otro lado, se seleccionaron elementos de hardware para el desarrollo del dispositivo con la finalidad de ejecutar los modelos de inteligencia artificial en el mismo dispositivo, lo que significa que los datos de audio recopilados no pueden ser escuchados por terceros.

Los resultados de los experimentos demuestran que es posible utilizar los ladridos de los perros como método de interacción animal-computadora. Esto podría impulsar el desarrollo de nuevas aplicaciones en el trabajo con perros, mientras se respeta el derecho a la privacidad de los usuarios y se brinda un trato digno a los animales involucrados.✷

REFERENCIAS

1. Lundqvist, M., Carlsson, P., Sjödahl, R., et al. (2017). Patient benefit of dog-assisted interventions in health care: a systematic review. *BMC complementary and alternative medicine*,17, 1-12.
2. Besser, A. y Zeigler-Hill, V. (2014). Positive personality features and stress among first-year university students: Implications for psychological distress, functional impairment, and self-esteem. *Self and identity*, 13(1), 24-44.
3. Blackwell, E.J., Twells, C., Seawright, A., et al. (2008). The relationship between training methods and the occurrence of behavior problems, as reported by owners, in a population of domestic dogs. *Journal of Veterinary Behavior*, 3(5), 207-217.
4. Rooney, N. J. y Cowan, S. (2011). Training methods and owner–dog interactions: links with dog behaviour and learning ability. *Applied Animal Behaviour Science*, 132(3-4), 169-177.
5. Fugazza, C. y Miklósi, Á. (2014). Should old dog trainers learn new tricks? The efficiency of the Do as I do method and shaping/clicker training method to train dogs. *Applied Animal Behaviour Science*, 153, 53-61.
6. Blanchard, A. y Taddeo, M. (2023). The Ethics of Artificial Intelligence for Intelligence Analysis: a Review of the Key Challenges with Recommendations. *Digital Society*, 2(1), 1-28.
7. Bibal, A., Lognoul, M., De Streel, A., et al. (2021). Legal requirements on explainability in machine learning. *Artificial Intelligence and Law*, 29, 149-169.
8. Mancini, C. (2011). Animal-computer interaction: a manifesto. *Interactions*, 18(4), 69-73.
9. Rodríguez-Vizzuett, L., Espinosa-Curiel, I.E. y Pérez-Espinosa, H. (2023). Digital Technology Supporting the Remote Human-Dog Interaction: Scoping Review. *Animals*, 13(4), 699-721.
10. Abrego-Ulloa, E.R., Aguilar-Lazcano, C.A., Pérez-Espinosa, H., et al. (2022). Towards a monitoring and emergency alarm system activated by the barking of assistant dogs. *In Proceedings of the Ninth International Conference on Animal-Computer Interaction*. Newcastle upon Tyne, United Kingdom, 1-10.

SOBRE LOS AUTORES



Carlos Alberto Aguilar-Lazcano es Doctor en inteligencia artificial por la Universidad Veracruzana (2021) y maestro en ingeniería electrónica por parte del Instituto Tecnológico de Orizaba (2015). Su principal línea de investigación es el análisis de bioseñales para reconocer gestos. Es investigador posdoctoral en la Unidad Nayarit del CICESE (CICESE-UT3) y es miembro del Sistema Nacional de Investigadores.



Humberto Pérez-Espinosa recibió el grado de Licenciado en Ciencias Computacionales en la Benemérita Universidad Autónoma de Puebla (BUAP) en 2004, el grado de Maestro en Ciencias Computacionales en 2006, y el grado de Doctor en Ciencias Computacionales en el Instituto Nacional de Astrofísica, Óptica y Electrónica (INAOE) en 2013. Sus intereses de investigación incluyen el aprendizaje de máquina, computación afectiva, computación paralingüística y el análisis inteligente de audio. Entre 2014 y 2023 fungió como investigador en CICESE-UT3. Actualmente se desempeña como investigador en el INAOE y es miembro del Sistema Nacional de Investigadores nivel I.



Verónica Reyes-Meza es Psicóloga con Maestría en Diagnóstico y Rehabilitación Neuropsicológica. En 2011 obtuvo su Doctorado en Neuroetología y en 2012 hizo un Postdoctorado en Ciencias Biológicas. Ha realizado estancias de investigación en instituciones nacionales y en el extranjero. Es miembro del SNI Nivel I. Actualmente es profesora de la Maestría y Doctorado en Ciencias Biológicas de la Universidad Autónoma de Tlaxcala. Pertenece a la Society for Neurosciences y a la Animal Behavior Society.

© 2023 - Sociedad Mexicana de Inteligencia Artificial ISSN 2007-0691



Mariel Urbina-Escalante es Psicóloga y Maestra en Ciencias Biológicas por la Universidad Autónoma de Tlaxcala, cuenta con un curso certificado de Adiestramiento Canino Profesional y Modificación de Conducta Canina por el Instituto de Formación y Capacitación para el Bienestar Animal (INCABA), así como certificación en el curso "Ansiedad por Separación Programa de 1era. Sesión", por Separation Anxiety Dogs Assoc. Actualmente se encuentra cursando el 1er. Semestre del Doctorado en Ciencias Biológicas en la Universidad Autónoma de Tlaxcala.



Imagen tomada de: https://diariodegastronomia.com/inteligencia-artificial-medir-bienestar-animal/.

© 2023 - Sociedad Mexicana de Inteligencia Artificial ISSN 2007-0691

ARTÍCULO ACEPTADO

Archivística e Inteligencia Artificial: posibilidades interdisciplinares y desarrollos responsables

Luis Miguel García Velázquez y Marisol Flores Garrido

En los últimos años los algoritmos de Inteligencia Artificial (IA) se han incorporado a la ciencia archivística en diversos ámbitos de aplicación; por ejemplo, la identificación automática de datos personales y sensibles dentro de documentos en entornos digitales, para su protección al momento de ponerlos a disposición para consulta [3]. Otra línea de trabajo es la automatización de procesos de valoración de documentos para determinar su conservación (temporal o permanente) o eliminación, reduciendo los tiempos al trabajar con grandes volúmenes. Una tercera aplicación es el desarrollo de herramientas automáticas para la descripción y organización de colecciones digitales que contienen imágenes, audios o videos entre sus componentes.

De forma paralela, el auge de la ciencia de datos ha permitido la exploración de grandes colecciones documentales con la finalidad de producir conocimiento nuevo; como ejemplo podemos referir las herramientas de análisis literario y de imágenes que permiten el estudio masivo de obras artísticas, amplificando la capacidad de investigación en humanidades al incluir en un mismo ejercicio grandes cantidades de registros digitales [10]. También en este rubro de indagación se encuentran los esfuerzos de investigación en salud para estudiar extensas bases de expedientes médicos, con intención de mejorar el entendimiento y la atención de padecimientos específicos a partir del reconocimiento de patrones en pacientes [2].

La ciencia archivística establece principios y procedimientos que promueven la preservación y la accesibilidad a la información contenida en conjuntos documentales, organizados y resguardados por personas e instituciones, que atienden a preocupaciones de interés ético como la transparencia, la rendición de cuentas y la integridad documental.

La creciente aplicación de la IA en los archivos ha motivado la articulación de proyectos de investigación con enfoque interdisciplinar, que promuevan desarrollos y usos responsables para garantizar la disponibilidad y la accesibilidad continuada de documentos en archivos confiables; este es el caso de la iniciativa multinacional InterPARES Trust AI, que incentiva el intercambio entre las herramientas archivísticas y los modelos de desarrollo de IA, a fin de establecer pautas de acción que puedan seguirse de forma sistemática [5]. Estas recomendaciones ofrecen rutas de actuación y reflexión que, si bien son de gran relevancia para la gestión documental, también proponen comportamientos éticos que se extienden al diseño e implementación de modelos con IA en otras áreas.

En la siguiente sección describiremos el esquema de paradatos propuesto por InterPARES Trust AI como una estrategia archivística para promover desarrollos de IA responsables, transparentes y confiables. Posteriormente, discutiremos sobre el potencial crítico de las primeras fases del aprendizaje automático, a fin de mostrar una aplicación concreta durante el análisis exploratorio de expedientes médicos, como parte del desarrollo de un modelo de aprendizaje supervisado para investigación en salud.

Paradatos como estrategia de preprocesamiento responsable

Una de las recomendaciones establecidas por los grupos de trabajo del proyecto InterPARES Trust AI es la documentación de paradatos en el desarrollo de algoritmos de IA, es decir, de una capa de información adicional que dé cuenta del propósito y del proceso de producción de un conjunto de datos, de la evolución de su entendimiento e interpretación, y del desempeño en las distintas fases de procesamiento, con fines de evaluar y refinar su funcionamiento [4]. En particular, se desea que los paradatos comuniquen los momentos de incertidumbre y decisión de las distintas etapas de análisis, con la intención de que sea posible en el futuro ampliar las posibilidades de interpretación, de manera pertinente e informada.

Adicionalmente, se pretende que la documentación de paradatos ofrezca una plataforma para entender el funcionamiento de los métodos de IA, dirigida también a especialistas en el dominio específico del conjunto de datos, aun cuando su disciplina profesional no sea cercana a la computación (por ejemplo, profesionales de la salud o el derecho). Esta característica le permitirá a una persona experta ampliar de la base de datos para robustecer el análisis y mejorar su interpretación. Una visualización apropiada, por ejemplo, permitirá incorporar una tipología particular en un conjunto de documentos de entrenamiento para una IA que quedaba invisible en los resultados preliminares; o bien eliminar inconsistencias al comparar con la bitácora de un sistema productor de documentos.

La incorporación de un esquema de paradatos favorece la transparencia operativa durante el análisis, robustece la anticipación de consecuencias y posibilita la

© 2023 - Sociedad Mexicana de Inteligencia Artificial ISSN 2007-0691

posterior rendición de cuentas al integrarse en el proceso de documentación durante el diseño, la implementación y la vida útil de un modelo de IA. Distintos enfoques metodológicos se han propuesto para sistematizar esta idea, que incluyen la adición de metadatos -datos acerca de los datos- desde las fases tempranas de un desarrollo de IA; éstos deben reflejar el seguimiento de políticas bien establecidas de recolección en términos de consentimiento, inclusión, colectividad y distribución de poder, transparencia, ética y privacidad [6, 8].

Si bien la etapa de recolección es crucial para la conformación de una base de datos, su construcción como objeto de estudio dentro de un análisis computacional incluye otras fases mucho más técnicas en la etapa de limpieza y preparación de los datos. De acuerdo con Lisa Gitelman, los datos nunca están crudos, sino que siempre están cocinados y las circunstancias de su creación, almacenamiento y transmisión pueden marcar una diferencia importante en las conclusiones que obtengamos al analizarlos [7].

Cada conjunto de datos responde a una selección de tecnologías y procesos que obedecen a un propósito específico, a los recursos disponibles y que, a través de su implementación, inducen una selección de información implícita que puede conducir a sesgos e injusticias. La fase conocida como *preprocesamiento*, que prepara los datos como entrada de un algoritmo de aprendizaje supervisado, es en realidad la prolongación de un procedimiento mucho más amplio que comenzó desde la recolección de datos que se almacenaron en una base.

Como señalan Jo y Gebru, la recolección de datos para análisis cumple con objetivos asociados a las implementaciones y las métricas de desempeño -por ejemplo, conseguir un alto volumen de información-, mientras que la vocación para preservar la memoria y dar acceso a la información incide en que los archivos se consoliden con mayor rigor [8]. Aun así, los archivos no son repositorios neutros de información, sino participantes activos en el génesis y la interpretación de una historia, puesto que en la creación de un archivo siempre hay un ejercicio de selección que define lo que se preserva y lo que queda excluido de la memoria, lo que propicia que la capacidad -y la autoridad- para interpretar los elementos que se almacenan en él se concentre en un reducido grupo de personas [1].

Esta diferencia de objetivos motiva que el acopio de datos en conjuntos para análisis tenga un crecimiento menos regulado -o indiscriminado-, donde la responsabilidad de limpieza y preparación termina por recaer en una persona ajena al proceso de recolección, que quizá no posea información suficiente para establecer criterios apropiados. Si bien es deseable promover la generación de paradatos que se incorporen desde la creación de una base de datos, también es importante implementar estrategias para explorar de manera crítica la colección recibida, a fin de ampliar el entendimiento de los datos durante la fase de preprocesamiento y añadir paradatos durante esa etapa que favorezcan la transparencia en fases sucesivas.

Los archivos son participantes activos en el génesis y la interpretación de una historia.

Análisis exploratorio y crítico de un conjunto de datos

El análisis exploratorio de datos (AED) es un paso fundamental en cualquier tarea relacionada con ciencia de datos. El AED tiene como objetivo examinar, muchas veces mediante visualizaciones, las características del conjunto de datos con el que se está trabajando y detectar patrones, anomalías o inconsistencias que podrían requerir limpieza y preprocesamiento [9].

Tradicionalmente se consideran como tareas básicas de AED la limpieza, la imputación, la transformación y la visualización. La limpieza permite remover o corregir errores e inconsistencias en los datos, mientras que la imputación completa datos faltantes con valores apropiados, a partir de métodos estadísticos. En la transformación los datos se trasladan a una representación apropiada para el análisis y el modelo que se desea utilizar, que puede discretizar y normalizar atributos o incrustar los datos originales en un espacio vectorial determinado. Finalmente, la visualización genera elementos gráficos que permitan percibir la relación entre los distintos atributos e identificar patrones o tendencias.

Además de los métodos computacionales y matemáticos anteriormente descritos, el AED debe implicar también un análisis que permita valorar si los datos son apropiados para el modelo que desea utilizarse y, si es el caso, las limitaciones que se esperan al hacerlo. Esta evaluación, de carácter cualitativo, debe buscar explícitamente fuentes potenciales de sesgo en los datos que podrían llevar a conclusiones equivocadas. Como parte de este proceso es conveniente realizar una observación cuidadosa de los datos, con intención de identificar instancias normales, sorprendentes y confusas; para ello es posible apoyarse de herramientas visuales y generar representaciones gráficas que habiliten la inspección por parte de especialistas de otras disciplinas.

En esta fase se debería incluir toda la información que se tiene sobre la construcción del conjunto de datos,

© 2023 - Sociedad Mexicana de Inteligencia Artificial ISSN 2007-0691

incluyendo diagramas sobre los procesos a los que han sido sometidos y una contextualización sobre los usos que se les ha dado previamente. Aunque dicha información no esté necesariamente contenida en los datos que alimentarán al modelo, es crucial para pensar críticamente sobre el contexto, el alcance y el significado que pueden atribuirse a los resultados del análisis que se lleve a cabo.

La información recopilada durante esta fase, que puede incluir hojas de registro de metadatos, diagramas de proceso, bitácoras de limpieza y preprocesamiento, entrevistas y visualizaciones gráficas de datos, también formará parte del conjunto de paradatos.

A continuación, presentamos un caso de estudio donde las tareas no tradicionales del AED y el registro de paradatos permitieron asumir una ética profesional para el desarrollo de modelos de IA en el área de investigación en salud.

Caso de estudio: Hipertensión en pacientes diabéticos

El personal médico de las Unidades de Medicina Familiar del Instituto Mexicano del Seguro Social registra la evolución de cada paciente con diabetes mellitus tipo 2 en el Sistema de Información de Medicina Familiar (SIMF), plataforma que gestiona la integración del expediente electrónico. Con la intención de identificar factores que inciden en el desarrollo de complicaciones de este padecimiento, un equipo de investigación interinstitucional[1] ha explorado distintos métodos de aprendizaje automático para analizar los datos históricos extraídos del SIMF que corresponden a pacientes en Michoacán desde 2006, fecha en que dio inicio el registro [2]. La base cuenta con el registro de más de 350,000 derechohabientes.

La hipertensión arterial es un padecimiento crónico que se caracteriza por el aumento de la presión con la que el corazón bombea la sangre a las arterias para que circule por el cuerpo. La presión arterial, que es la fuerza que la sangre ejerce contra las paredes arteriales, se registra con dos números: presión sistólica y presión diastólica. El primer número representa la presión causada cuando el corazón se contrae y empuja la sangre hacia afuera; el límite superior saludable para esta presión es 120 mmHG. El segundo número (presión diastólica) es la presión que ocurre cuando el corazón se relaja y se llena de sangre; esta presión debe estar por debajo de 80 mmHg.

Durante el análisis exploratorio de datos se quería establecer si sería posible encontrar relaciones entre el diagnóstico de diabetes y el de hipertensión, así como sobre la efectividad de los tratamientos. Como parte del AED se realizaron imágenes descriptivas como las que se muestran en las Figuras 1 y 2. La intención era generar paradatos que facilitaran la exploración conjunta desde las perspectivas médicas y de cómputo. Para interpretar la visualización de los valores de presión por paciente fue necesario documentar en los paradatos las reglas de imputación cuando pacientes faltaron a citas médicas, que reproducen los valores de la última medición registrada.

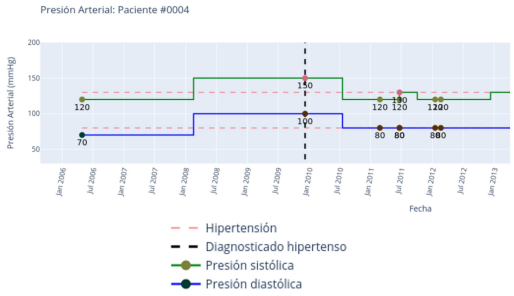


Figura 1. Visualización del historial de valores para la presión, sistólica y diastólica, del paciente #4, incluida en los paradatos del AED.

Al producir las primeras visualizaciones se observa que el tiempo que transcurre entre que alguien es diagnósticado con diabetes y desarrolla hipertensión, encontramos como patrón altamente dominante que los diagnosticos eran simultáneos (Figura 2). La conversación con el personal médico nos previno de que esta era una anomalía que imposibilitaba cualquier posibilidad de análisis sobre la relación entre ambos padecimientos.

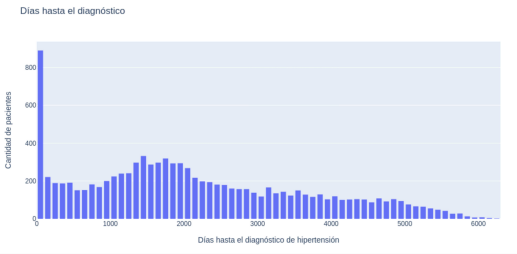


Figura 2. Visualización del tiempo transcurrido entre el diagnóstico de hipertensión y el de diabetes, incluida en la primera iteración de los paradatos del AED.

Como resultado de este primer análisis, se propuso una indagación sobre la implementación el expediente electrónico, realizada entre 2003 y 2006, que hizo notar que en el proceso de transición entre el expediente en papel y el electrónico no hay garantía sobre la homologación de las fechas de diagnóstico entre ambos soportes. A partir de esta información, se determinó ejecutar un proceso de limpieza focalizado para no incluir el conteo

[1] En el proyecto participa personal de investigación de la Universidad Michoacana de San Nicolás de Hidalgo (UMSNH), la Universidad Nacional Autónoma de México (UNAM), el Instituto Mexicano del Seguro Social (IMSS) y la empresa Amphora Health.

de las personas que en el día uno tienen ambas enfermedades, pues no existe un medio de verificación para identificar con precisión cuántas personas aparecen así por haber llegado con ambos padecimientos y cuántas por efecto del registro. Esta decisión quedó documentada también en los paradatos del AED, junto con una nueva visualización que incorpora este criterio (Figura 3).

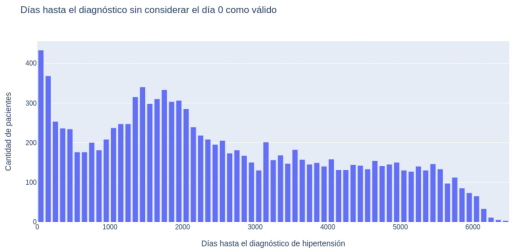


Figura 3. Visualización del tiempo transcurrido entre el diagnóstico de hipertensión y el de diabetes, incluida en los paradatos de la segunda iteración del AED. A esta figura se adjuntan los criterios de ajuste realizados en esta iteración, como paradatos indispensables para interpretarla.

Conclusiones

Cada colección de datos es creada en un contexto particular y, por tanto, responde a intereses específicos que han motivado su creación. Así, un ejercicio de cuantificación, por más cuidadoso que sea, no produce conocimiento pertinente sin una debida comprensión de las condiciones en las que el conjunto de datos se produjo. Modelar de la mejor forma posible requiere de una indagación iterativa entre el entendimiento de los datos y sus relaciones, para generar herramientas matemáticas que reflejen el contexto donde esas colecciones fueron recopiladas; al cerrar ese ciclo, conviene verificar la pertinencia de inferencias e interpretaciones que resultan del análisis con una mirada cercana al contexto de aplicación. A partir de esta valoración en espiral, será posible realizar ajustes sobre las herramientas propuestas y sus posibles usos.

En el caso que aquí se presenta, la exploración crítica de datos nos permitió ampliar el entendimiento de los datos recibidos y nos habilitó para documentar criterios de decisión que sumen a la transparencia en el análisis, faciliten la futura interpretación de resultados y promuevan el ejercicio de una ética profesional. Nos parece importante promover que los grupos de investigación adopten y desarrollen herramientas metodológicas situadas para el análisis exploratorio de datos desde una perspectiva crítica, con la intención de adquirir un mayor grado de confianza sobre el potencial y las limitaciones para hacer minería con ellos.

Dado que el objeto central de este estudio son expedientes médicos sistematizados dentro de una institución a través de prácticas específicas de documentación, retomamos aportes de la ciencia archivística para incorporar el esquema de paradatos como una herramienta de comunicación en un equipo interdisciplinario que reúne especialistas de las áreas de cómputo y salud. En particular, la implementación de este esquema a través de visualizaciones aportó beneficios directos hacia el refinamiento del modelo computacional, que a su vez generaron documentación útil para posteriores etapas de desarrollo e interpretación de resultados.

Finalmente, insistimos en que la ética profesional para el análisis de datos con herramientas computacionales requiere de una constante reflexión sobre la idoneidad de los modelos matemáticos que se producen, anticipando su funcionamiento en un contexto específico. El desarrollo de herramientas sistemáticas y flexibles para el AED con enfoque crítico y la documentación de paradatos promoverán la transparencia, la confiabilidad y la rendición de cuentas en el desarrollo de algoritmos de IA[2]. ✷

REFERENCIAS

1. Agostinho, D., D'Ignazio, C., Ring, A., Thilstrup, N. y Veel, K. (2019). Uncertain Archives: Approaching the Unknowns, Errors and Vulnerabilities of Big Data through Cultural Theories of the Archive. *Surveillance and Society*, 17(3/4), 422-441.

2. Alvarez, C., Arcos, C., Figueroa, K., Flores, M., García, L., Gómez, A., López, A., Mercado, J., Rodríguez, M. y Tripp, J. (2022). Estudio longitudinal para el desarrollo de modelos predictivos de complicaciones crónicas de la diabetes mellitus tipo 2. *Komputer Sapiens*, 14(3), 6-11.

3. Barnard, A., Bernal, Y., Cuéllar, R., Escoto, C. y García-Velázquez, L. (2022). Inteligencia artificial en los archivos. Consideraciones de diseño e implementación. *Tábula*, 25, 41-59.

4. Davet, J., Hamidzadeh, B. y Franks, P. (2023). Archivist in the machine: paradata for AI-based automation in the archives. *Archival Science*, 1-21.

5. Duranti, L., Abdul-Mageed, M., Hofman, D. y Sullivan, P. (2022). I Trust AI, the latest InterPARES research project. *Anuario Escuela de Archivología*, 13, 36-55.

6. Gebru, T., Morgesnstern, J., Vecchione, V., Jennifer, W., Hannah, H. y Crawford, K. (2021). Datasheets for datasets. *Communications of the ACM*, 64(12), 86-92.

7. Gitelman, L. (2013). Raw data is an oxymoron. *MIT press.*

8. Jo, E. y Gebru, T. (2020). Lessons from archives: Strategies for collecting sociocultural data in machine learning. *Proceedings of the 2020 conference on fairness, accountability, and transparency*, 306-316.

9. Myatt, G. y Johnson, W. (2014). Making Sense of Data I: A Practical Guide to Exploratory Data Analysis and Data Mining. *John Wiley & Sons.*

10. Rodríguez-Ortega, N. (2020). Image processing and computer vision in the field of art history. En *Kathryn Brown (Ed.), The Routledge Companion to Digital Humanities and Art History, New York, USA, Routledge*, 338-357.

[2]Este artículo se ha beneficiado de las reflexiones realizadas en el proyecto PE112523 *Casos de estudio en el aula: desarollos didácticos con metodologías feministas.* Quienes escriben este artículo agradecen a UNAM-DGAPA-PAPIME por el apoyo recibido.

© 2023 - Sociedad Mexicana de Inteligencia Artificial ISSN 2007-0691

SOBRE LOS AUTORES



Marisol Flores Garrido estudió Matemáticas Aplicadas en la Universidad Autónoma de Coahuila. Posteriormente obtuvo una Maestría en Cómputo Científico en la Universidad de British Columbia y un Doctorado en Ciencias Computacionales en el Instituto Nacional de Astrofísica, óptica y Electrónica. Es profesora en la Escuela Nacional de Estudios Supe- riores, Unidad Morelia (UNAM). Sus intereses de investigación se centran en el aprendizaje automático, el análisis de redes y la intersección entre la inteligencia artificial y las humanidades.



Luis Miguel García Velázquez es Ingeniero en Sistemas Computacionales graduado con Mención de Excelencia por el Tecnológico de Monterrey. Obtuvo el grado de Maestría por la Universidad Michoacana de San Nicolás de Hidalgo. Se doctoró en Ciencias Matemáticas por la Universidad Nacional Autónoma de México. Actualmente es profesor en la Escuela Nacional de Estudios Superiores, Unidad Morelia (UNAM). Sus intereses de investigación incluyen la modelación matemática y computacional para problemas interdisciplinares, la archivística computacional y las humanidades digitales. Desde el año 2020 colabora como investigador asociado al proyecto InterPARES Trust AI



Imagen tomada de: https://www.colaberry.com/tag/artificial-intelligence/.

© 2023 - Sociedad Mexicana de Inteligencia Artificial ISSN 2007-0691

ARTÍCULO ACEPTADO

Aplicaciones y Desafíos de la Inteligencia Artificial en la Aeronáutica

Romina Ruiz Macías, Rodolfo Garcia-Rodriguez y Luis Enrique Ramos-Velasco

Introducción

Imaginarse cómo será el mundo en un futuro ha sido una inquietud del ser humano desde sus inicios, siendo plasmada en libros, series o películas de ciencia ficción. De tal forma, no es casualidad que varios de los adelantos científicos y tecnológicos que tenemos hoy en día fueron descritos o imaginados previamente en la ciencia ficción. Por ejemplo, aunque el término de robótica se hizo popular con las obras de Isaac Asimov como *Yo, Robot* o *La Fundación*, donde se habla de un robot positrónico para controlar robots por medio de computadoras. Fue en la obra de Karel Čapek en 1920 donde se introduce el término *Robot*[1] por medio de su obra Rossum's Universal Robots. La obra menciona crear máquinas dotadas de cierta inteligencia que realicen el trabajo de los seres humanos. Ésta sería una de las primeras veces que se menciona la creación de máquinas inteligentes. La idea de tener máquinas inteligentes con capacidades de aprendizaje fue tomando forma debido al desarrollo de varios elementos en el siglo pasado. En este aspecto se destacan los trabajos de John Von Newman y Charles Babbage, que dieron paso a las computadoras que conocemos hoy en día, la aparición del transistor como elemento fundamental para la miniaturización y el almacenamiento de datos. También son relevantes los trabajos de Turing, y en especial de John McCarthy, Marvin Minsky y Claude Shannon, donde empiezan a hablar de cómo el aprendizaje puede llevarse a cabo por una máquina, dando origen a la Inteligencia Artificial, IA.

Durante las últimas décadas, en conjunto con el desarrollo de la tecnología, la IA se ha ido convirtiendo en una herramienta cada vez más importante en nuestra vida diaria. Sus capacidades de aprendizaje integran mecanismos como percepción, predicción, planificación y control; ésto en conjunto ha hecho que influya gradualmente en la forma de convivencia y comunicación de la sociedad. Más allá de los impactos positivos o negativos que tiene la IA en la vida diaria, muchos actores, incluso organizaciones como la UNESCO, han planteado la necesidad de definir los principios generales que guíen su comportamiento hacia el respeto a las costumbres de cada país. Evitando que estos principios impidan avanzar con su innovación y desarrollo. Un claro ejemplo lo vemos con la aparición en Noviembre de 2022 de ChatGPT de la empresa OpenAI[2], la cual tiene ciertas restricciones en algunos temas. Sin embargo, estas restricciones llegan a ser definidas bajo los valores y principios de sus propietarios, muchas veces empresas privadas, y no en un marco de valores y principios básicos que beneficien a la sociedad. Por lo tanto surge una pregunta de no fácil respuesta, ¿hasta qué punto o bajo qué criterios una empresa privada que ha destinado recursos humanos y económicos al desarrollo de la IA debe transparentar el diseño básico de sus algoritmos y de los datos con los que fueron entrenados? En el caso particular de OpenAI, empresa sin fines de lucro que promueve el desarrollo de la IA, hace consciencia sobre este dilema, pudiéndose percibir en su sitio web mediante un mensaje de la Directora de Tecnología, Mira Murati, que expresa que "Los sistemas de IA se están convirtiendo en parte de la vida cotidiana. La clave es asegurarse de que estas máquinas estén alineadas con las intenciones y los valores humanos."

La incorporación de la IA en la aeronáutica ha sido paulatina, principalmente en tareas de apoyo o supervisión. Sin embargo, la inclusión de la IA puede ser desde sus productos y servicios hasta el modelo de negocios que afectaría la certificación, elaboración de normas, y estandarización. Dentro de estas actividades, en particular en aeronáutica, el factor humano, la responsabilidad, la dimensión ética, y social son factores de gran importancia. De tal forma, desde hace varios años ha comenzado una serie de acciones y conversaciones que permitan llegar a un consenso a niveles local y global respecto a los principios sobre los cuales la IA debe basarse para la aeronáutica. Una de las primeras acciones que se han llevado a cabo por algunas agencias, como la Agencia de Seguridad Aérea de la Unión Europea, EASA, en el año 2019, es convocar a un grupo de expertos para dar algunas directrices éticas sobre el uso de la IA que permitan a los usuarios lograr cierta confiabilidad sobre la rendición de cuentas, robustez técnica, seguridad, supervisión, privacidad, gobierno de datos, no discriminación y equidad, y transparencia [1].

Asimismo, organismos mundiales como la Organización de Aviación Civil Internacional, OACI, y EASA, en los últimos años han expuesto los beneficios que tiene la IA para el sector aeronáutico, así como las necesidades

[1] La palabra *Robot* se deriva del sustantivo checo "robota" que significa "trabajo".

[2] Es un programa de computadora basado en IA y procesamiento de lenguaje natural que simula una conversación con humanos por internet.

© 2023 - Sociedad Mexicana de Inteligencia Artificial ISSN 2007-0691

para su adecuada integración. No obstante, existen diversas interrogantes que aún no tienen respuesta respecto al alcance que debería tener la IA en la aeronáutica y en especial cómo regirá la ética. Lo que parece evidente hoy en día es que la idea de que la IA reemplace por completo al ser humano en sus tareas no se vislumbra en un futuro cercano.

Sistemas basados en la IA y la ética

Uno de los objetivos de la IA es diseñar máquinas con ciertas capacidades de aprendizaje, interacción proactiva con el entorno e inferencia. Por otro lado, tenemos el Aprendizaje de Máquinas, ML[3], que es una subcategoría de la IA. Este tipo de aprendizaje se caracteriza por el diseño de algoritmos con capacidades de inferencia y predicción. Entonces, los sistemas de IA son una tecnología basada en algoritmos, encargados de procesar un conjunto de datos que son utilizados para la predicción, correlación y toma de decisiones. Aunque muchas de las aplicaciones de los sistemas de IA son virtuales, como diagnóstico médico, también pueden utilizarse en sistemas físicos, por ejemplo autos autónomos o vehículos aéreos no tripulados. Un gran número de los algoritmos de aprendizaje existentes se basan principalmente en tres tipos [2]:

- **Aprendizaje supervisado.** Estima un mapeo desconocido (dependencia) a partir de los datos de entrada y salida en un sistema, ver Figura 1. La estimación de este mapeo puede llevarse a cabo en línea o fuera de línea conocido también como entrenamiento. Una vez estimado el mapeo, es posible utilizarlo para la predicción de las salidas del sistema a partir de los datos de entrada. El aprendizaje supervisado busca la generalización del conocimiento, la cual requiere una gran cantidad de datos, demandando un gran poder de cómputo.

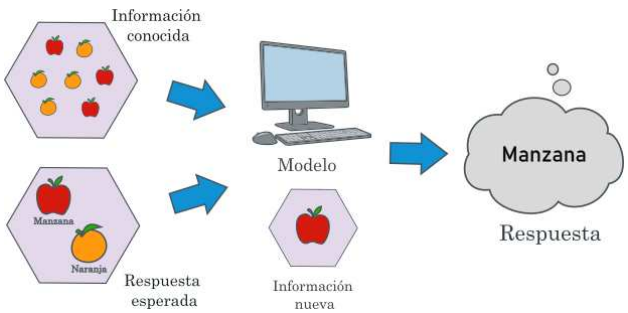


Figura 1. Aprendizaje supervisado.

A pesar de esta desventaja, el aprendizaje supervisado es uno de los tipos de aprendizaje más utilizados en IA. En el año 2006, como resultado de varios trabajos de Yan LeCun, Gerald Tesauro, entre otros, Geoffrey Hinton acuña el término *Aprendizaje Profundo*, DL[4] el cual se basa en transformar los datos de manera significativa con el fin de aprender representaciones útiles de los datos de entrada, es decir, se buscan representaciones ordenadas para múltiples conjuntos de datos complejos. A pesar de la gran cantidad de datos requerida por el DL su desempeño es superior al de algoritmos previos.

- **Aprendizaje no supervisado.** A diferencia del aprendizaje supervisado, no se cuenta con información clasificada ni etiquetada (datos de entrada-salida), ver Figura 2. En este tipo de aprendizaje, uno de los objetivos es descubrir grupos de datos similares dentro un conjunto de datos, lo que se denomina agrupación o *clustering*, o bien determinar cómo se distribuyen los datos en el espacio, lo que se conoce como estimación de densidad.

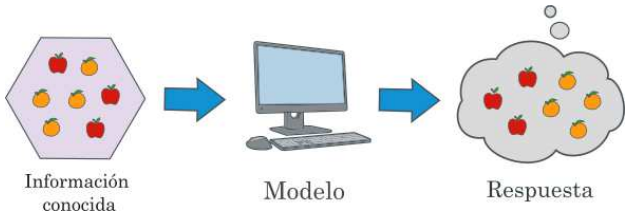


Figura 2. Aprendizaje no supervisado.

- **Aprendizaje reforzado, RL**[5] Este tipo de aprendizaje está inspirado en un mecanismo de aprendizaje natural, como el desarrollado por los seres humanos y animales, donde sus acciones se basan en recompensas o castigos recibidos por las acciones tomadas. Una de las principales características del RL es que el agente, que genera las acciones a realizar, puede equivocarse y no completar la tarea como parte del proceso de aprendizaje. En los últimos años se ha sugerido un tipo de aprendizaje denominado Aprendizaje Curricular, AC, [3] el cual se lleva a cabo de una manera muy similar a cómo aprende el ser humano. Aprende acumulando conocimientos a través de tareas simples en lugar de tratar de abordar una tarea compleja directamente. Los primeros resultados del AC son provisorios, requiriendo de más investigaciones.

[3] Machine Learning, ML.
[4] Deep Learning en inglés.
[5] Reinforcement Learning en inglés.

© 2023 - Sociedad Mexicana de Inteligencia Artificial ISSN 2007-0691

La relación de los sistemas de IA y la ética, a diferencia de lo que uno pudiera pensar, están unidos intrínsecamente. Es decir, desde la concepción de una idea, la investigación y desarrollo por parte de los investigadores, y su comercialización que incluye la parte jurídica de las instituciones; la ética estará presente como un actor fundamental, sobre todo en aplicaciones donde su público objetivo sean los seres humanos. Como en el área de la salud, donde la IA ha tenido gran desarrollo en los últimos años. Así, el uso de los sistemas basados en IA que se utilizan diariamente tienen un impacto en el desarrollo de la sociedad. Esto es así, ya que cambia la forma de entender el mundo y sus principios con los cuales lo explicamos. Si no existe una regulación apropiada, la IA deberá confrontar la experiencia y forma de actuar del ser humano para tomar decisiones, [4]. En aeronáutica esto resulta fundamental, ya que muchas operaciones de seguridad crítica, como el despegue y aterrizaje de aviones, requieren de la intervención humana. A diferencia de otras áreas en aeronáutica, el factor humano juega un rol crucial en sus operaciones más allá de buscar ser productiva y rentable. Por lo que se hace necesario una guía de principios y valores que regulen el uso de sistemas basados en IA, tanto a nivel global como local, en beneficio de todos y evitando la discriminación o desigualdad. En la siguiente sección mencionaremos cómo algunas áreas de la aeronáutica aplican la IA y su relación con la ética.

La relación de los sistemas de IA y la ética, a diferencia de lo que uno pudiera pensar, están unidos intrínsecamente.

Un trabajo conjunto: la ética y la IA en Aeronáutica

Aunque la aeronáutica se ha caracterizado como una industria a la vanguardia de la innovación. En la actualidad no existe una hoja de ruta acerca de la implementación y alcances de la IA en la industria aeronáutica a nivel global. En los últimos años, las agencias de varios países han comenzado a discutir acerca de los beneficios y riesgos de incluir sistemas basados en IA en sus diferentes programas, así como los medios de evaluación de su impacto con el fin de promover su fiabilidad. Una muestra es la EASA, que dio a conocer su visión al año 2035 de cómo incorporar la IA en vuelos autónomos. Si bien estas etapas no son aplicables a todas las áreas de la aeronáutica, nos dan una visión de sus posibles beneficios y limitaciones. En una primera etapa, se proponen escenarios de *asistencia*, hasta que gradualmente sea posible que las operaciones requieran un solo piloto en conjunto con el uso de copilotos virtuales. En una segunda etapa, se proponen vuelos donde exista una *colaboración hombre-máquina*, es decir, vuelos semi-autónomos bajo la supervisión del ser humano. Finalmente, en una tercera etapa se pretende tener un transporte aéreo comercial autónomo.

Cómo y porqué se desean incluir sistemas inteligentes en la aeronáutica, se debe al nivel de complejidad requerida en diversas tareas, principalmente ejecutadas por el ser humano. Por lo que la ética es trascendente en la realización de estas tareas, ya que la aeronáutica se basa en estándares estrictos con el fin de preservar la integridad del ser humano. Dentro de las áreas donde se aplica la IA, podemos mencionar las siguientes:

1. **Seguridad y Protección.** La seguridad es un elemento crítico y de continua evolución en la industria aeronáutica, tanto en la aviación comercial como privada. Aunque el desarrollo de sistemas de seguridad basados en IA ha comenzado por varias empresas, su aplicación en casos de uso críticos de seguridad, como vuelos autónomos o control de tráfico aéreo se ha tomado con mucha cautela. Deben existir mecanismos que evalúen constantemente las vulnerabilidades y daños no deseados del sistema para ser prevenidos y eliminados. Entendiendo como casos críticos de seguridad cualquier sistema que tenga un comportamiento no deseado y que podría ser extremadamente costoso, poniendo en riesgo a los seres humanos y el medio ambiente. Por ejemplo, un sistema de IA completamente autónomo que sea responsable de gestionar los despegues y aterrizajes de aviones. Estos sistemas se caracterizan por pasar extensas y costosas pruebas, evaluaciones, verificaciones y validaciones antes de estar certificados para su uso. Muchos expertos consideran que no se prevé la aplicación de un sistema 100 % basado en IA en un futuro próximo para estos casos. Más bien, se piensa en un sistema basado en IA apoyando al ser humano en sus tareas donde las decisiones de vida o muerte estén a cargo de un ser humano. Por ejemplo, aprovechando la capacidad de la IA, en especial del DL, de manejar, analizar y obtener correlaciones en grandes cantidades de datos; ha sido posible mejorar la inteligencia y gestión de la seguridad, como en detección de riesgos emergentes, clasificación de riesgos, y priorización de temas de seguridad. Por el otro lado, de manera reciente empresas como Airbus y Boeing han invertido en investigación y desarrollo en sistemas basados en IA con el objetivo de desarrollar eventualmente sistemas autónomos que

© 2023 - Sociedad Mexicana de Inteligencia Artificial ISSN 2007-0691

puedan cumplir con los altos y crecientes estándares de seguridad de la industria. Esta aparente contradicción responde muchas veces a las fuerzas del mercado en las cuales las empresas privadas están inmersas. Es decir, con el fin de reducir costos y aumentar su prestigio, las empresas buscan automatizar ciertos procesos, incluso los casos críticos de seguridad, que permitan minimizar el error humano. Sin embargo, también están consientes que una pérdida catastrófica, donde se vea involucrado un sistema basado en IA, significa una pérdida tanto económica como en la imagen de la empresa, [1], que difícilmente podría superarse. Algunos especialistas incluso recomiendan informar a los usuarios cuando una decisión se basa en IA.

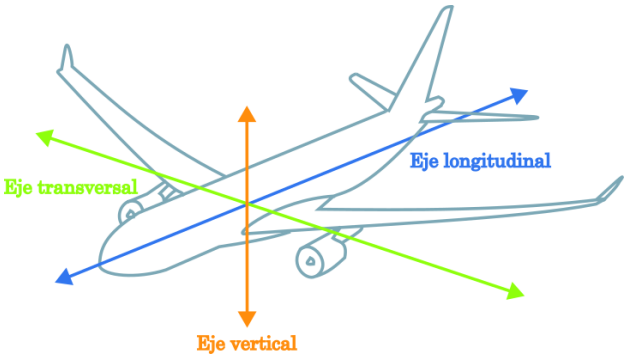


Figura 3. Ejes de un avión.

2. **Navegación.** Es el procedimiento mediante el cual se controla la ruta y posición de la aeronave, procurando mantenerla estable a lo largo de sus ejes longitudinal, vertical y transversal, ver Figura 3. Con el fin de mantener la aeronave en una determinada dirección, sin intervención del ser humano, durante vuelos largos, incluyendo todo lo que sucede durante el tiempo de traslado, como turbulencias, fallas en algún sistema, entre otros, se utiliza lo que se conoce como *piloto automático*[6]. No obstante, al presentarse un desperfecto, el pilotaje debe volverse manual, pues es aquí dónde el piloto automático se ve limitado, ya que no es capaz de tomar decisiones. En casos como el anterior, un sistema basado en IA debiese ser claro y transparente acerca de sus predicciones o decisiones específicas que garanticen la seguridad de los pasajeros. Asimismo, en caso de siniestros deben ser transparentes los códigos o conjuntos de datos utilizados. Los aviones modernos cuentan con una computadora que coordina toda la aviónica[7] del denominado Sistema de Gestión de Vuelo[8], ver Figura 4. Como parte de la computadora se tiene incorporado el piloto automático, que debe ser programado por los pilotos en cabina, previo al despegue, considerando el destino, las rutas que se pueden utilizar y el tráfico aéreo. Se espera que los sistemas basados en IA tendrán un alto potencial en la Gestión del Tráfico Aéreo[9]. Específicamente, en áreas que involucran la toma de decisiones bajo incertidumbre, por ejemplo, detección y resolución de conflictos, la predicción con información limitada, y la predicción de trayectoria, por mencionar algunos. Se considera, adicionalmente, que el reconocimiento de voz puede ser de utilidad como una herramienta de seguridad para detectar errores de colisión, y sincronización de trayectorias. En contraste, a pesar de la existencia de varios desarrollos en los sistemas de navegación aérea, no se ha estimado un consenso acerca de cómo deben ser presentados y certificados por las agencias correspondientes.

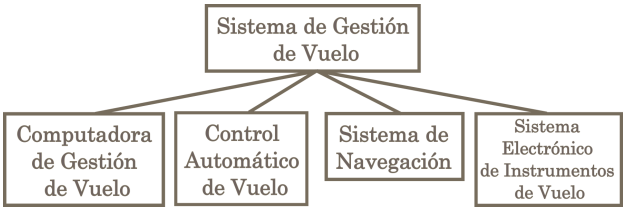


Figura 4. Subsistemas del FMS.

3. **Mantenimiento.** Es una tarea, tal vez de las más importantes, que tiene como objetivo la preservación de la integridad y operabilidad de la aeronave. Los encargados de estas actividades son los talleres aeronáuticos, que diseñan sus propios programas de acuerdo al modelo de la aeronave, manuales brindados por el fabricante, tiempo de vida de los componentes de la aeronave, regulaciones aeronáuticas del país, horas de vuelo y uso específico de la aeronave. El mantenimiento aeronáutico se divide en dos tipos: a)*Mantenimiento no programado:* acciones pensadas para detectar algún mal funcionamiento, falla, avería o daño imprevisto que comprometa las operaciones de la aeronave; b)*Mantenimiento programado:* sigue un plan de tareas específicas que son periódicas con base en las condiciones y necesidades de cada aeronave. Una

[6] Autopilot en inglés.
[7] Electrónica de un avión.
[8] Flight Management System, FMS, en inglés.
[9] Air Traffic Management, ATM, en inglés.

© 2023 - Sociedad Mexicana de Inteligencia Artificial ISSN 2007-0691

de las aplicaciones de la IA es el de reducir los costos de mantenimiento no planificado mediante el monitoreo de aeronaves y la detección de anomalías en tiempo real, alertando de posibles fallas, así como el evitar penalizaciones por demoras, y un aumento en la fiabilidad. Lo anterior permitirá a los gerentes de ingeniería realizar un seguimiento de los gastos, el inventario disponible, así como obtener información sobre las operaciones. En esta área, la empresa Airbus cuenta con un sistema de mantenimiento utilizado por más de cien clientes que monitorea constantemente y transmite fallas o mensajes de advertencia al control de tierra, brindando acceso rápido a documentos de mantenimiento y pasos de solución de problemas.

Conclusiones

El desarrollo de la IA en los últimos años ha permitido avances significativos en diversas áreas, entre ellas la aeronáutica. Hoy en día el principal uso de los sistemas basados en IA es como herramienta de apoyo para realizar diversas tareas no críticas como el mantenimiento predictivo y manufactura, por mencionar algunas. Aunque falta mucho camino por recorrer, se espera que la IA sea capaz de integrarse en diversas tareas de la aeronáutica, respetando los derechos humanos conforme a las legislaciones de cada país, así como de una correcta y objetiva evaluación de su impacto. Los sistemas basados en IA, en especial aplicados a la aeronáutica, deben ser caracterizados por su viabilidad y confiabilidad, minimizando sus riesgos en tareas críticas.

En el caso de aeronaves, la IA ya es capaz de controlar vehículos aéreos no tripulados, aunque su aplicación a sistemas autónomos en vuelos comerciales no se vislumbra en un futuro cercano. Lo que sí se prevé es que dichos sistemas deben ser sistemas de supervisión con una toma de decisiones centradas en el ser humano.

Para garantizar su operabilidad de los sistemas basados en IA deben existir un conjunto de objetivos comunes en lo humano, social, cultural y económico entre las diversas agencias y empresas, así como un conocimiento de posibles repercusiones de su uso. Es decir, buscar un equilibrio entre transparencia, privacidad y protección. Igualmente, deben diseñarse los mecanismos de supervisión, evaluación, y de auditoría que garanticen una rendición de la operación de los sistemas de IA. Finalmente, deben existir políticas para la gobernanza de datos a fin de conocer cuáles son los datos utilizados para el entrenamiento de los sistemas de IA, y de garantizar la privacidad de los datos personales. Todos estos cambios harán replantearse las nuevas competencias de los programas de estudio, en todos los niveles.✷

REFERENCIAS

1. EASA. (2020). AI Roadmap - A human-centric approach to AI in aviation. Recuperado de https://www.bibsonomy.org/bibtex/2e9452555954c3b15b770a2b5c7085aa2/nbienert.
2. Goodfellow, I., Bengio, Y., y Courville, A. (2016). Deep Learning. *MIT Press.*
3. Matiisen, T., Oliver, A., Cohen, T. y Schulman, J. (2020). Teacher–Student Curriculum Learning. *IEEE Transactions on Neural Networks and Learning Systems*, 31(9), 3732-3740.
4. EASA. (2020). Recomendación sobre la ética de la inteligencia artificial. Recuperado de https://unesdoc.unesco.org/ark:/48223/pf0000380455_spa.

SOBRE LOS AUTORES



Romina Ruiz-Macías egresada en el año 2023 del programa educativo de Ingeniería en Aeronáutica de la Universidad Politécnica Metropolitana de Hidalgo. Sus intereses de investigación incluyen: Aviónica, aerodinámica y desarrollo de algoritmos de IA para vehículos aéreos. Actualmente, es supervisora general en Astro Rocket, fomentando el estudio de la ingeniería y ciencia aeroespacial.



Rodolfo García-Rodríguez obtuvo el grado de doctor en Ciencias en Ingeniería Eléctrica en el Centro de Investigación y de Estudios Avanzados del IPN, CINVESTAV, en el año 2005. Actualmente, es profesor investigador en el programa educativo de Ingeniería en Aeronáutica y la Maestría en Ingeniería Aeroespacial de la Universidad Politécnica Metropolitana de Hidalgo. El Dr. García-Rodríguez es miembro del Sistema Nacional de Investigadores Nivel I. Sus intereses de investigación incluyen el control de aeronaves, aplicaciones de la Inteligencia Artificial en robots y aeronaves, principalmente el aprendizaje reforzado y aprendizaje profundo.

© 2023 - Sociedad Mexicana de Inteligencia Artificial ISSN 2007-0691



Luis Enrique Ramos-Velasco recibió el grado de Doctor en Filosofía (Ph.D.) en Ingeniería de Control y Robótica de la Facultad de Ingeniería Eléctrica de la Universidad Técnica Checa de Praga, República Checa en 2002. Actualmente, labora en la Universidad Politécnica Metropolitana de Hidalgo en el programa de Ingeniería Aeronáutica y en el posgrado en Ingeniería Aeroespacial. Cuenta con 31 tesis dirigidas de maestría, 3 de doctorado, más de 100 artículos en revistas y congresos internacionales. Su investigación se centra en: análisis y síntesis de los esquemas de control inteligente para sistemas robóticos, e interfaces hombre-robot. Es miembro del Sistema Nacional de Investigadores, Nivel I.

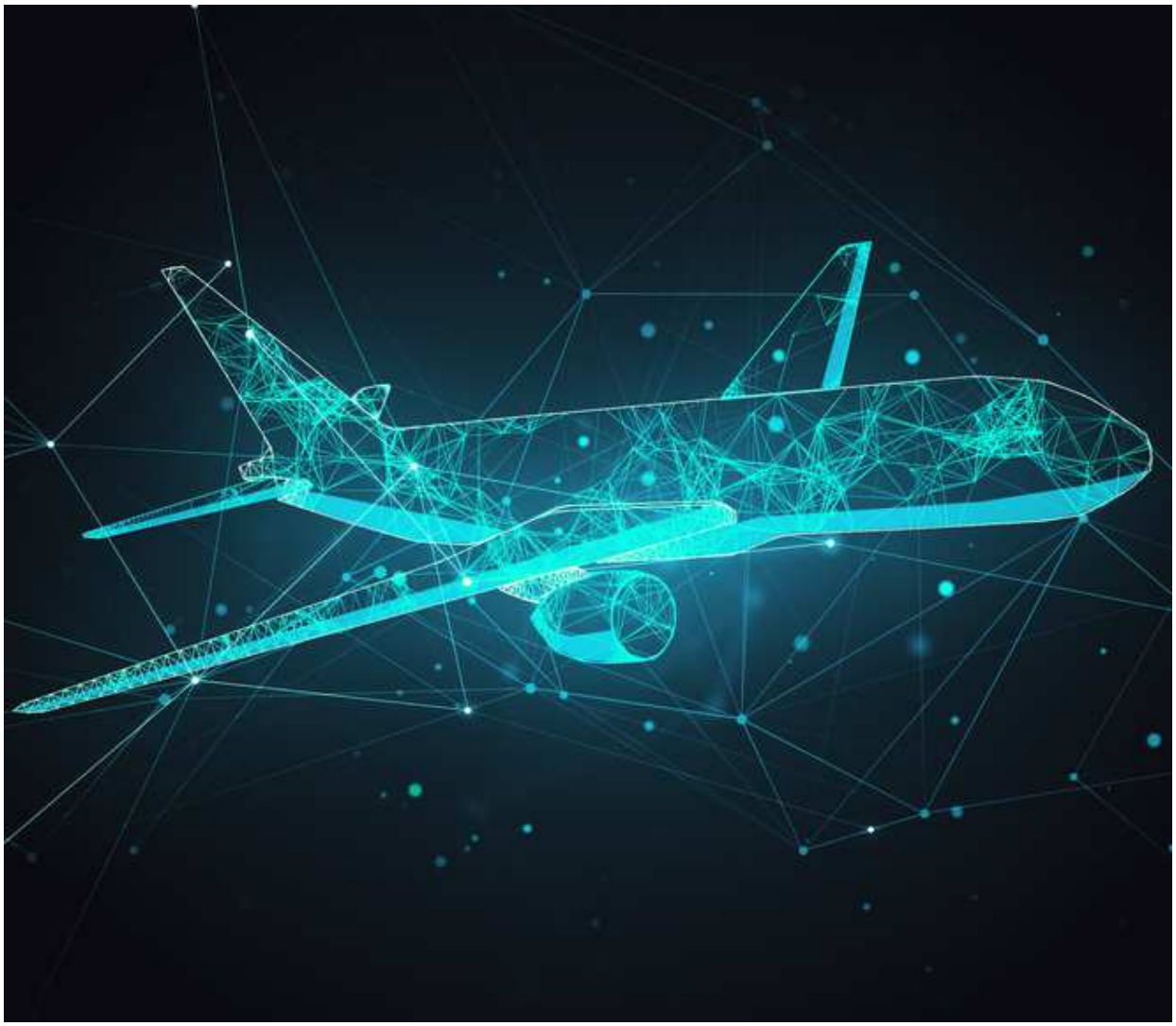


Imagen tomada de: https://www.allerin.com/blog/how-artificial-intelligence-is-transforming-the-aviation-industry.

© 2023 - Sociedad Mexicana de Inteligencia Artificial ISSN 2007-0691

ARTÍCULO ACEPTADO

Inteligencia Artificial para el diagnóstico primario de dependencia funcional en personas adultas mayores

Santiago Arceo Díaz, Elena Elsa Bricio Barrios, Jaime Alberto Bricio Barrios, Xóchitl Trujillo y Miguel Huerta

La ética de la Inteligencia Artificial

En años recientes, la inteligencia artificial (IA) ha comenzado a integrarse tanto a la vida cotidiana como al quehacer de profesionales de distintas ramas de la ciencia, ingeniería y humanidades. En muchos ámbitos se ha hecho común el uso de asistentes virtuales para facilitar tareas como encontrar la ruta más corta para llegar a un lugar, el servicio al cliente para generar estrategias que faciliten la toma de decisiones y, como se describe en este trabajo, auxiliar en la generación de diagnósticos médicos rápidos y eficientes para enfermedades potencialmente peligrosas para la vida de los pacientes.

La IA es capaz de proveer de estas ventajas gracias a la facilidad de generación de información a través de la transferencia, intercambio y análisis de datos. A pesar de que pudieran parecer sinónimos, la diferencia entre datos e información radica en que los datos son solo números contenidos en una tabla mientras que la información provee de medios para el aprendizaje a partir de estos datos. Este aprendizaje implica la generación de un modelo (una representación simplificada de una situación o problema del mundo real utilizando ecuaciones y fórmulas matemáticas). Si bien los modelos son solo aproximaciones, en muchos casos nos permitirán entender y predecir cómo funcionan todo tipo de fenómenos naturales.

Utilizar IA permite procesar grandes cantidades de datos en poco tiempo y generar modelos predictivos. Un ejemplo destacado es su aplicación durante la pandemia del virus SARS-CoV-2, donde los modelos de IA fueron utilizados para predecir la efectividad de medidas como el aislamiento voluntario y la vacunación. En el ámbito médico, también se ha logrado un éxito significativo con la creación de un banco de imágenes de radiografías de tórax para el diagnóstico temprano de neumonía relacionada con el COVID-19 en pacientes españoles [1].

Es importante tener en cuenta que si bien la IA nos brinda diagnósticos más rápidos y precisos, no debe ser vista como un reemplazo del personal médico, sino como una herramienta complementaria en la formación académica y laboral. La IA puede ser un valioso apoyo para los profesionales de la salud, proporcionándoles información adicional y ayudándoles a tomar decisiones fundamentadas.

Dado que la integración de IA en nuestra sociedad sigue en aumento, es esencial considerar el aspecto ético al desarrollar y aplicar modelos basados en esta tecnología. Por este motivo, la Organización de las Naciones Unidas para la Educación, la Ciencia y la Cultura (UNESCO) ha publicado la Recomendación sobre la Ética de la Inteligencia Artificial que busca garantizar que el uso de datos en la creación de modelos de IA sea justo, equitativo e imparcial, evitando sesgos y segregaciones y enfatiza la importancia de diseñar e implementar sistemas de IA basados en reglas, lógica y principios matemáticos, con el propósito de beneficiar a las personas en situaciones de vulnerabilidad [2].

La IA y el análisis de datos permiten identificar patrones para mejorar la eficiencia y precisión de diagnóstico médicos.

Diagnósticos primarios para aumentar la esperanza de vida

En México, las personas de 60 años o más son consideradas personas adultas mayores. Este grupo de la población es vulnerable y ha recibido un apoyo activo por parte del gobierno federal a través de diversas instituciones y organismos. La Secretaría de Bienestar, el Instituto Nacional de Geriatría, la Distribuidora e Impulsora Comercial CONASUPO y el Instituto Nacional de las Personas Adultas Mayores son solo algunos ejemplos de las instituciones gubernamentales que se dedican a brindar apoyo a los adultos mayores en el país. Además, también existen instituciones privadas, sociedades y asociaciones civiles que se enfocan en atender las necesidades de este grupo de la población .

Estas instituciones, además de brindar servicios médicos para la detección de enfermedades comunes en esta etapa de la vida, como la fragilidad y la sarcopenia que han

© 2023 - Sociedad Mexicana de Inteligencia Artificial ISSN 2007-0691

sido relacionados con un mayor riesgo de fallecimientos prematuros, se ha enfocado en la promover el envejecimiento activo, la participación social y el bienestar integral de los adultos mayores, así como el grado de dependencia funcional [3].

El grado de dependencia funcional en las personas adultas mayores es un factor importante para evaluar su capacidad de ser autosuficientes en las actividades de la vida diaria. En este sentido, el Instituto Nacional de Geriatría ha empleado la escala de Barthel, que es una escala que permite evaluar el grado de dependencia funcional en las personas adultas mayores. Esta escala consta de 10 preguntas relacionadas con las actividades cotidianas como vestirse, utilizar el baño, ducharse, alimentarse, desplazarse, subir y bajar escaleras, trasladarse de la silla a la cama, mantener el control de la función urinaria y fecal, así como bañarse [4].

Cada una de las actividades de la escala de Barthel se evalúa asignando un puntaje específico en función de la capacidad del individuo para realizarla. Estos puntajes reflejan la dificultad o incapacidad de llevar a cabo la actividad y se utilizan para determinar el grado de dependencia funcional de la persona.

Una vez categorizados los participantes según su grado de dependencia funcional, se determinan los cuidados recomendados para minimizar el riesgo de enfermedades crónicas degenerativas. Los adultos mayores independientes pueden realizar sus actividades diarias sin apoyo externo. Aquellos con dependencia leve o moderada requieren apoyo ocasional o frecuente, respectivamente. En casos de dependencia severa, se necesita asistencia permanente o total a cargo de personal geriátrico capacitado. Es importante adaptar los cuidados a las necesidades individuales y mantener una comunicación constante entre el equipo de salud, los cuidadores y el adulto mayor [5].

Incertidumbre sobre los datos obtenidos de la escala de Barthel

En el año 2021, un equipo de profesores del área de ciencias de la salud en el estado de Colima llevó a cabo un interesante proyecto de capacitación en colaboración con un grupo de estudiantes de la Facultad de Nutrición de la Universidad de Colima. El objetivo de esta iniciativa fue identificar las necesidades de las personas adultas mayores que acuden al Centro de Asistencia Social de Día para Personas Adultas Mayores Zona Oriente del DIF Estatal Colima ubicado en la zona metropolitana Colima-Villa de Álvarez.

Previo al proceso de entrevistas con las personas que acuden a este centro recreativo, se realizó un cribado de participantes donde no se consideraron aquellas personas con una edad inferior a 60 años, personas que requieren silla de ruedas para desplazarse o prótesis en pierna o brazos. De la población interesada en participar 155 de ellos cumplieron los requisitos mencionados y firmaron un acuerdo de consentimiento informado que indicaba el objetivo del proyecto, que actividades se comprometían a realizar, así como los riesgos de su participación.

Para garantizar la ética del proyecto, se siguieron los principios éticos establecidos en la Declaración de Helsinki y la Ley General de Investigación en Salud de México. Los investigadores se comprometieron a mantener la confidencialidad y privacidad de los datos, así como a garantizar la equidad, imparcialidad y transparencia en la selección de los participantes para evitar cualquier sesgo en el análisis y presentación de los resultados [6].

Durante el proceso de categorización y canalización de los pacientes con diagnóstico de dependencia funcional, se identificaron algunos patrones importantes como:

- La aplicación de la escala de Barthel para cada participante requería un tiempo variable, que iba desde 20 minutos hasta 2 horas. Este tiempo dependía de la capacidad física y cognitiva de cada persona adulta mayor.
- Físicamente, se notaba que los participantes se fatigaban a medida que se les pedía realizar las actividades o responder las preguntas de la escala de Barthel.
- Al preguntarles sobre su habilidad cognitiva para vestirse o realizar tareas de higiene personal, afirmaban ser capaces, aunque en realidad el evaluador observaba lo contrario. Por timidez o pudor, los pacientes negaban o minimizaban las dificultades relacionadas con el control de sus funciones corporales, como la incontinencia urinaria.
- Durante el procesamiento del diagnóstico primario, se notó que la categorización del grado de dependencia del paciente se obtenía sumando la puntuación de cada pregunta de la escala de Barthel. Esto significa que dos pacientes podrían tener el mismo diagnóstico, pero tener habilidades o capacidades diferentes para realizar actividades específicas.

Tomando en consideración que no es posible modificar el contenido de los cuestionamientos de la escala de Barthel, este estudio propone el uso de la técnica matemática conocida como conjuntos rugosos en este trabajo permitirá reducir o minimizar las preguntas de la escala de Barthel sin comprometer la fiabilidad del diagnóstico.

Mediante el enfoque de conjuntos rugosos, se puede identificar un patrón o estructura en los puntajes de las respuestas de los participantes. Esto se logra mediante la construcción de un árbol de decisión, que actúa como un sistema de IA.

Este enfoque permite optimizar el proceso de evaluación, reduciendo el tiempo, las limitantes físicas de los participantes y sesgo en la interpretación de los resultados de la

estala de Barthel y así obtener un diagnóstico confiable. A continuación se describen las reglas, lógica y principios de estas herramientas, así como el uso de la base de datos de personas adultas mayores que participaron el el estudio para el diagnóstico primario de dependencia funcional.

La teoría de conjuntos clasifica y organiza datos, los arboles de decisión crean reglas de clasificación a partir de estos conjuntos.

Teoría de conjuntos rugosos para reducir cuestionamientos

La teoría de conjuntos es una rama de las matemáticas que se usa para clasificar objetos de acuerdo con una lista de características bien definidas. Es posible concebir un conjunto como una especie de contenedor (una caja) en el que se colocarán aquellos objetos de una lista, de acuerdo con las características que poseen en común (por ejemplo, en el caso de figuras geométricas, si tienen la misma forma, tamaño o color). La teoría de conjuntos no está limitada a clasificar objetos, ya que puede aplicarse también al diagnóstico médico. Bajo la teoría de conjuntos, se propone un modelo objetivo en que el diagnóstico de una enfermedad, puede definirse puramente en función de los síntomas. Estos modelos excluyen todas las características que puedan generar sesgos (tales como prejuicios raciales, de género o socioeconómicos) y decisiones equivocadas (ya que el diagnóstico podría verse influenciado por acciones precipitadas que se basan más en la subjetividad y experiencia previa que en la evidencia que dan las pruebas médicas).

Sin embargo, para aplicar la teoría de conjuntos al diagnóstico médico, debe superarse el reto que implica la incertidumbre. A diferencia de la teoría de conjuntos clásica, que ofrece respuestas claras de "sí" o "no" sobre la pertenencia de un objeto a un conjunto, en el diagnóstico médico a menudo nos encontramos con situaciones inciertas. Estas situaciones pueden surgir cuando un paciente presenta síntomas que sugieren varias enfermedades diferentes o cuando no ha sido posible realizar todas las pruebas necesarias para determinar el diagnóstico de una enfermedad.

La teoría de conjuntos rugosos es una modificación a la teoría clásica que propone una metodología para tratar con la incertidumbre en la clasificación. Esta teoría clasifica los objetos en gránulos de conocimiento (ver Figura 1). Cada gránulo agrupa objetos que tengan valores idénticos para una lista definida de características. A diferencia de la teoría clásica (en la que cada objeto pertenece completamente a un conjunto o no pertenece en absoluto), los gránulos permiten una clasificación más flexible y gradual. Los objetos pueden ser asignados a diferentes gránulos según su grado de similitud. El proceso de clasificación en gránulos de conocimiento implica identificar características relevantes o atributos de los objetos que se van a clasificar. Estas características pueden ser cualitativas o cuantitativas y a partir de ellas se define un criterio de similitud que determina qué objetos deben agruparse en un mismo gránulo, según reglas específicas [7].

Una vez que se establecen los gránulos, cada uno representa un conjunto de objetos que comparten características comunes y se consideran más similares entre sí que con los objetos de otros gránulos. Así se define que gránulos conforman tres regiones de interés: la aproximaciones inferior y superior y la frontera. La aproximación inferior incluye aquellos gránulos que sin duda pertenecen al conjunto de interés (por ejemplo, grupos de pacientes que sin duda alguna sufren de una enfermedad), mientras que la aproximación superior incluye también a aquellos gránulos que podrían pertenecer (grupos de pacientes en los que, con cierta posibilidad, pudiera haber enfermos). La frontera es la diferencia entre ambas listas y corresponde a los gránulos con clasificación incierta (que en el caso médico implica a pacientes con un diagnóstico dudoso, véase la Figura 1). En muchos casos, la frontera está conformada por objetos con atributos idénticos y diagnósticos contradictorios. Al examinar los atributos y características específicas de estos objetos ambiguos, podemos realizar una evaluación más exhaustiva y considerar otros factores relevantes que podrían ayudar a determinar su verdadera clasificación [8].

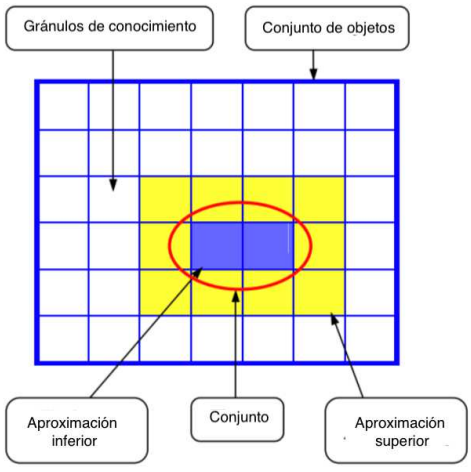


Figura 1. Aproximaciones inferior y superior a un conjunto clásico, los objetos se clasifican dentro de los gránulos de conocimiento [7].

© 2023 - Sociedad Mexicana de Inteligencia Artificial ISSN 2007-0691

Además de permitir la clasificación de objetos, la teoría de conjuntos rugosos también permite reducir la complejidad en el análisis de un conjunto de datos, al identificar y eliminar aquellos atributos que pudieran considerarse "irrelevantes", es decir, que no ayudan realmente a clasificar objetos. Así, los atributos restantes, aportarán información valiosa para realizar el diagnóstico. Esta lista es conocida como "reducto". Para identificar el reducto de un conjunto de objetos, se utiliza el concepto de dependencia de atributos. Un atributo es dependiente de otro si puede ser inferido a partir de él. El reducto se define como el conjunto mínimo de atributos que son necesarios para inferir todos los demás atributos de la tabla de datos, es decir, aquellos que no son dependientes de los otros.

Conjuntos rugosos aplicados a la valoración de la dependencia funcional

El diagnóstico de la dependencia funcional, puede verse beneficiado de la aplicación combinada de la teoría de conjuntos rugosos y la IA. Cada una de las pruebas que comprenden escalas de valoración, como la de Barthel, trae consigo un grado de complejidad adicional al problema de la asignación de diagnósticos. Debido a su naturaleza sumativa, diferentes combinaciones de respuestas pueden llevar a un mismo puntaje final, lo que puede dificultar la distinción entre pacientes con necesidades de atención significativamente diferentes. Aquí es donde entra en juego la teoría de conjuntos y la IA. La teoría de conjuntos proporciona un marco lógico para identificar y clasificar conjuntos de pacientes con características y necesidades similares. Por su lado, la inteligencia artificial puede desempeñar un papel crucial al aplicar algoritmos de aprendizaje automático a los datos recopilados de los pacientes. Estos algoritmos pueden analizar patrones y relaciones complejas entre las respuestas en las escalas de valoración, lo que permite una asignación más precisa de los diagnósticos. La IA puede aprender a reconocer combinaciones de respuestas que implican necesidades de atención específicas y ayudar a diferenciar entre conjuntos de pacientes con necesidades distintas [9].

La teoría de conjuntos rugosos puede aplicarse a la escala de Barthel para la búsqueda de un conjunto de reglas que faciliten el diagnóstico de la dependencia funcional en personas adultas mayores. Partiendo de la muestra de personas adultas mayores que asisten a los centros recreativos, primero identificamos el reducto del sistema (siguiendo el método descrito por Yao y Zhao en [10]). En este caso, de los 10 cuestionamientos que conforman el test de Barthel, el reducto corresponde a las pruebas de: "subir y bajar escales sin ayuda"("B. Escal"), "velocidad de desplazamiento"("B. Desplaz"), "usar el retrete sin ayuda"("B. Retrete"). "trasladarse de la silla a la cama sin ayuda"("B. Silla") y "frecuencia de episodios de incontinencia urinaria"("B. Orina"). Gracias a la identificación del reducto, es posible obtener los perfiles más comunes para los pacientes que pueden clasificarse como "independientes", "dependientes leves", "dependientes moderados.º "dependientes severos". Los perfiles que con mayor frecuencia se observaron en los participantes se muestra en la Figura 2 y 3.

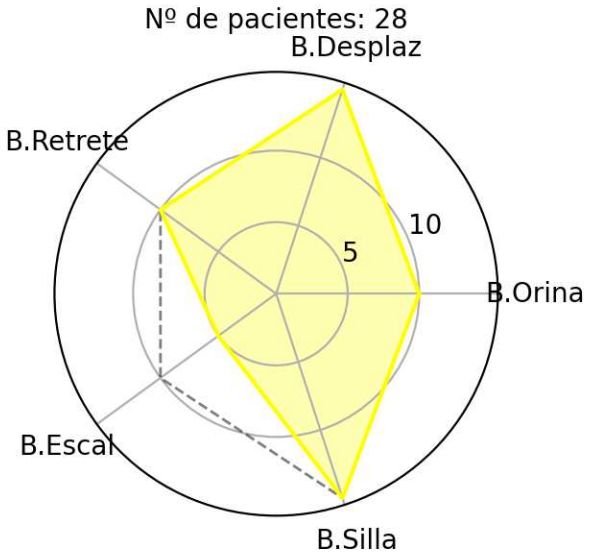


Figura 2. Gráfico de radar que muestra los resultados del grupo más numeroso de pacientes que con signos de dependencia moderada.

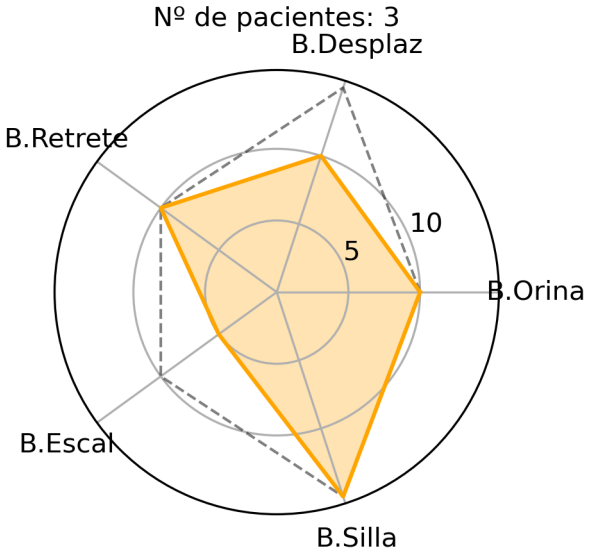


Figura 3. Gráfico de radar que muestra los resultados del grupo más numeroso de pacientes que con signos de dependencia moderada.

© 2023 - Sociedad Mexicana de Inteligencia Artificial ISSN 2007-0691

El reducto contiene características mínimas para diagnóstico eficaz de la dependencia funcional.

Aprendizaje supervisado a través de árboles de decisión

El reducto es una herramienta que nos permite crear un árbol de decisión, el cual es un modelo predictivo construido mediante un diagrama de flujo. Este diagrama muestra cómo las decisiones conducen a diferentes resultados. Los árboles de decisión son muy útiles en el diagnóstico médico, ya que nos ayudan a identificar patrones complejos en grandes conjuntos de datos y a tomar decisiones de manera consistente. Además, los árboles de decisión nos permiten explicar de manera más clara a los pacientes el diagnóstico y las opciones de tratamiento disponibles [10]. También brindan una forma rápida de realizar el diagnóstico y podemos visualizar un ejemplo del modelo predictivo de árbol de decisión para el test de Barthel en la Figura 4.

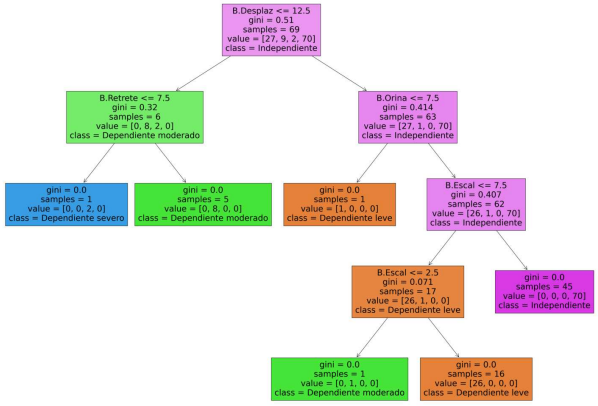


Figura 4. Árbol de decisión que modela la dependencia funcional en términos de los puntajes obtenidos en las pruebas que conforman el reducto de Barthel.

El árbol de decisión utilizado para clasificar a los pacientes según su grado de dependencia. Se basa en condiciones aplicadas a características que conforman el reducto. Cada nodo en el árbol representa una decisión que se toma en función de una variable, y cada rama representa el resultado de esa decisión. El árbol se divide en dos ramas principales, con diferentes condiciones para determinar el grado de dependencia en cada rama. Por ejemplo, si el nivel de escalera es menor o igual a 7.50 y el desplazamiento es menor o igual a 12.50, se aplican diferentes condiciones para determinar si el paciente es un dependiente severo o moderado, basado en su capacidad para usar el retrete. Si el desplazamiento es mayor a 12.50 y el nivel de escalera es menor o igual a 2.50, el paciente es clasificado como dependiente moderado. Este modelo es una herramienta útil para los profesionales de la salud para clasificar rápidamente a los pacientes y proporcionarles el tratamiento y los cuidados adecuados, siendo una alternativa cuando no es posible realizar la batería completa de pruebas.

Conclusiones

En este trabajo, mostramos como puede utilizarse la teoría de conjuntos para el desarrollo de un sistema de diagnóstico de dependencia funcional basado en la prueba de Barthel. La muestra de adultos mayores de 155 pacientes reveló una distribución de dependencia en la población estudiada, con la mayoría de los pacientes siendo independientes. El reducto de pruebas compuesto por las pruebas de “escalera”, “orina”, “desplazamiento”, “silla” y “retrete” se identificó como un conjunto clave para clasificar el grado de dependencia de los pacientes.

A partir de este reducto, se desarrolló un modelo de árbol de decisión que puede ser utilizado para clasificar rápidamente a los pacientes según su grado de dependencia y proporcionar el tratamiento y los cuidados adecuados. Además, se planea implementar este modelo en una IA de aprendizaje supervisado, lo que permitiría una mayor precisión y rapidez en la clasificación de los pacientes. En conjunto, estos hallazgos pueden contribuir a una mejor atención y cuidado de los adultos mayores con diferentes niveles de dependencia. ✳

REFERENCIAS

1. López-Cabrera, J. y Pérez-Díaz, M. (2022). Inteligencia artificial explicable, una perspectiva al problema de la clasificación automática de COVID-19 mediante radiografías de tórax. *MediSur*, 20(2), 341-351.
2. Flores-Vivar, J. y García-Peñalvo, F. (2023). Reflexiones sobre la ética, potencialidades y retos de la Inteligencia Artificial en el marco de la Educación de Calidad. Recuperado el 3 de junio de 2023, de `https://acortar.link/avGcG8`.
3. Prado, L., González-Villegas, G., et al. (2023). Calidad de vida en adultos mayores con enfermedades crónico degenerativas. Ciencia Latina Revista Científica. *Multidisciplinar*, 7(1), 2802-2812.
4. Cid-Ruzafa, J. y Damián-Moreno, J. (1997). Valoración de la discapacidad física: el índice de Barthel. *Revista española de salud pública*, 71, 127-137.
5. Carrillo-Cervantes, A., Medina-Fernández, I., et al. (2022). Sarcopenia como factor predictor de dependencia y funcionalidad en adultos mayores mexicanos. *Index de Enfermería*, 31(3), 170-174.

© 2023 - Sociedad Mexicana de Inteligencia Artificial ISSN 2007-0691

6. Salazar-Arenas, M. Á. y Viveros-Contreras, R. (2022). El uso de información en el área de la salud. *Estancias*, 2(4), 189-207.

7. Pawlak, Z. (1982). Rough sets. *International journal of computer & information sciences*, 11, 341-356.

8. Raj, S. y Vigneshwaran, M. (2020). Analysing diabetes using rough sets. *Advances in Mathematics: Scientific Journal*. 9m(2), 533-538.

9. Vargas, M., Biggs, D., Larraín, T. et al. (2022). Inteligencia artificial en medicina: Métodos de modelamiento (Parte I). *Revista Chilena de Anestesiología*, 51(5), 527-534.

10. Elhoseny, M., Nabil, A. et al. (2018). Hybrid rough neural network model for signature recognition". En Hassanien A., y Oliva D. (Eds.) *.Advances in Soft Computing and Machine Learning in Image Processing"*, Springer Cham, 295-318.

SOBRE LOS AUTORES



Santiago Arceo Diaz Licenciado en Física, Maestro en Física y Doctor en Ciencias (Astrofísica). Posdoctorante de la Universidad de Colima y profesor del Tecnológico Nacional de México campus Colima. Cuenta con el perfil deseable, pertenece al núcleo académico y es colaborador del cuerpo académico Tecnologías Emergentes y Desarrollo Web de la Maestría Sistemas Computacionales. Ha dirigido tesis de la Maestría en Sistemas Computacionales y en la Maestría en Arquitectura Sostenible y Gestión Urbana.



Elena Elsa Bricio Barrios Ingeniera Química Metalúrgica, Maestra en Ciencias en Ingeniería Química y doctorante en Ingeniería Química. Actualmente es profesora del Tecnológico Nacional de México campus Colima. Cuenta con el perfil deseable, es miembro del cuerpo académico Tecnologías Emergentes y Desarrollo Web y ha codirigido tesis de la Maestría en Sistemas Computacionales.



Jaime Alberto Bricio Barrios Licenciado en Nutrición, Maestro en Ciencias Médicas, Maestro en Seguridad Alimentaria y Doctor en Ciencias Médicas. Profesor e Investigador de Tiempo Completo de la Facultad de Medicina en la Universidad de Colima. miembro del Sistema Nacional de Investigadores en el nivel 1. Miembro fundador de la asociación civil DAYIN (Desarrollo de Ayuda con Investigación).



Xóchitl Trujillo Bióloga, Maestra y Doctora en Ciencias Fisiológicas con especialidad en Fisiología. Es Profesora-Investigadora de Tiempo Completo de la Universidad de Colima. Cuenta con perfil deseable y es miembro del Sistema Nacional de Investigadores en el nivel 3. Su línea de investigación es en Biomedicina en la que cuenta con una producción científica de más de noventa artículos en revistas internacionales, varios capítulos de libro y dos libros. Imparte docencia y ha formado a más de treinta estudiantes de licenciatura y de posgrado en programas académicos adscritos al Sistema Nacional de Posgrado del CONAHCYT.



Miguel Huerta Doctor en Ciencias con especialidad en Fisiología y Biofísica. Es Profesor-Investigador Titular "C" del Centro Universitario de Investigaciones Biomédicas de la Universidad de Colima. Es miembro del Sistema Nacional de Investigadores en el nivel 3 emérito. Su campo de investigación es la Biomedicina, con énfasis en la fisiología y biofísica del sistema neuromuscular y la fisiopatología de la diabetes mellitus. Ha publicado más de cien artículos revistas indizadas al Journal of Citation Reports y ha graduado a más de 40 Maestros y Doctores en Ciencias en programas SNP-CONAHCyT.

© 2023 - Sociedad Mexicana de Inteligencia Artificial ISSN 2007-0691

ARTÍCULO ACEPTADO

La necesidad de la ética ante la expansión de la IA

Fabio Morandín-Ahuerma

Introducción

La palabra "ética" tiene su origen en la palabra griega .^{et}hos", que se traduce comúnmente como "costumbre", pero también es "comportamiento" e incluso "carácter" de la persona. La ética, como disciplina filosófica, se ocupa del estudio de la moral y de los principios que guían el quehacer humano. La moral, por su parte, es un conjunto de normas, valores y principios de conducta que determinan qué es considerado como "correcto" e "incorrecto" en una determinada sociedad o contexto. La ética examina las razones y los principios que justifican las normas y los juicios morales. Busca respuestas a preguntas como "¿por qué algo es bueno o malo?" y "¿cuáles son los fundamentos de la moralidad?". De esto surgen las llamadas "éticas aplicadas" como en el campo de la Inteligencia Artificial, en adelante IA, la cual, como se ha visto en este 2023, está cambiando las reglas de juego. La ética de la IA se entiende como un conjunto de directrices (reglas de acción), principios (valores) y regulaciones (estándares) que deben o deberían ser observadas en el desarrollo, distribución, comercialización y uso intensivo de algoritmos de aprendizaje automático, *Machine Learning* (ML) y *Deep Learning* (DL).

Propuestas Normativas

Hay al menos una docena de propuestas que contienen principios éticos y normativos para el desarrollo responsable de IA. Estos documentos podrían llegar a convertirse en reglas e incluyen recomendaciones para la programación de algoritmos de ML, DL, redes neuronales, *big data*, uso y recogida de datos biométricos, computación en la nube, apps, conducción autónoma, armas autónomas y otras áreas relacionadas. Sin embargo, por ahora, para autores como Hagendorff et al. [1], estos documentos, desafortunadamente, fungen solo como "catálogos de buenas intenciones".

Hasta el momento las propuestas más importantes para crear normativas son: los *Principios de inteligencia artificial de Asilomar; la Declaración de Montreal para la IA responsable; la Recomendación de la UNESCO sobre la ética de la IA; la Recomendación del Consejo de la OCDE sobre inteligencia artificial; el Reglamento del Parlamento Europeo y del Consejo por el que se establecen normas armonizadas en materia de inteligencia artificial; y, la Iniciativa global sobre ética de los sistemas autónomos e inteligentes d*e la IEEE, entre otras.

¿Influyen las pautas en ética de IA en la toma de decisiones de los desarrolladores? Hasta ahora, no mucho.

También son dignas de mención las normativas gubernamentales como la Ley de iniciativa nacional de IA de 2020 (Estados Unidos); Código de ética de IA (Rusia); las Normas éticas para una nueva generación de inteligencia artificial (China); el Marco de ética, transparencia y rendición de cuentas para la toma de decisiones automatizada (Reino Unido); y la Directiva sobre la toma de decisiones automatizada (Canadá), entre otras.

Por último, algunos "Códigos de ética" de las propias empresas que desarrollan IA como: Google-Alphabet, Deep Mind, Microsoft, Facebook-Meta, OpenAI, Anthropic, por mencionar las más importantes.

Utilidad de las propuestas

La pregunta incómoda es: ¿Son estas iniciativas influyentes o útiles para los desarrolladores? Las pruebas sugieren que estas normativas a menudo se ignoran debido a su naturaleza abstracta, polisémica o ambigua [1]. En este sentido, McNamara et al. [2] hicieron un trabajo de investigación en la Universidad de Carolina del Norte en donde demostraron que las normativas no tienen mayor efecto, al menos en el caso del Código de Ética de la ACM (Assocation for Computing Machinery). Concluyeron en su trabajo: "Descubrimos que instruir explícitamente a los participantes para que tuvieran en cuenta el código ético de la ACM en su toma de decisiones no tuvo ningún efecto observado, en comparación con el grupo control" (p. 729).

Del mismo modo, Rességuier y Rodrigues [3] afirmaron que la creación de múltiples códigos de ética ha sido una forma de evitar la imposición de normativas jurídicamente vinculantes a las empresas. En su lugar han desarrollado sus propias reglas y formado asociaciones, como Partnership on AI (PAI), para protegerse de posibles obligaciones legales relacionadas con el uso de la IA y los derechos humanos. También argumentan que estas políticas carecen de mecanismos de aplicación, lo que las convierte en un sustituto descafeinado de la regulación. Mientras las normativas no sean vinculantes, esto es, obligatorias legalmente, las empresas seguirán dando

© 2023 - Sociedad Mexicana de Inteligencia Artificial ISSN 2007-0691

prioridad a sus intereses mercantiles sobre las prácticas responsables de la IA [4].

Principios y valores específicos

Un examen de los principios articulados en las normas de ética pone de manifiesto que muchas de ellas expresan valores similares; lo anterior, me parece, lejos de ser una debilidad sugiere que existe un consenso sobre la dirección que debería tomar el desarrollo responsable de la ciencia de datos, IA, ML y DL. ¿Cuáles son los valores mayormente compartidos? De acuerdo con la revisión de la literatura la "protección de la privacidad" aparece en primer lugar [5].

Existen al menos ocho principios éticos que son coincidentes en las iniciativas de formulación de buenas prácticas de IA.

En segundo lugar, la "justicia", entendida como ausencia de prejuicios de algoritmos "mal calibrados", igualdad de oportunidades, equidad de criterios de decisión y representatividad de todos los posibles involucrados. Un algoritmo "mal calibrado" se refiere aquel que produce sesgos y resultados incorrectos, inexactos o inapropiados debido a una configuración o ajuste intencional o no, de sus parámetros o reglas. Por ejemplo, que se decida otorgar una línea de crédito más alta basada en el sexo del solicitante como parámetro definitorio.

En tercer lugar, la "ciberseguridad", esto es, generar sistemas robustos, confiables, que gestionen los riesgos y puedan defenderse contra ataques [6].

En cuarto lugar, el principio del "bien común"suele considerarse uno de los más difíciles de interpretar, ya que implica esforzarse por conseguir un beneficio global para la humanidad, lo cual es un valor abstracto que debe traducirse en acciones concretas. Esto puede significar la promoción del mayor bien para el mayor número de personas.

En quinto lugar, "la transparencia y la apertura" lo que lleva al debate entre código abierto, seguridad, patentes, derechos de autor y derechos industriales. ¿Cuál debería ser el justo medio? Si bien una "caja negra" puede representar un riesgo para la toma de decisiones por sus pasos estocásticos, también es verdad que revelar algunos códigos de programación puede generar un doble propósito, lo que significa que pueden ser utilizados tanto para hacer el bien, como para causar daño.

Un "algoritmo de caja negra" es aquel en el que se conocen sus datos de entrada y los resultados de salida, pero no cómo se llegó a ellos. Estocástico significa azaroso, aleatorio, sin control. Por eso, el sexto principio "supervisión humana" es fundamental para no permitir que decisiones que puedan llegar a ser vitales —que pueda depender de ellas el patrimonio, la salud y hasta la vida de una persona— deba ser siempre tomada por humanos y no dejarlas en manos de la IA, cuyos valores, como ha señado Nick Bostrom [7], podrían no estar alineados con los valores de la humanidad.

El séptimo principio se refiere a la "trazabilidad" que, de alguna manera, también implica "explicabilidad". Si bien la trazabilidad se centra en la procedencia de los datos, modelos y procesos que han conducido a resultados; la explicabilidad trata de articular las razones de la decisión o predicciones concretas. Ambas ayudan a entender los resultados.

El octavo principio es la cadena de responsabilidad. Es importante saber quién ha tenido la intencionalidad, ya sea en la producción, en la comercialización, como en el uso de los algoritmos. Respuestas tales como "el sistema lo hizo" deben terminarse: siempre existe un responsable, por difícil que sea ubicarlo. En el ámbito militar, por ejemplo, la pregunta sería: "¿Quién dio la orden?" [8].

La solidaridad, la inclusión, y la cohesión social son otros aspectos relevantes que aparecen reiteradamente en las normas de ética de la IA, y que deben ser también valorados para evitar sesgos de los algoritmos.



Figura 1. Cadena de valores compartidos para una ética de la IA. Se puede agregar justicia, solidaridad, inclusión y cohesión social, aunque se consideran aquí parte del bien común.

© 2023 - Sociedad Mexicana de Inteligencia Artificial ISSN 2007-0691

Discusión

Los filósofos y expertos debaten activamente sobre las normas y directrices éticas que deberían prevalecer sobre la IA. Desde el escepticismo, estas normas y sus máximas resultan ambiguas y abstractas porque no ofrecen orientaciones concretas para los complejos problemas del mundo. Existen múltiples escenarios y particularidades que no pueden abarcarse todas con principios generales.

Por otra parte, los constructivistas sostienen que las normas éticas desempeñan un papel importante en la sensibilización, el establecimiento de un conjunto de valores compartidos y la observancia de reglas en el sector. Por ejemplo, Luciano Floridi y Mariarosaria Taddeo [9] de la Universidad de Oxford, sostienen que las máximas deontológicas son perfectibles y pueden evolucionar para adecuarse con el tiempo a los casos específicos. La deontología es una rama de la ética que se ocupa del estudio de los deberes, obligaciones y principios morales que guían la conducta profesional o laboral.

Tal vez el término medio sea que estos códigos tienen fuerza como un punto de partida, y deben complementarse con análisis más profundos, razonamientos basados en casos y procesos estructurados para abordar cuestiones prácticas.

Algo parecido a lo que la IEEE ha hecho a través del *7000TH-2021: Proceso del modelo estándar para abordar las preocupaciones éticas durante el diseño del sistema*, en donde señalan la necesidad de incluir los siguientes procesos: a) De elicitación y priorización de valores éticos; b) De definición de requisitos éticos y, c) De diseño ético basado en riesgos.

Además, la ética de la IA evoluciona constantemente a medida que la tecnología se vuelve más sofisticada. Esto significa que los filósofos siempre tendrán que estar preparados para actualizar sus marcos éticos a las nuevas tecnologías, pero también en sentido opuesto, la tecnología debe adecuarse a la ética.

Se necesita más trabajo empírico para determinar qué funciona, para quién y en qué contextos puede incidir realmente la ética para que los intereses económicos de las grandes empresas no estén por encima de los derechos de los consumidores.

Los expertos tienen opiniones encontradas sobre el potencial de las normas éticas para configurar de manera significativa el desarrollo y uso de la IA de forma responsable. El debate sigue abierto. Se considera en lo general que las normas son un punto de partida, no un punto final.

Sin embargo, para el filósofo inglés Tomás Hobbes en su obra El Leviatán (1651), en una sociedad sin leyes, no se tienen en cuenta los derechos de los demás, y las decisiones y acciones se guían únicamente por el egoísmo. Sin la ética, la confianza y la cooperación entre individuos y comunidades se erosiona, pues no hay normas compartidas para defender la equidad, la justicia y la compasión. Un mundo así estará marcado por el caos, la desigualdad y el desprecio de los principios que promueven la armonía y el florecimiento de la humanidad.

Del mismo modo, se puede decir que una IA sin ética, sin regulación, podría conducir a un futuro plagado de consecuencias imprevisibles y peligros potenciales. Sin normas éticas y supervisión, los sistemas de IA podrían preferir la eficiencia y la optimización a expensas del bienestar humano y los derechos individuales, como Bostrom [7] insiste.

Una IA sin control puede perpetuar los prejuicios, reforzar la discriminación y exacerbar las desigualdades sociales. Además, sin regulación, la IA puede utilizarse con fines maliciosos, permitiendo ciberataques, abusos de vigilancia o incluso armas que decidan a quién matar [9].

Conclusión

Se necesita más trabajo empírico para determinar qué funciona, para quién y en qué contextos puede incidir realmente la ética para que los intereses económicos de las grandes empresas no estén por encima de los derechos de las personas.

En ese trabajo, se hizo un repaso somero de los valores de los códigos normativos y su potencial para configurar de manera significativa el desarrollo y su uso responsable de la IA. Se considera que las normativas son un punto de partida, pero no un punto final, y por ello se necesita mayor análisis sobre la resolución de problemas prácticos, especialmente en el caso del autoaprendizaje y toma de decisiones automatizadas.

No es, en definitiva, pidiendo "treguas" [10] que se podrá detener la bola de nieve en las fronteras indistinguibles del doble propósito en que la IA se encuentra ahora.

La vigilancia es fundamental, pero no debe paralizar un trabajo prometedor de la ingeniería computacional que, sin duda, está beneficiando a la humanidad de muchas formas.

La historia parece sugerir que no debe desarrollarse ninguna tecnología o descubrimiento sin considerar cómo limitarlo y acotarlo adecuadamente para garantizar que se mantenga en línea con los valores y el bienestar humanos.

La ética de la IA es un trabajo en ciernes, pero necesario. Se espera que las iniciativas de normativas sean influyentes y útiles para los desarrolladores, como este número de *Komputer Sapiens*. ✷

REFERENCIAS

1. Hagendorff, T. (2020). The Ethics of AI Ethics: An Evaluation of Guidelines. *Minds and Machines*, 30(1), 99-120.

2. McNamara, A., Smith, J. y Murphy-Hill, E. (2018). Does ACM's code of ethics change ethical decision making in software development?. *ESEC/FSE*, 729-733.

© 2023 - Sociedad Mexicana de Inteligencia Artificial ISSN 2007-0691

3. Rességuier, A. y Rodrigues, R. (2020). AI ethics should not remain toothless! A call to bring back the teeth of ethics. *Big Data & Society*, 7(2).

4. Lauer, D. (2021). Facebook's ethical failures are not accidental; they are part of the business model. *AI Ethics*, 1, 395–403.

5. Véliz, C. (2022). Privacy Is Power: Why and How You Should Take Back Control of Your. New York, Penguin, ISBN 978-055-217-771-9.

6. Pearce, W. y Kumar, S. (2021). Best practices for AI security risk management. Recuperado el 30 de junio de 2023, de: `https://bsu.buap.mx/b31`.

7. Bostrom, N. (2014). Superintelligence: Paths, Dangers, Strategies. *Oxford, Oxford University Press*, ISBN 978-019-873-983-8.

8. Galliott, J. y Scholz, J. (2020). The Case for Ethical AI in the Military. En Dubber M., et al. (Eds.) *The Oxford Handbook of Ethics of AI. Oxford University Press*, 684–702.

9. Taddeo, M. y Floridi, L. (2021). How AI Can Be a Force for Good – An Ethical Framework to Harness the Potential of AI While Keeping Humans in Control. En L. Floridi (Ed.), *Ethics, Governance, and Policies in AI, Springer*, 91-96.

10. Future of Life Institute. (2023). Pause Giant AI Experiments: An Open Letter". Recuperado el 2 de julio de 2023, de: `https://bsu.buap.mx/b8d`.

SOBRE EL AUTOR



Fabio Morandín-Ahuerma es doctor en Filosofía y profesor Investigador de tiempo completo de la Benemérita Universidad Autónoma de Puebla. Autor de los libros: «Principios normativos para una ética de la Inteligencia Artificial» (Concytep, 2023) y «Neuroética fundamental y teoría de las decisiones» (Concytep, 2021). Líder del Cuerpo Académico BUAP-CA-354. SNI 1. Página web: `fabiomorandin.net`.



Imagen tomada de: https://pharmaboardroom.com/articles/ethics-in-artificial-intelligence-better-prepare-now/.

IA & Educación

Lucía Barrón , Julieta Noguez y Yasmín Hernández

Implicaciones éticas de la IA en la educación

La educación tiene como objetivo el pleno desarrollo de la personalidad humana incluyendo aspectos cognitivos, artísticos y éticos entre otros, con el fin de propiciar que los individuos se desarrollen bajo una cultura de igualdad de oportunidades[1]. La Real Academia de la Lengua Española define la palabra ética como "parte de la filosofía que trata del bien y del fundamento de sus valores " o bien como el "conjunto de reglas morales que rigen la conducta de la persona en cualquier ámbito de su vida." [2]. Específicamente en el ámbito escolar la ética involucra aspectos tales como respeto, justicia y honestidad, entre otros, relacionados con el comportamiento de las y los estudiantes.

La popularidad de internet desde hace más de dos décadas, la aparición de Wikipedia en 2001, el creciente número de blogs, páginas web institucionales y universidades virtuales pusieron a disposición de los usuarios materiales educativos de diferentes disciplinas, algunos con alta calidad y veracidad en sus contenidos, pero también otros con niveles de calidad inaceptables o de dudosa veracidad. Desde entonces, los estudiantes de los diversos niveles educativos recurren frecuentemente a estos medios digitales en internet para consultar y recolectar información que les permita cumplir con algunas actividades escolares como son: elaborar tareas, aprender técnicas para solución de problemas, escribir reportes de investigación, etc. Esto propicia que en algunos casos estos trabajos escolares carezcan de rigor científico en su contenido y peor aún que se reproduzca y propague información falsa haciendo uso del famoso "copiar-pegar" cuando se elabora un reporte, incurriendo además en faltas éticas que limitan el desarrollo humano.

Más recientemente la aparición de sistemas construidos con grandes modelos lingüísticos (LLMs por sus siglas en inglés) han demostrado el potencial de estos sistemas para automatizar procesos de generar y analizar de forma eficiente gran cantidad de contenido en formato de texto liberando a los educadores de tareas laboriosas como evaluar ensayos o proporcionar retroalimentación [3]. Por otra parte, los sistemas como Bert [4] y ChatGPT [5] que son capaces de producir texto para responder a preguntas fácticas o generar ensayos, argumentos y justificaciones, han motivado a los educadores a señalar los riesgos de que los estudiantes usen estas herramientas porque inhiben el desarrollo de habilidades de pensamiento crítico y resolución de problemas que son indispensables para su formación.

Algunas escuelas, instituciones y empresas alrededor del mundo han decidido temporalmente limitar el uso de estas herramientas en sus entornos, porque existe preocupación de que los usuarios (estudiantes o trabajadores) durante su interacción con la herramienta realicen plagio o provean información sensible.

Los educadores se enfrentan hoy a una nueva forma de un viejo problema: los estudiantes intentan realizar sus tareas escolares sin mucho esfuerzo. Esto crea la necesidad de que las herramientas sean capaces de reconocer los textos generados por ellas mismas. Sin lugar a dudas, se requiere una formación centrada en las necesidades particulares de cada persona y con enfoque humanista.✷

REFERENCIAS

1. Hernández, Silvestre Manuel. (2010). Educación y ética. Sociológica (México), 25(72), 215-227. Recuperado en 20 de mayo de 2023, de http://www.scielo.org.mx/scielo.php?script=sci_arttext&pid=S0187-01732010000100010&lng=es&tlng=es.

2. Diccionario de la lengua española. (s.f.). Ético. en el Diccionario de la lengua española de la Real Academia Española. Recuperado el 25 de mayo, 2023, enhttps://dle.rae.es/%C3%A9tico.

3. Yan, L., Sha, L., Zhao, L., Li, Y., Maldonado, R.M., Chen, G., Li, X., Jin, Y. y Gašević, D. (2023). Practical and Ethical Challenges of Large Language Models in Education: A Systematic Literature Review. ArXiv, abs/2303.13379. arXiv:2303.13379v1

4. Devlin, J., Chang, M., Lee, K., Toutanova, K. (2018). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. arXiv:1810.04805v2 [cs.CL]

5. ChatGPT: Optimizing Language Models for Dialogue. http://openai.com/blog/chatgpt(consultada el 24 de enero de 2023).

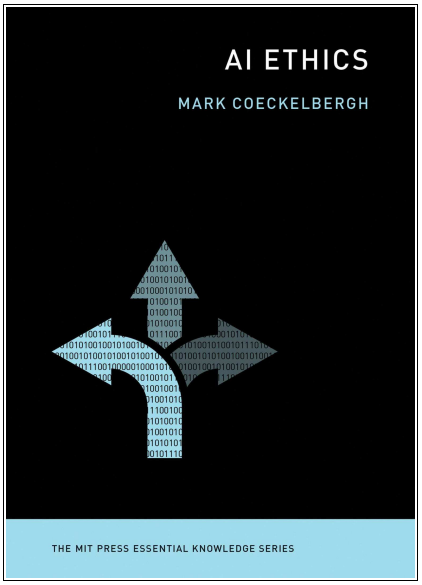
© 2023 - Sociedad Mexicana de Inteligencia Artificial ISSN 2007-0691

Deskubriendo Konocimiento

Alejandro Guerra Hernández y **Gildardo Sánchez Ante**
deskubriendokonocimiento@komputersapiens.org

Revisión de "AI Ethics" *

Dr. Alejandro Guerra Hernández
Universidad Veracruzana
Instituto de Investigaciones en Inteligencia Artificial



Portada del libro. *

Supongamos que recibe usted la invitación a firmar una carta en apoyo a detener temporalmente el desarrollo de GPT4, el conocido modelo de lenguaje. Firmar, o no firmar, es la cuestión. Pues bien, tal decisión va más allá de consideraciones estrictamente científicas y necesariamente involucra aspectos sociales, culturales, filosóficos y éticos. Pero la verdadera cuestión no es firmar o no firmar el manifiesto, sino en que medida estos aspectos son considerados actualmente en el desarrollo de la Inteligencia Artificial (IA). Y ese es el tema central del libro que hoy nos ocupa.

Su autor, el Profesor Mark Coeckelbergh (Bélgica, 1975), es un estudioso de la ética de las nuevas tecnologías, robótica e inteligencia artificial incluidas; y un consultor de políticas en estas áreas. Su justificación del texto es la siguiente: Los espectaculares avances de la IA han dado lugar a mucha especulación acerca de nuestro futuro, afortunadamente, acompañada de una interesante discusión filosófica en torno a lo que significa ser humano. Al mismo tiempo, la IA ya está presente en un amplio rango de dominios de conocimiento y prácticas humanas, lo cual genera cierta urgencia entre los estudiosos de la ética y los elaboradores de políticas por asegurar que estas tecnologías sean benéficas, evitando que se generen problemas insuperables para determinados individuos y sociedades. Estas preocupaciones son más prácticas y parecen más inmediatas que las anteriores. La cuestión es entonces, cómo jerarquizamos todas estas preocupaciones y en qué medida las atendemos en el desarrollo de la IA.

Para dar cuenta de ello, el libro se organiza en doce capítulos. El primero de ellos introduce la justificación que comento arriba. El segundo documenta la super inteligencia y otros monstruos propios de la IA apocalíptica. La propuesta aquí es identificar por qué ciertas narrativas son tan frecuentes, quién las ha creado y a quién benefician; con el objetivo de construir nuevas narrativas sobre el futuro de la IA donde usemos filosofía y ciencia para examinar críticamente el papel que juegan las IAs y los humanos en estos escenarios; mirar con más detalle lo que las IAs pueden hacer actualmente; discutir los problemas éticos y sociales más prácticos y urgentes generados por la IA en estos momentos; investigar las políticas para la IA en un futuro cercano; y considerar si otros problemas son más urgentes que la IA, como el cambio climático.

El tercer capítulo plantea el problema de si la IA general es posible. ¿Pueden las computadoras tener inteligencia, conciencia y creatividad? Aquí se revisan las críticas conocidas a la IA y otras más recientes relacionadas con los escenarios post-humanistas y trans-humanistas. El cuarto capítulo aborda, en este contexto, una cuestión central al libro ¿Qué clase de capacidades morales tienen, o debería tener, nuestros agentes inteligentes artificiales? Y en consecuencia ¿Cómo debemos tratar a tales agentes? Estas cuestiones pueden considerarse como menos urgentes, pero no por ello carecen de importancia. En particular, nos enseñan cosas nos solo sobre la IA, sino sobre cómo los humanos pensamos, actuamos y nos relacionamos con los no-humanos.

Los capítulos quinto y sexto pueden resultar menos interesantes

*M. Coeckelbergh. AI Ethics. MIT Press, Cambridge, MA, USA, 2021

© 2023 - Sociedad Mexicana de Inteligencia Artificial ISSN 2007-0691

para un especialista en IA, pero buscan que el libro sea auto contenido. Aquí se introduce la IA, el aprendizaje automático y la ciencia de datos. Ahora, el cierre de esta parte es de antología: La estadística no solía considerarse una disciplina sexy, hoy, como parte de la ciencia de datos y en su forma de IA trabajando con grandes volúmenes de datos, ¡no solo es sexy, es candente!

El capítulo séptimo es el primero que trata con las cuestiones éticas prácticas y urgentes, incluyendo la privacidad; la manipulación y explotación de usuarios vulnerables; las noticas falsas, los riesgos totalitarios y el impacto en las relaciones personales; y la seguridad. El capítulo octavo se concentra en un problema que resulta de reforzar la agencia de nuestra IAs o de substituir la agencia humana con estas tecnologías: la atribución de responsabilidad y las condiciones asociadas de transparencia y explicabilidad. El capítulo noveno discute otros problemas que además de éticos son sociales. Uno de ellos nos es muy conocido, los problemas y efectos de los sesgos en la IA basada en aprendizaje automático. El otro no, y tiene que ver con el sentido la vida. ¿Qué aspectos de lo que nos define como humanos serán afectados por la IA? Por ejemplo, el trabajo es una fuente de ingresos, pero también define de alguna forma lo que son las personas. Destruir empleo, afecta ambas dimensiones del trabajo.

Los capítulo décimo y undécimo abordan las consideraciones y retos en el diseño de políticas sobre la IA. En todas ellas debe ser claro no solo lo que se tiene que hacer, sino también quién tiene que hacerlo, cuando y en qué medida. El texto ofrece muchos ejemplos de estos requerimientos. Los retos incluyen anticiparse a la tecnología en cuestión, posiblemente embebiendo la ética en su diseño con la participación temprana de los interesados: desarrolladores, usuarios, empresas, etc. En esto es fundamental cambiar la narrativa de la ética como una fuente de restricciones, promoviendo una visión donde es el soporte de la innovación responsable y el desarrollo sustentable a largo plazo, tanto de empresas como de la sociedad.

El capítulo doce ofrece las opiniones finales del autor, discutiendo si la ética de la IA debe ser homocéntrica; cómo establecer nuestras prioridades al respecto correctamente; la relación de la IA con el cambio climático y algunas consideraciones finales sobre la inteligencia, la sabiduría y lo deseable de su incorporación en la IA.

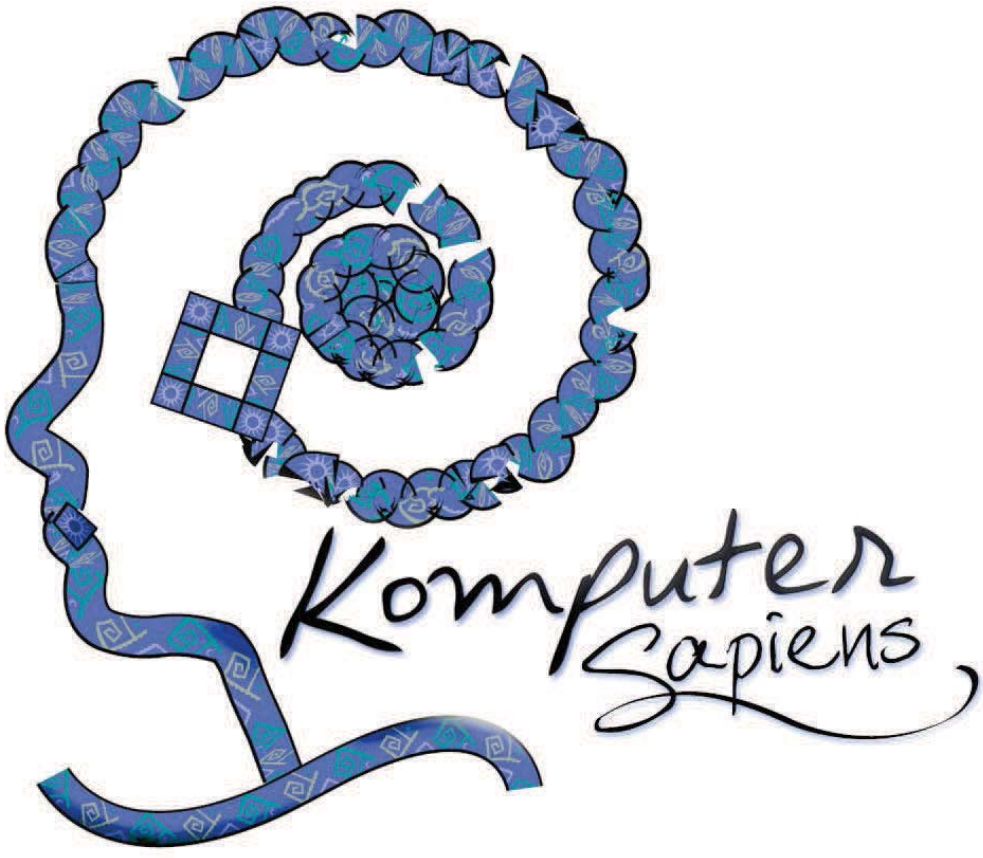
Este texto está publicado en la serie de Essential Knowledge de MIT Press, por lo que uno puede esperar una presentación accesible, concisa y elegante del tema en cuestión. Se ofrece además un glosario muy útil y una lista de lecturas complementarias, más allá de las numerosas referencias. Me parece que esta lectura es un buen punto de partida para considerar las implicaciones éticas, y filosóficas, de la IA y reconsiderar el papel de estas disciplinas en nuestra práctica profesional.✷



Imagen tomada de:
https://swisscognitive.ch/2023/04/19/ai-ethics-in-focus-balancing-innovation-security-and-responsibility-swisscognitive-ai-radar/.

© 2023 - Sociedad Mexicana de Inteligencia Artificial ISSN 2007-0691

Indizada en el IRMDCT de CONACYT y en Latindex



¡Publique en Komputer Sapiens!

CONACYT
ÍNDICE DE REVISTAS MEXICANAS DE DIVULGACIÓN CIENTÍFICA Y TECNOLÓGICA (IRMDCT)

Komputer Sapiens solicita artículos de divulgación en todos los temas de Inteligencia Artificial, dirigidos a un amplio público conformado por estudiantes, académicos, empresarios, tomadores de decisiones y consultores. Komputer Sapiens es patrocinada por la **SMIA, la Sociedad Mexicana de Inteligencia Artificial**

SMIA
www.smia.org.mx

Instrucciones para autores e información general: http://www.komputersapiens.org
Síguenos en las redes sociales: www.facebook.com/Komputer.Sapiens, twitter.com/KomputerSapiens