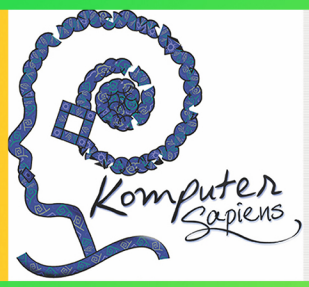
www.komputersapiens.org



KOMPUTER SAPIENS

Revista de Divulgación de la Sociedad Mexicana de Inteligencia Artificial

Año 17
Volumen 3
Septiembre-Diciembre 2025



ISSN 2007-0691

©Komputer Sapiens, Año XVII Volumen III, septiembre-diciembre, 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 del 2025. 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 SECIHTI y en Latindex.

Directorio SMIA

Presidenta	María Lourdes Martínez Villaseñor
Vicepresidenta	Hiram Eredín Ponce Espinosa
Secretario	Roberto Antonio Vázquez Espinosa de los Monteros
Tesorero	Gilberto Ochoa Ruiz
Vocales:	Gustavo Arroyo Figueroa
	Noé Alejandro Castro Sánchez
	Hugo Jair Escalante Balderas
	Sofia Natalia Galicia Haro
	Antonio Marín Hernández
	Bella Citlali Martínez Seis
	Iris Iddaly Méndez Gurrola
	Leobardo Morales Tiburcio
	César Jesús Núñez Prado
	Iván Alejandro Olvera Rodríguez
	Obdulia Pichardo Lagunas
	David Eduardo Pinto Avendaño
	Victor Josué Ruiz Martínez

Komputer Sapiens

Directora general	María Lourdes Martínez Villaseñor
Editora en jefe	Karina Mariela Figueroa Mora
Editor científico	Héctor Gabriel Acosta Mesa
Editores asociados	Marco Antonio Aguirre Lam
	Laura Cruz Reyes
Editores invitados	J. Antonio García Macías
	Cuauhtémoc Rivera-Loaiza
Coordinadora de producción	Viridiana Mena Gómez
e-Tlakuilo	Karina Figueroa
	Rosa María Valdovinos
Kalixxotl	José Manuel Morelos Villegas
	Fernando Piña Campos
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
Corrección de estilo	Claudia Leticia Díaz González
	Gilberto Rivera Zárate
	Guillermo de Jesús Hoyos Rivera
	Luis Miguel García Velázquez
	Marcela Quiroz Castellanos
	Nelson Rangel Valdez
	Nicandro Cruz Ramírez
	Sonia Lilia Mestizo Gutiérrez
	Susana Aurora Velasco Montiel
Portada	Fernando Piña Campos
	José Manuel Morelos Villegas

Directores Fundadores

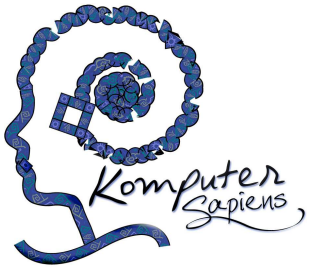
Carlos Alberto Reyes García
Ángel Kuri Morales

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

Adriana Sainz Navarro
Adriana Laura López Lobato
Adriana Lorena Iñiguez Carrillo
Aldo Márquez-Grajales
Ana Isabel Martínez García
Angel Leonel Ortiz Herrera
Armando Ayala Corona
Cuauhtémoc Rivera-Loaiza
Diego Alfredo Pérez Rivas
Edgard Ivan Benítez-Guerrero
Efrén Mezura Montes
Evelio Martinez
Everardo Gutiérrez López
Federico Miguel Cirett Galán
Guadalupe Carmona Arroyo
Gustavo Adolfo Vargas Hakim
Isaac de Jesus Palazuelos Rojo
Isabel I López Hurtado
Jessica Beltrán Márquez
Jesus Horacio Pacheco Ramirez
Jesús Nestaly Marín Nevárez
José Luis Morales Reyes
Karina Gutierrez Fragoso
Karina Mariela Figueroa Mora
Laura Sanely Gaytán Lugo
Leocundo Aguilar Noriega
Lizbeth Escobedo
Luis Adrián Castro
María Del Carmen Mezura-odoy
María Margarita Reyes Sierra
Mario Castelán
Marisol Flores Garrido
Martha Lorena Avendaño-Garrido
Nancy Pérez Castro
Nora Susana Reyes
Pedro Gonzalez Zamora
Pedro César Santana Mancilla
Raquel Torres
René Navarro
Rocio Erandi Barrientos Martínez
Rolando Menchaca Méndez
Sandra Edith Nava Muñoz
Valeria Soto-Mendoza
Victor Hugo Benitez Baltazar



Contenido

Robots y su aplicación en la rehabilitación de Hombro Congelado
Pág. 8 ⇒Jimena Sánchez Guerrero, Edwin Sebna Albarrán García, Adriana H. Vilchis González y Juan C. Ávila Vilchis

Innovación en la rehabilitación de hombro mediante juegos de realidad virtual
Pág. 15 ⇒ Citlali Irais Cano Del Razo, Juan Manuel Jacinto Villegas, Monserrat Ríos Hernández y Adriana Herlinda Vilchis González

¿Eres tú, Juan?: La amenaza de los *deepfakes* y la investigación que busca mitigarlos
Pág. 21 ⇒ Edgar Macías García, David Gómez Gutiérrez, Paulo López Meyer, Rodrigo Aldana-López y Julio Zamora-Esquivel

El Efecto Mandela Digital: Cómo los Ataques Adversarios Engañan a las Redes Neuronales
Pág. 27 ⇒ Monserrat Vázquez Hernández, Luis Alberto Morales Rosales e Ignacio Algredo Badillo

Detectando las sombras del engaño: Cómo la IA responsable puede desarmar los principios de persuasión en ataques de phishing
Pág. 31 ⇒ Viviana I. Fuentes-Fuentes, Luisa F. Agudelo-Fuentes, Vitali Herrera-Semenets, Miguel Á. Álvarez- Carmona y Lázaro Bustio-Martín

Generación Adversaria: Cómo Contribuye a una IA Responsable en Medio de Desafíos Éticos y Riesgos
Pág. 37 ⇒ Samantha Acosta Ruiz, Mireya Tovar Vidal y José Alejandro Reyes Ortiz

Programando la creatividad: El caso de AARON, el robot pintor
Pág. 42 ⇒ Diana Fernández Duque, Itzel Aranguren-Navarro, Guillermo Barrios del Valle y Andrea Fernández Duque

Inteligencia artificial y los sesgos machistas
Pág. 47 ⇒ Mónica Jaimes Gaona

De la ficción a la realidad: La Inteligencia Artificial a través de los ojos de Isaac Asimov
Pág. 51 ⇒ Oscar Ramos-Soto, J. Antonio Ruvalcaba, Diego Oliva, y Sandra E. Balderas-Mata

¿Autonomía o Control? Caminos hacia una Inteligencia Artificial Responsable
Pág. 55 ⇒ J. Antonio García Macías y Antonio de Jesús García Chávez

Emociones Dirigidas: Manipulación Inteligente a través de la Segmentación
Pág. 59 ⇒ Angel Gaspar, Luis Pérez Gaspar, Diego Oliva e Itzel Aranguren-Navarro

Sesgo en los Datos y los Prejuicios en Algoritmos de Aprendizaje Automático
Pág. 64 ⇒ Daniel Sánchez Ruiz, Ricardo Ramos Aguilar, Eric Ramos Aguilar y Jesús García Ramírez

Columnas

Sapiens Piensa. Editorial pág. 2

Kalixxotl pág. 4

e-Tlakuilo pág. 5

Estado del IArte pág. 6

Sakbe pág. 7

IA & Educación pág. 69

Deskubriendo Konocimiento pág. 71

Sapiens Piensa

J. Antonio García Macías y Cuauhtémoc Rivera-Loaiza

Conforme fue avanzando el presente siglo y se retomaron con éxito técnicas como el aprendizaje profundo, la Inteligencia Artificial (IA) tomó nuevos bríos y hoy en día se ha convertido en tópico de discusión, no sólo en entornos académicos, sino en medios no especializados e incluso en charlas de sobremesa. Como suele suceder con los tópicos de moda que tienen un trasfondo complejo, las discusiones se polarizan: en el caso de la IA, los polos opuestos lo constituyen, por una parte, quienes adoptan una posición tecnofóbica y argumentan que la IA pone a la humanidad en un riesgo existencial y que constituye nuestro último gran invento antes de la desaparición de la raza humana; en el otro extremo, están quienes toman una postura más bien tecnofílica ven en la IA la mejor herramienta para ayudarnos a resolver los grandes problemas que aquejan a la humanidad, como son el cambio climático, las pandemias, la desigualdad económica, los conflictos bélicos, la falta de acceso a recursos básicos como agua potable y alimentos. Consideramos que estas posturas, en esencia binarias, ocultan todos los matices que existen en la realidad. La IA podrá ser un factor de gran peso para resolver esos grandes problemas, a la vez que podemos evitar resultados catastróficos de su uso, siempre y cuando tomemos las medidas necesarias en ambos casos.

En medio de la efervescencia en el desarrollo de nuevas técnicas (y el reciclamiento y actualización de otras), el enfoque se ha puesto en los aspectos técnicos tendientes a mejorar el desempeño, minimizar el consumo de energía, aumentar la precisión, reducir el uso de recursos de procesamiento y concebir algoritmos más eficientes y elegantes. Sin embargo, es necesario tomar las medidas adecuadas para guiar el desarrollo de la IA de manera que esté alineada con los valores y objetivos humanos, todo esto mientras se ha ido tornando más compleja y autónoma; es decir, se trata no solamente de atender los problemas técnicos, sino también aquellos referentes al impacto actual y futuro sobre la humanidad y su entorno de vida. Estamos, pues, ante un problema complejo que debe abordarse como un ecosistema de carácter socio-técnico: esta es en esencia la naturaleza de la IA Responsable.

¿De qué tipo de responsabilidades hablamos cuando abordamos la IA Responsable? Una de gran importancia es la de guiar su desarrollo en compatibilidad con los valores y objetivos humanos, el ya señalado problema del alineamiento. Aunque los principios constituyentes de la IA Responsable varían según los autores, en general se suele coincidir en los siguientes: equidad, confiabilidad, privacidad y seguridad, inclusión, transparencia y rendición de cuentas. Ya de inicio, traducir estos conceptos a procesos y controles del mundo real resulta ser un desafío mayúsculo. También resulta primordial ubicar sobre quiénes recaen las responsabilidades; aquí entran en juego desarrolladores, usuarios, organizaciones, formuladores de políticas e incluso el público en general, todos los cuales deberían tener interés en garantizar que los sistemas de IA se desarrollen y utilicen de forma ética y responsable. Estamos ante un problema *multistakeholder*, donde son varias las partes interesadas y responsables.

En este sentido vimos la necesidad de proveer de un espacio que permitiera exponer la manera en que el uso de la IA, en particular de la IA Responsable, está siendo tratada en México y latinoamérica. Komputer Sapiens nos encomendó dirigir la presente edición de la revista sobre este tema, teniendo una excelente respuesta a la convocatoria que lanzamos. En total se recibieron 36 trabajos, de los cuales se aceptaron 10 trabajos; también incluímos dos artículos invitados. En su conjunto, los 12 artículos de esta edición muestran una perspectiva general de cómo los principios que componen al Cómputo Responsable están siendo abordados en nuestra región.

Los artículos que se publican en esta edición cubren una amplia gama de temas, donde el uso responsable de la IA es el tema toral. Una de las principales preocupaciones sobre el uso generalizado de los sistemas de IA actuales son los temas de seguridad y privacidad; esto se ve reflejado en la temática abordada por la mayoría de los artículos aceptados. El primero de ellos trata sobre los deep fakes (“¿Eres tú, Juan?: La amenaza de los "Deepfakesz la investigación que busca mitigarlos”) y otro más sobre el problema del phishing (“Detectando las sombras del engaño: Cómo la IA responsable puede desarmar los principios de persuasión en ataques de phishing”). Un tercer artículo presenta el caso de la potencial manipulación de los sistemas de IA para generar información falsa (“Generación Adversaria: Cómo contribuye a una IA Responsable en medio de desafíos éticos y riesgo”) y, por último, se tiene un artículo que expone las consecuencias de la manipulación de redes neuronales con fines potencialmente dañinos (“El Efecto Mandela digital: Cómo los ataques adversarios engañan a las redes neuronales”).

Dentro del cómputo médico se tienen dos artículos (ambos por invitación) que exploran el uso de la IA en labores de rehabilitación física: “Robots y su aplicación en la rehabilitación de hombro congelado”, e “Innovación en la rehabilitación de hombro mediante juegos de realidad virtual.”.

© 2025 - Sociedad Mexicana de Inteligencia Artificial ISSN 2007-0691

Dos de los artículos aceptados exploran el impacto de la IA en la creación artística, ya sea desde la creación de obras pictóricas ("Programando la creatividad: El caso de AARON, el robot pintor") o con la intersección de la IA con ficción y la realidad ("De la ficción a la realidad: La Inteligencia Artificial a través de los ojos de Isaac Asimov").

Otros dos artículos presentan argumentos sobre los efectos negativos del sesgo en IA: "Inteligencia artificial y los sesgos de género", y "Sesgo en los datos y los prejuicios en algoritmos de Aprendizaje Automático".

Y finalmente, dos artículos se abocan a discutir los aspectos éticos del uso de la IA sobre la sociedad en su conjunto: "Autonomía vs. control para IA Responsable", y "Emociones dirigidas: Manipulación inteligente a través de la segmentación".

Como editores de esta edición de la revista agradecemos enormemente la participación de los autores de los artículos que se enviaron a consideración, así como un gran reconocimiento a todos los revisores de los trabajos. Esperamos que este número contribuya a la divulgación de lo que consideramos es un tema fundamental: el hacer uso de manera responsable de la Inteligencia Artificial. ✳

SOBRE LOS AUTORES



J. Antonio García Macías Obtuvo su doctorado en el INP-Grenoble, Francia. Desde el 2002 es investigador en el Departamento de Ciencias de la Computación del CICESE, Baja California. Es miembro del SNII nivel I y miembro titular de la Amexcomp. Actualmente coordina la red de investigación en Justicia, Ética, Democracia e Inclusión en las Sociedades Digitales (JEDI-SD). Su investigación se ha enfocado en la aplicación del cómputo ubicuo a problemas de comunidades marginadas.



Cuauhtémoc Rivera-Loaiza Es profesor jubilado de la Facultad de Ciencias Físico-Matemáticas de la Universidad Michoacana de San Nicolás de Hidalgo, en Morelia, Michoacán. Obtuvo su Doctorado en Educación por la UNICLA, y es miembro adherente de la Amexcomp. Su área de investigación es la Interacción Humano-Computadora, habiendo sido co-fundador de la AMEXIHC. Ha sido parte del comité organizador de CHI (ACM/SIGCHI) en dos ocasiones (2005 y 2022), además de los congresos MexIHC, CLIHC y ENC, entre otros.



© 2025 - Sociedad Mexicana de Inteligencia Artificial ISSN 2007-0691

Kalixxotl: el arte en la portada

José Manuel Morelos Villegas

La conciencia: una referencia conocida

La imagen de la portada vincula un dispositivo simulado y la figura de un ícono de la cultura popular; esto a efecto de sugerir la necesidad de actuar éticamente con respecto a la Inteligencia Artificial, que es materia hoy día tanto de investigación y usos recurrentes, como de debates necesarios.

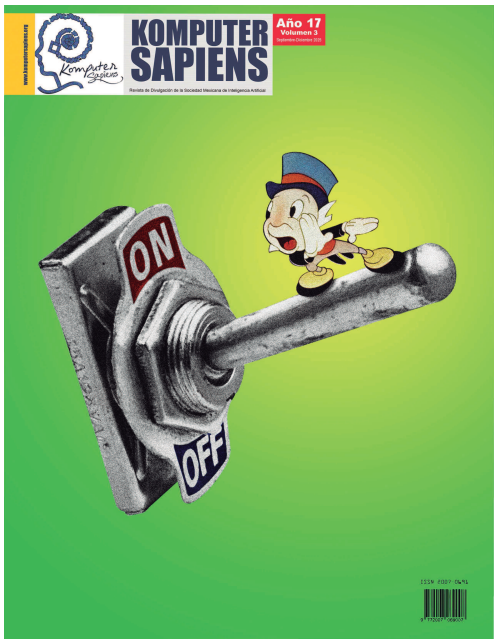
Replicamos a Pepe Grillo (Jiminy Cricket, en inglés), originalmente Il Grillo Parlante creado por Carlo Collodi en 1883 para su novela Las aventuras de Pinocho. En ella, el personaje de Pepe Grillo es quien advierte a Pinocho, el títere de madera que cobra vida gracias a los poderes de un Hada Azul, de las consecuencias de sus acciones. Pepe Grillo es para Pinocho su voz interior y su conciencia, es decir, aquello que permite distinguir lo correcto de lo incorrecto.

Hallamos una relación entre conciencia y ética, y es que aunque son conceptos distintos, ambos connotan de alguna manera la idea de responsabilidad. Entonces, empleamos al personaje de Pepe Grillo por ser ampliamente reconocible; además, éste evocaría sus atributos, consignados sobre todo en la película de Disney de 1940, la cual incluso las nuevas generaciones conocen hoy en día gracias a su difusión en diversas plataformas de streaming (transmisoras de contenido a través de Internet).

El switch de encendido de la imagen, que muestra las alternativas on y off, y que simula la nariz de Pinocho, refiere a las posibilidades de accionar, lo que, junto con el personaje, expresan el propósito que puede sintetizarse en la necesidad de reflexionar. Es como si Pepe Grillo (la conciencia) adjetivara al sustantivo (el switch, el algoritmo, el mecanismo, el dispositivo).

En el fondo, de lo que se trata es de aquello que señaló Potter Stewart, juez de la Suprema Corte de Estados Unidos, que dice "*Ethics is knowing the difference between what you have a right to do and what is right to do*" y cuya traducción es "Ética es saber la diferencia entre lo que tienes derecho a hacer y lo que es correcto hacer".

La lectura o interpretación de la imagen depende como siempre de referentes y de contexto, pero nos parece suficientemente ilustrativa de la intención.✷



© 2025 - Sociedad Mexicana de Inteligencia Artificial ISSN 2007-0691

e-Tlakuilo: Cartas de nuestros lectores

Karina Figueroa y Rosa María Valdovinos
etlakuilo@komputersapiens.org

Los autores agradecemos el interés de los lectores. A continuación, respondemos las preguntas que amablemente nos hicieron llegar.

Estudio longitudinal para el desarrollo de modelos predictivos de complicaciones crónicas de la diabetes mellitus tipo 2

Pregunta: Diana Karen Tapia Figueroa

En el artículo describen el proyecto en general. Muy interesante. ¿tienen resultados del modelo predictivo? de ser así, ¿cuál es el porcentaje de eficiencia?, respecto a la base de datos con la que trabajaron ¿se encuentra disponible en algún sitio de internet? ¿con cuántos datos trabajaron?

Estimada lectora, muchas gracias por las preguntas.
Para contextualizar la respuesta debemos decir que se desarrolló un estudio observacional, retrospectivo de cohorte longitudinal, en donde se siguieron los criterios de la declaratoria de STROBE. Se estudiaron a todos los derechohabientes del IMSS que han sido atendidos en las Unidades de Medicina Familiar de Morelia, Uruapan, Zamora, Lázaro Cárdenas, La Piedad y Zacapu. Esto es, pacientes con código de diagnóstico de DM2 entre el 1 de enero del 2006 y el 31 de diciembre del 2020, según la clasificación internacional de enfermedades (CIE). Este proyecto se desarrolló en Tres Fases:

1. Extracción y anonimización mediante la eliminación de elementos de información identificable (lo descrito en el artículo que detona esta pregunta).
2. Construcción de modelos predictivos.
3. Validación del modelo.

En la primera fase, la extracción de los datos de los expedientes electrónicos entregados por el Departamento Delegacional de la OOAD Michoacán se realizó al 100 % con una N= 2,255,994 expedientes. Se excluyeron aquellos expedientes de pacientes sin diabetes; pacientes con DM1 o solo con otro diagnóstico clínico; menores de 18 años; mujeres embarazadas o en puerperio tardío; sin datos registrados de IMC o expedientes duplicados. La anonimización de la información se realizó en 44,937 expedientes de pacientes con DM2. Se obtuvo la información de complicaciones registradas en el expediente electrónico de acuerdo con el CIE (Nefropatía: E11.2; Retinopatía: E11.3; Neuropatía: E11.4; Pie diabético: E11.5). Además de la edad del paciente al diagnóstico, sexo, IMC registrado durante 3 años y si tuvo diagnóstico de hipertensión arterial sistémica (HAS). La edad del paciente al diagnóstico de la DM2 fue de 48.37 ± 14.72 años. 58.7 % fueron mujeres y 41.3 % fueron hombres. La distribución del IMC es en promedio: bajo peso: 0.3 %; Peso normal: 16.3 %; Preobesidad: 39.8 %; Obesidad Grado 1: 24.2 %; Obesidad Grado II: 10.4 % y Obesidad Grado III: 4.4 %. Glucosa en ayuno: 136.85 ± 48.11 mg/dL. Presión arterial sistólica: 120.43 ± 6.26 mmHg. Presión arterial diastólica: 75.68 ± 3.67 mmHg La frecuencia de las complicaciones registradas en el expediente electrónico son de Nefropatía (E11.2): 11.1 %; Retinopatía (E11.3): 9.6 %; Neuropatía (E11.4): 6 %; Pie diabético (E11.5): 13.5 %. Sin complicaciones: 59.8 %.

Para la 2a fase, se construyeron y generaron varios modelos de predicción de riesgos en fenotipos complejos con varios grupos de algoritmos que son el estándar de la literatura en aprendizaje automático supervisado, que incluyen redes neuronales artificiales, redes bayesianas, máquinas de soporte vectorial y árboles de decisión. A continuación se muestra el área bajo la curva de los mejores 2 (de 22) modelos probados.

Modelos	Nefropatía E11.2	Retinopatía E11.3	Neuropatía E11.4	Pie diabético E11.5
Gaussian NB	0.87	0.87	0.87	0.87
Bernoulli NB	0.82	0.80	0.81	0.80

Dado que el primero modelo Gaussian NB tuvo una mayor área bajo la curva para identificar las probables complicaciones, se realizó la validación interna del modelo con todos los datos extraídos y anonimizados. Se obtuvo una sensibilidad del 97 % y una especificidad del 87 % para la identificación de las complicaciones crónicas en los pacientes con DM2.

Finalmente, para la Fase 3. Las fases anteriores tuvieron su propia validación interna para definir la sensibilidad y especificidad de los modelos construidos.

A la fecha, el modelo mejor evaluado está en proceso de implementarse en el sistema SIMF del IMSS nacional para los médicos de primer nivel.✷

© 2025 - Sociedad Mexicana de Inteligencia Artificial ISSN 2007-0691

Estado del IArte

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

Inteligencia Artificial Responsable

En esta columna hemos abordado temas como la Regulación de la IA (número 1,2024), su uso Razonable (número 3,2024), Ética en la IA (número 3,2023), y la IA Sustentable (número 3,2022). En esta ocasión el tema central es la IA Responsable (IAR). La responsabilidad es una cualidad humana que está ligada con la conciencia e intencionalidad, la capacidad de juicio moral y autónomo que conlleva el comprender contextos sociales, valores humanos y consecuencias éticas. Por otro lado, la responsabilidad legal recae sólo en personas físicas o jurídicas, en otras palabras empresas, asociaciones o fundaciones. Cualquier Inteligencia Artificial (IA) actúa según el diseño, los datos y los objetivos definidos por sus creadores o usuarios, es decir aquellos que controlan y supervisan su funcionamiento. Podemos entonces considerar que existen sistemas de IA responsables cuando éstos son creados y/o utilizados dentro de marcos éticos que aseguran justicia, transparencia, privacidad, y respeto por los derechos humanos, entre otros. La IAR busca maximizar los beneficios de la IA mientras se mitigan riesgos éticos, sociales y técnicos. Áreas relevantes que todo sistema de IAR debe considerar son: marcos normativos y éticos, transparencia y explicabilidad, mitigación de sesgos y equidad, privacidad y protección de datos, sostenibilidad ambiental, regulación global y cooperación, gobernanza corporativa y adopción empresarial, impacto social y laboral de la IA, derechos humanos, bien común, tendencias emergentes y desafíos, etc. A continuación, elegimos y reflexionamos sobre uno de los más relevantes y novedosos en la actualidad.

En la agenda 2030 de la UNESCO[1] se subraya que la responsabilidad de la IA siempre recae en los humanos que desarrollan, implementan o regulan estas tecnologías. Por lo anterior, minimizar los sesgos, garantizar una equidad de uso, educar sobre el uso ético de las IAs, crear y cumplir marcos tanto legales como técnicos para evitar daños siguen siendo fundamentales. Promover la inclusión, la diversidad cultural, la equidad de género y el desarrollo sostenible, dentro de marcos éticos globales, participativos y orientados al bien común. Hoy más que nunca debemos conservar nuestro lado humano. Utilizar la IA de manera responsable aplicando pensamiento crítico, verla como un complemento y no como único recurso, dosificar su uso, y ser conscientes de sus limitaciones. Realizar diagnósticos sobre el uso de las herramientas de IA y con base a ello proponer un plan estratégico y desarrollar programas que capaciten a interesados, p.e. alumnos y profesores en el caso de la educación, sobre su uso adecuado. Favoreciendo con ello, el desarrollo en los usuarios de habilidades en el uso de aplicaciones de IAR para evitar dependencia tecnológica. Dicho de otra manera, no transferir nuestra responsabilidad a la IAR. Pedirle a una IA apoyo para analizar un problema y nosotros realizar una síntesis que eventualmente nos lleve a una solución, no es lo mismo que pedirle que haga la síntesis y dé solución al problema, y menos cuando hay temas delicados que debemos considerar.

Finalmente, podemos decir que la IAR se caracteriza por buscar un equilibrio entre regulación, estándares técnicos, prácticas operacionales e innovación ética. Las organizaciones avanzan desde principios hacia implementación efectiva: gobernanza integrada, participación de diversos actores, monitoreo constante y rendición de cuentas. En sí, la IAR requiere colaboración entre gobiernos, empresas y sociedad, especialmente las grandes empresas de tecnología, que por el momento parece que sí están comprometidas con el tema. Cada país debe establecer sus propias iniciativas hacia una IAR. En México, una de ellas es la reciente creación del *Laboratorio Nacional de Inteligencia Artificial* (LNIA[2]) el cual es una propuesta de la Presidencia de la República que a través de la Agencia de Transformación Digital y Telecomunicaciones (ATDT), así como de la Secretaría de Ciencia, Humanidades, Tecnología e Innovación (SECIHTI). Su objetivo es fomentar la colaboración intersectorial, interinstitucional e interdisciplinaria para acelerar la incorporación, el aprovechamiento y los beneficios de la Inteligencia Artificial en México.

Esta columna ha sido creada por humanos, sin el apoyo de IA generativa alguna.✷

[1] Disponible también es enpañol, ver https://unesdoc.unesco.org/ark:/48223/pf0000381137
[2] https://secihti.mx

© 2025 - Sociedad Mexicana de Inteligencia Artificial ISSN 2007-0691

Sakbe

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

El camino hacia una IA responsable

La inteligencia artificial (IA) se ha convertido en un tema central en el ámbito tecnológico, político y social. Sin embargo, el mundo necesita reglas para que la IA beneficie a la humanidad. Por un lado, la ***ética de la IA*** define valores mientras que la ***IA responsable*** los operacionaliza: muestra cómo aplicarlos y cómo rendir cuentas. La investigación y el debate público funcionan como engranajes que conectan ambas capas: verifican y legitiman el paso de los principios a la práctica.

Ética de la IA ⇒ La UNESCO tiene la visión de «centrar la IA en el ser humano» mediante cuatro ejes (dignidad humana, equidad, sostenibilidad y gobernanza efectiva) que actúan como brújula normativa para cualquier iniciativa tecnológica. En su documento *Recomendación sobre la Ética de la Inteligencia Artificial* (2021) insiste en que la ética debe incorporarse desde el diseño y mantenerse durante todo el ciclo de vida, de modo que cada sistema pueda demostrar de forma auditable que respeta los derechos humanos y genera beneficios sociales.

https://unesdoc.unesco.org/ark:/48223/pf0000381137_spa.locale=en

IA Responsable ⇒ Un marco regulatorio eficaz combina *ley dura* (ej. una ley federal) y *ley blanda* (normas voluntarias). Canadá lo ejemplifica: aplica la *Directiva obligatoria sobre decisiones automatizadas y la Evaluación de Impacto Algorítmico*, opcional fuera del ámbito federal. Para aterrizar los valores éticos en prácticas auditables, la ISO/IEC 42001, primer estándar de gestión de IA, exige documentar el ciclo de vida del modelo, gestionar riesgos y garantizar la trazabilidad de datos, convirtiéndose en la guía operativa para una IA responsable.

https://www.iso.org/obp/ui/en/#iso:std:iso-iec:42001:ed-1:v1:en

Ética de la IA ⇔ IA Responsable: Investigación ⇒ En México, el *AI for Good Lab de C Minds* audita algoritmos con ≪caja gris*gg* (solo entradas y salidas) para detectar sesgos (ej. sistemas de reclutamiento que discriminan por género); impulsó Open Loop México (2021), piloto regulatorio de empresas cuyas reglas de transparencia se publicaron en 2023, y el *hub fAIr LAC Jalisco* (2020-2023), que dejó una biblioteca de IA responsable y Retina-IA, sistema que acortó el periodo de diagnóstico de retinopatía diabética.

hhttps://www.cminds.co/ai-for-good-lab

Ética de la IA ⇔ IA Responsable: Debate público ⇒ Los debates sobre IA buscan regularla, ética y legalmente, para frenar riesgos—privacidad, sesgos y desinformación—sin frenar su aporte al desarrollo sostenible. La UNESCO los encauza en foros anuales: París 2023 (acceso vs. control algorítmico), Eslovenia 2024 (cooperación y supervisión) y Bangkok 2025 (revisión de avances y nuevas acciones). En México, la discusión sigue dispersa: falta articular a gobierno, academia, empresas y sociedad civil, definir estándares claros y concientizar a la población sobre los riesgos.

https://comunicacionsocial.senado.gob.mx/informacion/comunicados/8982-presentan-en-el-senado-agenda-nacional-de-la-inteligencia-artificial-para-mexico-2024-2030

© 2025 - Sociedad Mexicana de Inteligencia Artificial ISSN 2007-0691

ARTÍCULO ACEPTADO

Robots y su aplicación en la rehabilitación de Hombro Congelado

Jimena Sánchez Guerrero, Edwin Sebna Albarrán García, Adriana H. Vilchis González y Juan C. Ávila Vilchis

Introducción

El hombro, está compuesto por cinco articulaciones. Entre ellas, la articulación glenohumeral, encargada de unir la cavidad glenoidea con la cabeza del húmero con apoyo de la cápsula glenohumeral y ligamentos asociados, permite realizar movimientos de flexión, extensión, abducción, aducción, rotación interna, rotación externa y circunducción [1].

Hombro Congelado

El hombro congelado (HC) es una patología común, poco investigada; se caracteriza por la contractura de la cápsula glenohumeral, generando dolor y restricción de movimiento. Suele presentarse alrededor de los 50 años, principalmente en mujeres; aumenta el riesgo en personas con antecedentes de cirugía de hombro, pecho o vértebras cervicales, con diabetes o enfermedades tiroideas. La prevalencia de esta patología es del 2 % al 5 % de la población en general [2].

La progresión de la patología se divide en tres etapas: 1) Fase de congelación, entre 2 a 9 meses, predomina el dolor y se pierde gradualmente la movilidad; 2) Fase congelada, entre 4 y 12 meses, el dolor y el movimiento se reducen progresivamente, 3) Fase de deshielo, entre 5 y 26 meses, el paciente presenta mejora en la movilidad articular [3].

Los tratamientos buscan reducir el dolor y preservar la movilidad; se dividen en quirúrgicos y no quirúrgicos. Estos últimos incluyen la fisioterapia, que acorde a la fase de la enfermedad emplea cinesiterapia, electroterapia analgésica, ultrasonido, etcétera. La cinesiterapia consiste en movilidad pasiva y activa para recuperar función en la articulación [1].

Dentro de la movilidad pasiva, los ejercicios pendulares con el brazo extendido a favor de la gravedad (ver figura 1) favorecen la movilización temprana de la articulación y disminuyen el dolor. Existen tres tipos: Codman, Chandler y Sperry [4].

Posteriormente, aplican ejercicios de movilidad activa para fortalecer la musculatura; puede incluir una fuerza de oposición que genere resistencia [5].

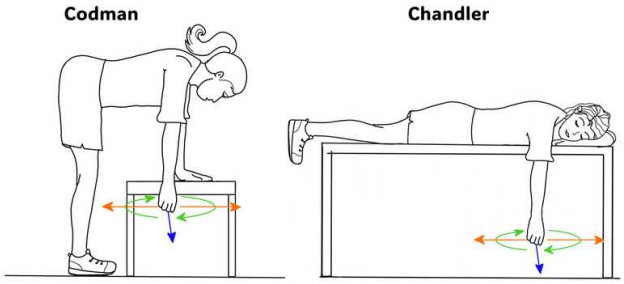


Figura 1. Ejercicios pendulares, movimientos de Codman y Chandler.

Robots en la rehabilitación

La red europea de monitoreo de prospectiva define a los robots de atención a la salud como "sistemas capaces de realizar acciones mecatrónicas coordinadas (fuerza o movimiento) para apoyar la [...] rehabilitación de los pacientes..."; describe 3 categorías principales [6]:

1. Robots quirúrgicos.
2. Robots de asistencia.
3. Robots de servicio del cuidado de la salud.

Los robots de asistencia coexisten como robots de asistencia social de terapia y robots de asistencia física de rehabilitación de usuario; ambos se aplican en entornos terapéuticos como parte de la robot-terapia. La diferencia es que los primeros se orientan a terapia cognitiva o psicológica mediante interacciones sociales con robots, y los segundos hacia una rehabilitación física, como la requerida en la rehabilitación basada en cinesiterapia del HC [6].

La literatura, particularmente lo presentado en [7], [8], [9], [10], clasifica a los robots de rehabilitación para extremidad superior (ES) según su estructura en dos vertientes principales (ver Figura 2):

A) Exoesqueleto (E)

Inspirados en la anatomía del cuerpo humano. Su característica principal es el contacto directo entre las

© 2025 - Sociedad Mexicana de Inteligencia Artificial ISSN 2007-0691

articulaciones y el exoesqueleto. Adaptado a la anatomía transmite el movimiento [7].

B) Efector final (EF)

Considerados movilizadores externos que interactúan con un solo punto del usuario a través del elemento terminal del robot (ej. agarradera); proporcionan movimientos amplios y buscan ejercer la fuerza adecuada para guiar el movimiento requerido durante la rehabilitación [7] [8] [9].

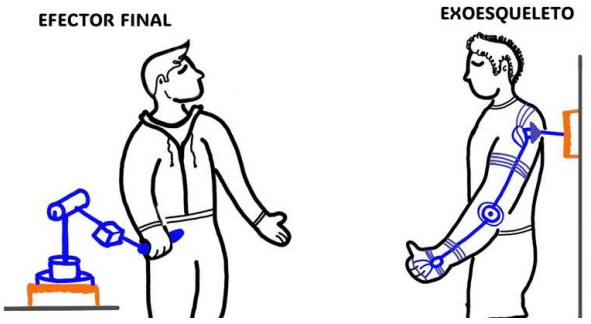


Figura 2. Clasificación de los robots de miembro superior.

La década de 1990 marcó el inicio del desarrollo e investigación de los robots de rehabilitación [9]. Algunos robots que consideran el movimiento del hombro, no enfocados a rehabilitar HC, recopilados por [10], y algunas características que contemplan (ver Tabla 1) se sintetizan a continuación:

Dampace

Exoesqueleto desarrollado en 2007. El paciente participa de forma activa en los movimientos de la ES, dando seguridad al paciente durante su uso. Para compensar el peso de la ES cuenta con un sistema separado de compensación de la gravedad.



Figura 3. Dampace [10].

Therapy system Wilmington Robotic Exoskeleton (T-WREX)

Desarrollado para ambos brazos, uso en casa y como entrada para interactuar con entornos virtuales. El software simula actividades funcionales que mejoran el movimiento de la ES tras lesiones neurológicas (ej. hemiparesia crónica).

Se opone al peso del brazo con bandas elásticas; disponible comercialmente bajo el nombre Armeo Spring, y la versión Armeo Spring Pro.



Figura 4. 4 T-WREX [10].

Maryland-Gerogetown-Army (MGA)-exoskeleton

Desarrollado en 2005. Opera en dos modos: 1) modo de realidad virtual, simula y apoya en tareas de la vida diaria y 2) modo de terapia física, rota el brazo en un eje arbitrario a través del hombro con resistencia, y soporte seleccionado por el terapeuta.

Por seguridad, la articulación del robot asociada al codo ajusta el rango de torque con un embrague. Un torque excesivo, desacopla el eje del codo.

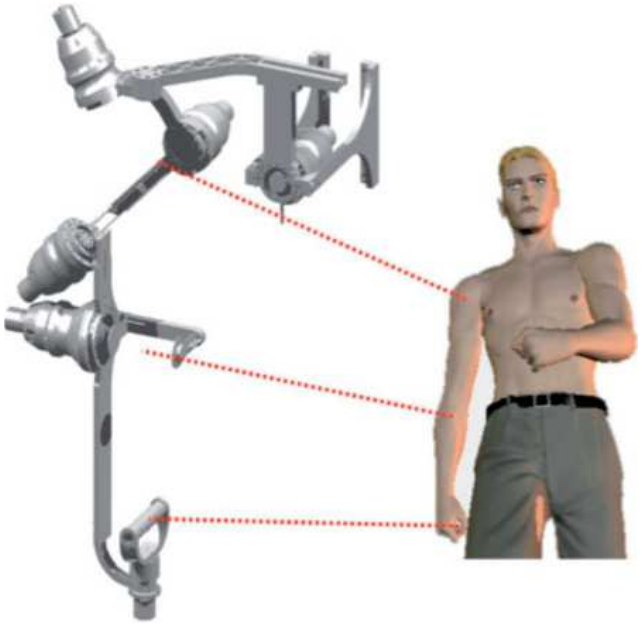


Figura 5.MGA exoskeleton [10].

© 2025 - Sociedad Mexicana de Inteligencia Artificial ISSN 2007-0691

L-EXOS

Exoesqueleto de retroalimentación de fuerza para el brazo derecho, desarrollado en 2007. El paciente puede interactuar de forma básica con un entorno virtual, además, la ubicación de motores en la base aligera el peso para el paciente.



Figura 6. L-EXOS [10].

ARMin

Desarrollado entre 2003 y 2006. El robot asiste al paciente y proporciona medidas cuantitativas del rango de movimiento (ROM), fuerza muscular y precisión espacial de la posición de la mano.

Principalmente brinda: 1) movilización pasiva y activa en una trayectoria predefinida, 2) terapia de juego con tecnología de realidad virtual y 3) entrenamiento para actividades de la vida diaria.

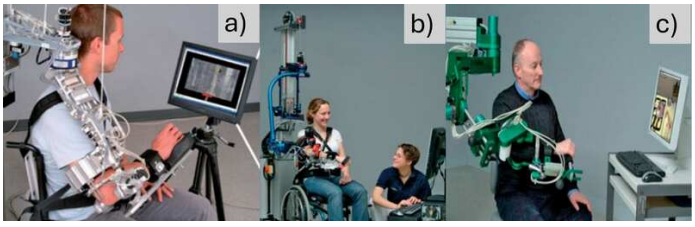


Figura 7.a) ARMin I, b) ARMin II, c) ARMin III [10].

Presenta tres evoluciones:

ARMin I: El brazo se conecta a un robot tipo efector final y el antebrazo a un exoesqueleto.

ARMin II: Permite la pronación-supinación del antebrazo y la flexión-extensión de muñeca.

ARMin III: Configurable para ambos brazos, compensa parte del peso del exoesqueleto con un resorte. Permite ajustar el centro de la articulación glenohumeral con punteros láser como referencia.

Comercializado con el nombre de Armeo Power.

Mit Manus

De los primeros robots desarrollados para rehabilitación. En operación desde 1994.

Permite tratar solo la zona afectada de la ES con sus 4 módulos: 1) módulo activo planar; 2) módulo activo vertical; 3) módulo activo de muñeca; 4) módulo de agarre pasivo.

Con agarraderas hechas a medida para la sujeción del paciente al robot. Es portátil y está diseñado para ser seguro en contacto cercano con personas durante rehabilitación clínica.



Figura 8. Mit Manus [10].

Mirror image motion enabler (MIME)

Diseñado para neuro-rehabilitación de hombro y codo. Probado por primera vez en 2002 con un sistema Puma 260, y reemplazado por el Puma 560.

Con el paciente sentado en una silla de ruedas frente a una mesa de altura ajustable con correas que limitan el movimiento del torso, lo apoya con 4 modos de operación: 1) modo pasivo, 2) modo activo-asistido, 3) modo activo-resistido y 4) modo bilateral, donde el robot asiste la extremidad afectada reflejando el movimiento realizado con la extremidad sana.

© 2025 - Sociedad Mexicana de Inteligencia Artificial ISSN 2007-0691

El hombro congelado es una patología prolongada que afecta las actividades de la vida diaria.



Figura 9. MIME [10].

GENTLE's

Presentado en 2003. El paciente sentado con el brazo suspendido y solo la muñeca conectada a una ortesis en comunicación a una interfaz gráfica. El terapeuta selecciona los ejercicios adecuados que generan trayectorias entre dos puntos.



Figura 10. GENTLE's [10].

Assisted Rehabilitation and Measurement guide (ARM)

Herramienta de diagnóstico y terapéutica que evalúa y entrena el ROM y la generación de fuerza del brazo después de accidentes cerebrovasculares y otras lesiones cerebrales Su diseño con un solo motor disminuye la carga en el brazo.

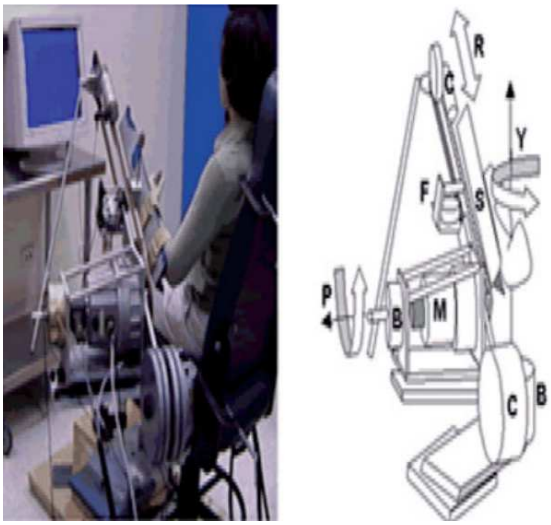


Figura 11. ARM [10].

REHAROB

Completado en 2003. Proporciona fisioterapia pasiva para el brazo hemiparético espástico, utiliza dos robots industriales: uno mueve la parte superior del brazo y el otro la inferior. El terapeuta acompaña libremente el primer movimiento del paciente, el robot memoriza y realiza las siguientes repeticiones.



Figura 12. REHAROB [10].

© 2025 - Sociedad Mexicana de Inteligencia Artificial ISSN 2007-0691

Tabla 1. Comparativa de las características de los robots [10].

Robot	TR +	Tipo de terapia	Movimientos	Retroalimentación	GDL
Dampace	E	Activa	Movimientos de hombro y codo.	Interacción con videojuegos	4
Therapy system Wilmington Robotic Exoskeleton (T-WREX)**	E	Activa	Movimientos de ES.	Interacción con entorno virtual. Sensor de presión en la mano.	5
Maryland-Gerogetown-Army (MGA)-exoskeleton	E	Activa y activa-resistida.	Rotación del hombro. Movimiento del codo y la escápula.	Entorno virtual. Sensor de fuerza/torque en la agarradera.	5
L-EXOS	E	Pasiva y activa	Movimientos de ES.	Interacción con entorno virtual. Sensor de fuerza.	5
ARMin I	SE	Pasiva y activa.	Movimiento de hombro en tres dimensiones y flexión-extensión de codo.	Interacción con juegos.	4
ARMin II	E	Pasiva, activa-asistida y activa.	Pronación-supinación del antebrazo y la flexión-extensión de muñeca.	Interacción con juegos.	6
ARMin III **	E	Terapia pasiva, activa asistida y activa.	Flexión-extensión de hombro, codo y muñeca; Abducción-aducción y rotación del hombro, pronación-supinación del antebrazo y apertura y cierre de la mano (opcionalmente).	Interacción con juegos.	6
Mit Manus **	EF	Activa y pasiva.	Movimientos horizontales de muñeca y antebrazo, Movimientos verticales y de fuerza.	Interacción con videojuegos. Sensor de fuerza.	7 6 activos 1 pasivo
Mirror image motion enabler (MIME) *	EF	Pasiva, activa asistida, activa resistida y activa	Movimientos de hombro, codo. Movimiento bilateral del brazo no afectado y el afectado.	Sensor de fuerza/torque.	6
GENTLE's	EF	Pasiva, activa asistida y activa resistida	Movimientos de ES.	Técnicas de realidad virtual y háptica.	6 3 activos 3 pasivos.
Assisted Rehabilitation and Measurement guide (ARM)	EF	Pasiva y activa	Guía linealmente la ES y extensión-flexión, abducción-aducción pasiva del hombro.	Monitor que informa al usuario sobre el movimiento y la generación de fuerza en el hombro.	3 1 activo 2 pasivos
REHAROB *	EF	Pasiva.	Movimientos de la ES definidos por el terapeuta.	NA	12

+ *Tipo de Robot, *Utiliza robots industriales; **Se comercializa; E: Exoesqueletos; SE: Semi exoesqueleto; EF: Efector final; ES: Extremidad superior; GDL: Grados de libertad; NA: No aplica / Falta información.*

© 2025 - Sociedad Mexicana de Inteligencia Artificial ISSN 2007-0691

Tabla 2. Requerimientos identificados para un robot rehabilitador de hombro congelado.

Población	Adultos mayores y personas laboralmente productivas con dolor de hombro y disminución de ROM del hombro. Considerar un rango de medidas antropométricas (Estatura, longitudes de brazo y mano) y la masa del brazo.
Postura	Flexión anterior de cadera (45° a 90°) en bipedestación o decúbito prono en una camilla. Brazo en extensión a favor de la gravedad, manteniendo la distención del húmero. Posición neutra de la mano con agarre permitido.
Movimientos	Flexión, extensión, abducción, aducción y circunducción del brazo. Progresivos y repetitivos y velocidad (baja o media).
Espacio de uso	Dentro de un área de rehabilitación física (ej. Consultorio de fisioterapia). Portabilidad Seguridad en interacción robot – paciente/fisioterapeuta.

Añadir peso adicional en la articulación podría resultar en aumento de dolor.

Conclusiones

Considerando la literatura revisada, múltiples robots de rehabilitación muestran ser herramientas valiosas en el tratamiento y movilización de la ES, principalmente en pacientes con accidentes cerebro vasculares o lesiones cerebrales.

Sin embargo, para acoplarse al HC, una patología exclusivamente musculoesquelética; cuyo tratamiento se basa en movilizaciones progresivas de hombro y la reducción del dolor con ayuda de la postura del paciente, es adecuado considerar algunos requerimientos para apuntar a una correcta rehabilitación. La tabla 2 incluye los principales.

De estos requerimientos principales derivan otros necesarios o deseables, como la elección del tipo de estructura y GDL que permita al robot asistir los movimientos deseados, adecuarse en dimensiones a la población objetivo y al espacio de uso, además, en este caso, a la postura del paciente.

Los exoesqueletos, como el MGA, podrían mantener la postura del brazo necesaria en los movimientos de Codman y Chandler, sin embargo, también añadir peso adicional en la articulación y el brazo, lo que podría resultar en incomodidad y aumento del dolor.

Mientras que un EF guía la ES sin relativamente agregar peso, por el contrario, amplia la libertad de movimiento de las articulaciones del paciente. Asimismo, se considera más simple planificar trayectorias y aplicar métodos de control.

Al considerar la aplicación del robot durante la fase de congelación, buscando atenuar la gravedad de la fase congelada y la elongación de la patología con los movimientos progresivos considerados adecuados por el terapeuta; se deben contemplar los modos de operación a fines para lograr el objetivo, por ejemplo, un modo orientado a la terapia pasiva permitiría iniciar el tratamiento de forma temprana con las movilizaciones suaves y controladas de los ejercicios de Codman y Chandler. Con el progreso a ejercicios de fuerza, se pueden contemplar los modos de operación activo-resistido o activo, donde el paciente ejerce fuerza o genera el movimiento.

Como se menciona, algunos de los robots emplean robots industriales, adaptados para el trato directo con humanos. Al proponer la estructura y su espacio de trabajo, se debe tomar en cuenta, el entorno de trabajo (Área de terapia) y el usuario para minimizar riesgos en la interacción.

Sería adecuado que, como el ARMIN III, se ajuste al mayor rango posible de dimensiones antropométricas de la población objetivo para no causar dolor o daño a la articulación, inhibiendo el movimiento y afectando la terapia.

Se podría añadir retroalimentación con realidad virtual para motivar al paciente a finalizar la terapia, debido a que, la rehabilitación de HC presenta deserción al tender a ser repetitiva y prolongada.

En resumen, los robots de rehabilitación son una prometedora adición a las herramientas disponibles para el tratamiento del HC, permitiendo la movilización controlada, guiada y repetitiva que ayude a mejorar el ROM, evitar el gesto de compensación de los pacientes y, reducir la rigidez y el dolor.

Sin embargo, para llegar a su implementación clínica, se debe contar con un enfoque que integre en el desarrollo del robot, los aspectos vislumbrados anteriormente con resultados de otros estudios que sustenten características terapéuticas, antropométricas y de diseño para lograr la aplicación segura, eficiente y eficaz de un dispositivo robótico en la rehabilitación del HC.✷

© 2025 - Sociedad Mexicana de Inteligencia Artificial ISSN 2007-0691

REFERENCIAS

1. Vanessa, E. O. (2021). Hombro Congelado. Tesis de grado. Universidad Inca Garcilaso de la Vega.
2. Ramirez, J. (2019). Adhesive Capsulitis: Diagnosis and Management. *American Family Physician*, 99(5), 297–300.
3. Morales, J. (2021). Tratamiento fisioterápico en pacientes con hombro congelado. Revisión bibliográfica. Tesis de grado, Grado en Fisioterapia. Universidad Miguel Hernández.
4. Dávila De Santis, A. D. (2016). Tratamiento Fisioterapéutico en Capsulitis Adhesiva de Hombro. Análisis de caso. Tesis de grado. Universidad Laica Eloy Alfaro de Manabí. Manta, Ecuador.
5. Chamorro, Y. M. P. (2019). Ejercicios de Codman en el tratamiento de lesiones del hombro. Centro de Salud Santa Rosa. Riobamba 2018 – 2019. Tesis de grado. Licenciatura en Ciencias de la Salud en Terapia Física y Deportiva. Universidad Nacional de Chimborazo. Riobamba, Ecuador.
6. Fosch-Villaronga, E. y Drukarch, H. (2021). On healthcare robots. Concepts, definitions, and considerations for healthcare robot governance. *Leading Fellows Marie Curie COFUND fellowship*.
7. Garcia-Gonzalez, A., Fuentes-Aguilar, R. Q., Salgado, I., y Chairez, I. (2022). A review on the application of autonomous and intelligent robotic devices in medical rehabilitation. *Journal of the Brazilian Society of Mechanical Sciences and Engineering*, 44(9).
8. Dupont, P. E. et al. (2021). A decade retrospective of medical robotics research from 2010 to 2020. *Sci. Robot*, 6, 8017.
9. Meng, Q., Xie, Q., y Yu, H. (2019). Upper-Limb Rehabilitation Robot: State of the Art and Existing Problems. 1-4.
10. Babaiasl et al. (2016). A review of technological and clinical aspects of robot-aided rehabilitation of upper-extremity after stroke. *Disability and Rehabilitation: Assistive Technology*, 11(4):263-80, 1–18.

SOBRE LOS AUTORES



Jimena Sánchez Guerrero es egresada de la licenciatura en Bioingeniería Médica de la Universidad Autónoma del Estado de México. Se desempeña actualmente en el desarrollo de una plataforma robótica efector final para rehabilitación de hombro congelado. Sus áreas científicas de interés incluyen el Diseño Mecánico y Robótica Médica.



Edwin Sebna Albarrán García es egresado de la licenciatura en Bioingeniería Médica de la Universidad Autónoma del Estado de México. Se desempeña actualmente en el desarrollo de una plataforma robótica efector final para rehabilitación de hombro congelado. Sus áreas científicas de interés incluyen dispositivos para rehabilitación física, medicina en el deporte y robótica médica.



Adriana H. Vilchis González obtuvo el grado de doctor en Imagen, Visión y Robótica en el Instituto Nacional Politécnico de Grenoble, Francia. Actualmente es profesor investigador en las Facultades de Ingeniería y de Medicina de la Universidad Autónoma del Estado de México. Sus intereses científicos incluyen Robótica Médica, Realidad Virtual, Robots Suaves y Sistemas Mecatrónicos.



Juan C. Ávila Vilchis obtuvo el grado de doctor en el Instituto Nacional Politécnico de Grenoble, Francia. Actualmente es profesor investigador en las Facultades de Ingeniería y de Medicina de la Universidad Autónoma del Estado de México. Sus intereses científicos incluyen Modelado, Análisis y Control de sistemas no lineales y Robótica Médica.

© 2025 - Sociedad Mexicana de Inteligencia Artificial ISSN 2007-0691

ARTÍCULO ACEPTADO

Innovación en la rehabilitación de hombro mediante juegos de realidad virtual

Citlali Irais Cano Del Razo, Juan Manuel Jacinto Villegas, Monserrat Ríos Hernández y Adriana Herlinda Vilchis González

Introducción

Los trastornos del hombro afectan entre el 16 y 26 % de la población adulta, causando dolor y pérdida funcional que impacta la calidad de vida [1][2]. La rehabilitación tradicional busca recuperar flexibilidad y fuerza muscular [1], a partir de ejercicios repetitivos, pero presenta inconvenientes como una evaluación subjetiva del progreso del paciente y una baja adherencia debido a la monotonía de las sesiones [2], volviendo menos efectivo el tratamiento [3].

Los juegos serios, que son videojuegos con un fin distinto del entretenimiento, hacen más atractiva y motivadora la rehabilitación [3]. Además, la realidad virtual (RV), que consiste en un entorno 3D generado por computadora [7], y la realidad aumentada (RA), que permite superponer objetos virtuales al mundo real [1], ofrecen entornos inmersivos y multisensoriales, transformando la experiencia de rehabilitación con mayor eficacia y compromiso emocional [1][4] comparada con la rehabilitación tradicional.

Este artículo explora el uso actual de juegos serios con RV, RA y diferentes dispositivos en la rehabilitación de hombro. Para ello se realizó una búsqueda en las bases de datos: Google Académico, IEEE, PubMed, Scopus y Web of Science, utilizando palabras clave como *Serious games*, *Rehabilitation*, *Shoulder*, *Leap Motion*, *Virtual Reality*, *Augmented Reality*, *Kinect*, *Custom* y *HDM*. Abarcando desde 2016 hasta 2024.

Se han clasificado tres grupos de juegos serios para la rehabilitación del hombro, de acuerdo con el dispositivo de interacción que emplean: i) los que utilizan Kinect, entre los que destacan The Airplane, *The Ball y The Slicer* de Belotti et al., el juego de Bai et al., y el juego de Arif et al.; ii) juegos serios que emplean visores montados en la cabeza (HMD por las siglas en inglés "Head Mounted Displays"), como el juego de Moreira et al., el de Longo et al., y *Painting Discovery* de Condino et al.; y iii) los que utilizan otros dispositivos, como *Painting Discovery* con Leap Motion Control de Viglialoro et al., el uso de PlayBall®en el juego de Rizzato et al., y el sistema de captura de movimiento OptiTrack en *Sleeve AR* de Sousa et al.

Juegos serios con Kinect

El Kinect desarrollado por Microsoft, es un dispositivo compuesto de una cámara RGB, un sensor de profundidad, un micrófono multidireccional y una serie de motores y acelerómetros que permiten la captura y seguimiento del cuerpo sin necesidad de marcadores físicos. Puede rastrear 20 articulaciones en tiempo real e integrarse con motores de juegos como Unity3D y Unreal Engine [6].

Un sistema que lo utiliza es el desarrollado por Belotti et al. [2], cuyo objetivo es tratar trastornos musculoesqueléticos del hombro a partir de tres juegos serios: *The Airplane*, en el que el paciente mueve un avión con sus movimientos del brazo, *The Ball*, en el que se coloca una pelota en un cubo y *The Slicer*, en el que el paciente corta sandías (ver Figura 1). Utiliza el Azure Kinect DK, el cual se distingue por su sensor de profundidad, Time-of-Flight y su diseño compacto para captar de manera precisa los movimientos de abducción, flexión y rotación externa del hombro. El entorno proporciona retroalimentación visual, acompañada del almacenamiento de datos en dos aplicaciones, una para fisioterapeutas para programar sesiones, ajustar la dificultad y dar seguimiento a los resultados, y otra para pacientes, que presenta un calendario de actividades. Los juegos se instalaron en una computadora con GPU Nvidia GeForce RTX 2060 y las aplicaciones en dos tabletas Samsung Galaxy Tab A [2].

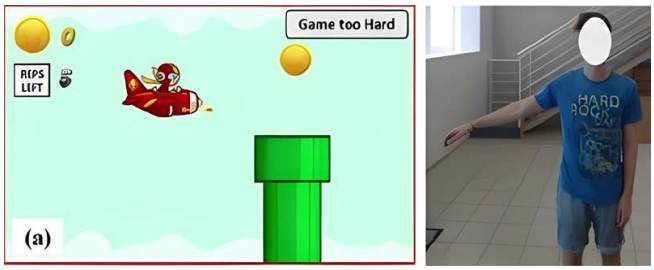


Figura 1. The Airplane, tomada de [2].

En [6] se presenta un artículo de revisión en el cual se destacan dos artículos. El primero, presentado por Bai et al. [6], es un juego para pacientes con accidente cerebrovascular. Los pacientes recolectan peces y monedas mediante movimientos de abducción, aducción y flexión de hombro como se observa en la Figura 2. Fue desarrollado con Unity3D y utiliza Kinect v1.0, que tiene un sensor de profundidad con proyección de luz infrarroja y una cámara RGB. Ofrece retroalimentación visual y auditiva. Al ejecutar el movimiento se graba un video que se guarda en la nube junto a la evaluación de parámetros

© 2025 - Sociedad Mexicana de Inteligencia Artificial ISSN 2007-0691

de rendimiento. Permite realizar sesiones en el hogar al facilitar la supervisión remota por parte de terapeutas [6].



Figura 2. Juego serio de Bai et al. tomada de [6].

El segundo artículo se enfoca en el tratamiento del hombro congelado que describe Arif et al. [6], también utilizando el dispositivo Kinect. En este juego, los pacientes mueven el brazo según las instrucciones proporcionadas como se muestra en la Figura 3. Ofrece retroalimentación visual y evalúa el movimiento con el ángulo máximo alcanzado [6].

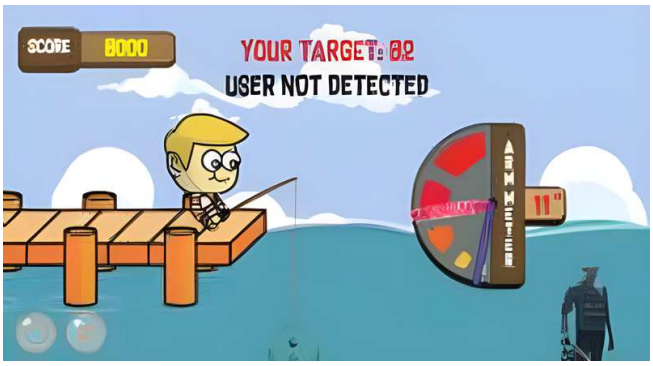


Figura 3. Juego serio de Arif et al. tomada de [6].

Los juegos antes mencionados tienen una cadena cinemática de retroalimentación abierta, ya que el paciente no recibe retroalimentación kinestésica ni háptica, dificultando la propiocepción de la orientación y ubicación espacial en el ambiente virtual.

Juegos serios con visores montados en la cabeza (HMD)

En la actualidad los HMD y sus joysticks han tomado relevancia en el ámbito de la RV, lo que ha llevado a implementarlos en diversas áreas, entre ellas en el área de la rehabilitación.

Entre las aplicaciones actuales está la de Moreira et al. [3] donde utilizan el Oculus Quest 2 (OQ2) con su juego serio centrado en la dinámica de cortar un tronco con una sierra para hacer rotación externa pasiva del hombro, como se muestra en la Figura 4. Desarrollado en Unity3D, con el Oculus SDK para Unity3D. Tiene RV inmersiva y permite retroalimentación háptica ya que los joysticks del OQ2 vibran de acuerdo a la interacción con el entorno. Se almacenan datos del movimiento en un servidor SQL como la rotación, posición y marcas de tiempo, e información del paciente como la edad, condiciones médicas y número de sesiones asistidas [3].

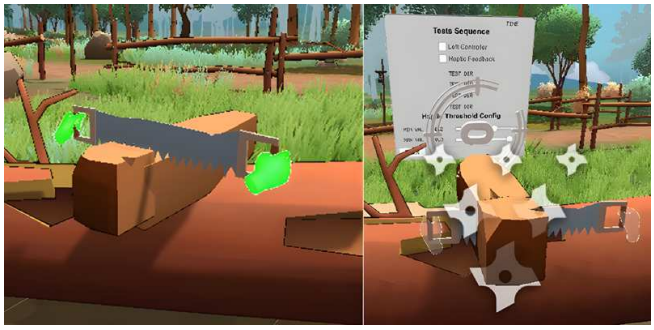


Figura 4. Juego serio de Moreira et al. tomada de [3]

Longo et al. [7], también utilizaron el OQ2 para su juego serio centrado en la rehabilitación de desgarros del manguito rotador y osteoartritis, implementando un programa de rehabilitación virtual creado a partir de un protocolo de la Sociedad Estadounidense de Terapeutas de Hombro y Codo. En este, los pacientes realizan actividades que requieren movimientos de flexión, abducción y rotaciones del hombro, como obtener manzanas de un árbol y atrapar frutas y mariposas, como se observa en la Figura 5. La retroalimentación proporcionada es visual y háptica mediante vibraciones de los joysticks del OQ2 [7].

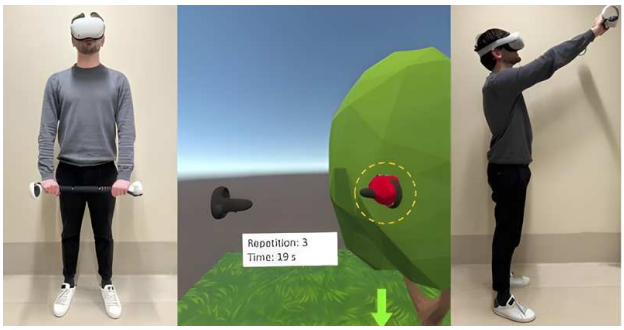


Figura 5. Juego serio de recolectar manzanas de Longo et al., modificada de [7].

Igualmente, Condino et al. [4], presentaron el juego *Painting Discovery* como la primera aplicación de juego de rehabilitación de realidad aumentada portátil para la rehabilitación del hombro que permite un seguimiento del movimiento sin marcadores en tiempo real. Este juego se desarrolló con Microsoft HoloToolkit y Unity3D.

La rehabilitación tradicional suele ser monótona y tiene baja adherencia.

Los pacientes realizan abducción horizontal del hombro siguiendo una trayectoria. Si lo hacen bien, los mosaicos muestran parte de una pintura; si no, se vuelven rojos (ver Figura 6). La retroalimentación es visual y auditiva, y los datos del ejercicio, como el número de errores, la puntuación y el tiempo, se almacenan en un archivo de texto. El movimiento se evalúa en función del ROM ideal. Los fisioterapeutas pueden configurar la trayectoria, el tiempo máximo de finalización y el número de repeticiones, proporcionando una rehabilitación ajustable y específica [4].

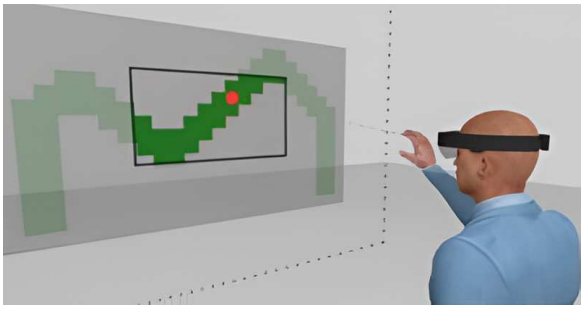


Figura 6. Funcionamiento del *Painting Discovery*, tomada de [4].

Otros dispositivos de interacción en juegos serios

Existen otros dispositivos utilizados en los juegos serios que ofrecen distintas funcionalidades y ventajas que pueden ser aprovechadas para diseñar ejercicios terapéuticos efectivos y motivadores. Tales como el desarrollado por Viglialoro et al. [8], que es una adaptación del juego Painting Discovery mencionado anteriormente, utilizando el dispositivo Leap Motion Control como se muestra en la Figura 7. Este juego mueve un cursor virtual a lo largo de una trayectoria, mediante la aducción horizontal del hombro. Los datos del ejercicio se almacenan en la nube y se evalúan con base en la velocidad, aceleración y jerk, que mide cómo cambia la aceleración durante el movimiento y permite analizar si el movimiento es fluido o no [8].

Rizzato et al. [5] evaluaron la eficacia del dispositivo PlayBall, que es similar a una pelota y mide la fuerza y movilidad del hombro mediante contracciones isométricas y sus movimientos, para rehabilitar patologías como síndrome de pinzamiento, capsulitis y lesiones tendinosas. Además, incluye un juego serio en el que el paciente maneja una nave espacial para golpear estrellas y evitar asteroides usando el PlayBall, que ofrece retroalimentación visual y háptica en tiempo real como se observa en la Figura 8. El estudio comparó esta terapia con la tradicional en 22 adultos con patologías del hombro, concluyendo que son igualmente efectivas, mejorando fuerza, reduciendo dolor, y aumentando satisfacción, autoeficacia e intención de continuar el ejercicio del paciente [5].

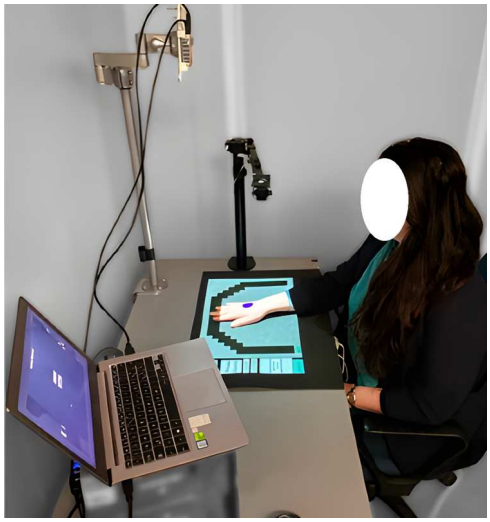


Figura 7. *Painting Discovery* con Leap Motion, modificada de [8].

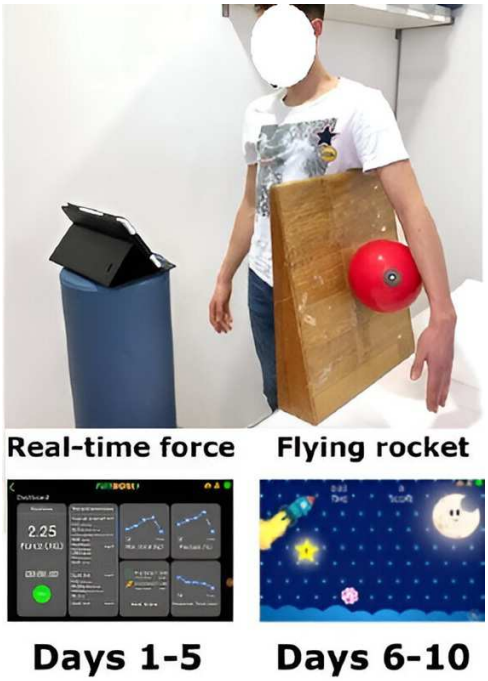


Figura 8. Uso del PlayBall, modificada de [5].

Otro dispositivo, desarrollado por Sousa et al. [9], se denomina Sleeve AR, mostrado en la Figura 9, para la rehabilitación de hombro después de un accidente cerebrovascular con un entorno virtual que involucra movimientos como abducción-aducción, elevación-depresión, flexión-extensión de hombro y flexión-extensión de codo-muñeca. Utiliza el OptiTrack Motion Capture que es un sistema de captura de movimiento que emplea cámaras de alta precisión para rastrear los movimientos del cuerpo en 3D. La retroalimentación en el juego es visual y auditiva, y el movimiento del paciente se evalúa comparándolo con un patrón pregrabado por el fisioterapeuta [9].



Figura 9. Juego serio Sleeve AR, modificada de [9].

Discusión

La rehabilitación con juegos serios con RV y RA ha tomado gran relevancia por su capacidad de ofrecer una experiencia inmersiva y motivadora que contrarresta las limitaciones actuales de la rehabilitación convencional, pues ofrece un entorno multisensorial que hace más atractivo el tratamiento y aumenta la adherencia del paciente. Además, ofrece una evaluación objetiva del progreso gracias al almacenamiento de datos cuantitativos de los movimientos realizados por el paciente durante su sesión de rehabilitación.

Uno de los aspectos más destacados de la rehabilitación virtual es su naturaleza multimodal, ya que permite integrar diferentes dispositivos para interactuar con el entorno virtual de rehabilitación.

Por otra parte, los estudios donde se utilizó cadena cinemática de retroalimentación cerrada, en la que el paciente recibía retroalimentación háptica, como en el caso de los estudios de Moreira et al. [3], Longo et al. [7] y Rizzato et al. [5], demuestran que es particularmente beneficiosa para el proceso de rehabilitación porque proporciona una mejor propiocepción y ubicación en el espacio virtual para el paciente, contribuyendo significativamente a la precisión y eficacia de los movimientos rehabilitadores. Además, la retroalimentación multisensorial mejora la percepción del paciente sobre sus propios movimientos, lo que es crucial para la corrección y mejora de la motricidad.

En cuanto a los dispositivos, el Kinect destaca por su capacidad de capturar movimientos sin necesidad de marcadores físicos o dispositivos adheridos al cuerpo, lo que facilita su implementación en diversos entornos, es ideal para capturar movimientos complejos y grandes ROM. Además, se puede integrar con plataformas como Unity3D y Unreal Enginee, lo que facilita la creación de juegos para rehabilitación. Presenta desventajas como solo proporcionar retroalimentación visual y auditiva, su precisión puede comprometerse por una iluminación deficiente y ha sido descontinuado y reemplazado por otros dispositivos.

Por otro lado, los HDMs ofrecen una inmersión total en entornos virtuales, lo que puede aumentar significativamente la motivación y el compromiso del paciente con la terapia, también proporciona retroalimentación háptica mediante la vibración de sus joysticks lo que mejora la propiocepción y corrección de los movimientos. Sin embargo, suelen ser más costosos en comparación con otros dispositivos de rehabilitación, el uso prolongado del dispositivo puede causar incomodidad, pueden presentarse efectos secundarios como mareos, necesitan de una computadora y su uso presenta limitaciones para usuarios con alta graduación en lentes.

En comparación, el Leap Motion es una opción precisa para el seguimiento de movimientos de manos y muñecas, no obstante, su capacidad de captura se limita a estos segmentos del cuerpo, por lo que no permite capturar y evaluar movimientos directos del hombro, al requerir un rango de movimiento más amplio.

El PlayBall es particularmente útil para medir la fuerza isométrica y la movilidad del hombro, su retroalimentación táctil en tiempo real es crucial para mejorar la propiocepción del paciente. Sin embargo, su enfoque en ejercicios isométricos y de fuerza puede limitar su aplicación en programas de rehabilitación que necesiten un enfoque más integral del movimiento.

Finalmente, el OptiTrack Motion Capture sobresale por su capacidad para capturar movimientos tridimensionales con alta precisión. Sin embargo, su costo varía entre $360,000 y $1,800,000 MXN dependiendo de su configuración específica y necesita un entorno controlado para su operación, lo que limita su accesibilidad.

Conclusión

La rehabilitación virtual ha emergido como una evolución significativa en el ámbito terapéutico, ofreciendo una plataforma multimodal que se adapta a las diversas necesidades de los pacientes mediante la integración

© 2025 - Sociedad Mexicana de Inteligencia Artificial ISSN 2007-0691

de distintos dispositivos tecnológicos. En este contexto, la RV y la RA juegan un papel crucial, ya que contrarrestan las limitaciones de la rehabilitación tradicional y fomentan la adherencia al tratamiento.

Un componente fundamental de esta rehabilitación es el uso de entornos de juego de RV, pues al transformar las sesiones de terapia en experiencias lúdicas, se reduce la monotonía de la rehabilitación convencional, también permite a los pacientes concentrarse en objetivos alcanzables a corto plazo, facilitando la percepción de sus propias mejoras. Además, su retroalimentación en tiempo real enriquece el proceso de aprendizaje y promueve una práctica más efectiva en un entorno seguro.

Por otra parte, la multimodalidad que ofrece la RV y la RA es de suma importancia al momento de personalizar la rehabilitación, pues el poder usar diferentes dispositivos facilita la adaptación de los programas de rehabilitación a las necesidades específicas de cada paciente, patología e incluso fisioterapeuta. Sin embargo, para elegir un dispositivo es esencial considerar las necesidades individuales de cada paciente, las características del entorno de rehabilitación y los recursos disponibles, así como sus ventajas y desventajas.

Se ha demostrado que al estimular múltiples sentidos se optimiza la experiencia del paciente y los resultados terapéuticos de la rehabilitación, especialmente al implementar una cadena cinemática de retroalimentación cerrada, que combina estímulos visuales, auditivos y hápticos. Esta combinación mejora la propiocepción y la ubicación espacial del paciente, contribuyendo a mejorar los resultados terapéuticos.

La implementación de juegos serios RV y RA garantiza un tratamiento efectivo y motivador, lo que contribuye a la recuperación del paciente. Además, como se ha señalado a lo largo del artículo, estos juegos serios ofrecen ventajas notables en comparación con la rehabilitación tradicional. Por lo que se considera, es recomendable continuar con la investigación y desarrollo de estos juegos, especialmente si cumplen con criterios de multimodalidad, retroalimentación multisensorial y la integración de cadenas cinemáticas de retroalimentación cerrada.✷

REFERENCIAS

1. Viglialoro, R. M., Condino, S., Turini, G., Carbone, M., Ferrari, V. y Gesi, M. (2019). Review of the augmented reality systems for shoulder rehabilitation. *Information*, 10(5).

2. Belotti, N., Bonfanti, S., Locatelli, A., Rota, L., Ghidotti, A. y Vitali, A. (2022). A Tele-Rehabilitation Platform for Shoulder Motor Function Recovery Using Serious Games and an Azure Kinect Device. *Studies in Health Technology and Informatics*, 293, 145–152.

3. Moreira, M., Duque, D. y Carvalho, V. (2023). Moreira et al., 3rd Symposium of Applied Science for Young Researchers. *3rd Symposium of Applied Science for Young Researchers*, 83–87.

4. Condino, S., Turini, G., Viglialoro, R., Gesi, M. y Ferrari, V. (2019). Wearable Augmented Reality Application for Shoulder Rehabilitation. *Electronics*, 8(10).

5. Rizzato, A., Pizzichemi, M., Gobbi, E., Gerardi, A., Fortin, C., Copcia, A., Paoli, A. y Marcolin, G. (2023). Effectiveness and therapeutic compliance of digital therapy in shoulder rehabilitation: a randomized controlled trial. *Journal of NeuroEngineering and Rehabilitation*, 20(1).

6. Koutsiana, E., Ladakis, I., Fotopoulos, D., Chytas, A., Kilintzis, V. y Chouvarda, I. (2020). Serious gaming technology in upper extremity rehabilitation: Scoping review. *JMIR Serious Games*, 8(4).

7. Longo, U. G., Carnevale, A., Andreoli, F., Mannocchi, I., Bravi, M., Sassi, M. S. H., Santacaterina, F., Carli, M., Schena, E. y Papalia, R. (2023). Immersive virtual reality for shoulder rehabilitation: evaluation of a physical therapy program executed with Oculus Quest 2. *BMC Musculoskeletal Disorders*, 24(1), 859.

8. Viglialoro, R. M., Turini, G., Carbone, M., Condino, S., Mamone, V., Coluccia, N., Dell'Agli, S., Morucci, G., Ryskalin, L., Ferrari, V. y Gesi, M. (2023). A Projected AR Serious Game for Shoulder Rehabilitation Using Hand-Finger Tracking and Performance Metrics: A Preliminary Study on Healthy Subjects. *Electronics*, 12(11).

9. Sousa, M., Vieira, J., Medeiros, D., Arsenio, A. y Jorge, J. (2016). SleeveAR: Augmented Reality for Rehabilitation using Realtime Feedback. En *Proceedings of the 21st International Conference on Intelligent User Interfaces*, 175–185.

SOBRE LOS AUTORES



Citlali Irais Cano del Razo es egresada de Bioingeniería Médica por la Universidad Autónoma del Estado de México. Realizó su servicio social en el laboratorio de robótica de la Facultad de Ingeniería de la misma universidad, donde desarrolló un juego serio de realidad virtual para la rehabilitación de hombro doloroso. Sus intereses incluyen la rehabilitación, la robótica médica y los entornos virtuales.

© 2025 - Sociedad Mexicana de Inteligencia Artificial ISSN 2007-0691



Juan Manuel Jacinto Villegas obtuvo el título de Ingeniero en Electrónica (por promedio) y la Maestría en Ingeniería Electrónica (con mención honorífica) por el Instituto Tecnológico de Orizaba en 2010 y 2012, respectivamente. En 2017, obtuvo el grado de Doctor (con mención honorífica) en Tecnologías Digitales Emergentes, en el área de investigación de Robótica Perceptual, por la Scuola Superiore Sant'Anna, en Pisa, Italia. Desde 2018, se desempeña como Profesor-Investigador del programa "Investigadoras e Investigadores por México" del SECIHTI, adscrito a la Facultad de Ingeniería de la Universidad Autónoma del Estado de México (UAEMéx), en Toluca, México. Es miembro del Sistema Nacional de Investigadoras e Investigadores (SNII) nivel 1. Sus áreas de interés incluyen la robótica, control háptico, interacción humano-robot, exoesqueletos, ambientes virtuales, diseño mecánico y electrónico, mecatrónica y automatización.



Monserrat Ríos Hernández realizó sus estudios de Maestría en la Facultad de Ingeniería de la Universidad Autónoma del Estado de México con una estancia en el LIRMM (Laboratoire d'Informatique, de Robotique et de Microélectronique de Montpellier), Francia, enfocándose en el desarrollo de simuladores médicos. Actualmente, es profesor de la Facultad de Medicina y estudiante doctoral en la Facultad de Ingeniería. Sus intereses incluyen la aplicación de la realidad virtual en el área médica y control háptico.



Adriana H. Vilchis González obtuvo el grado de doctor en Imagen, Visión y Robótica en el Instituto Nacional Politécnico de Grenoble, Francia. Actualmente es profesor investigador en las Facultades de Ingeniería y de Medicina de la Universidad Autónoma del Estado de México. Sus intereses científicos incluyen Robótica Médica, Realidad Virtual, Robots Suaves y Sistemas Mecatrónicos.



La rehabilitación virtual. Imagen tomada de: `https://www.gob.mx/conadis/articulos/la-rehabilitacion-virtual?idiom=es`.

© 2025 - Sociedad Mexicana de Inteligencia Artificial ISSN 2007-0691

ARTÍCULO ACEPTADO

¿Eres tú, Juan?: La amenaza de los *deepfakes* y la investigación que busca mitigarlos

Edgar Macías García, David Gómez Gutiérrez, Paulo López Meyer, Rodrigo Aldana-López y Julio Zamora-Esquivel

El término *deepfake* surge de la combinación de "aprendizaje profundo"(del inglés *deep learning*) y "falsificaciones"(del inglés *fakes*)

Introducción

La Inteligencia Artificial (IA) ha transformado profundamente nuestra vida cotidiana, dando lugar a herramientas con aplicaciones en campos tan diversos como la medicina, el entretenimiento, la educación y la seguridad. Sin embargo, estos avances también han introducido desafíos éticos, sociales y legales que debemos abordar. Por ejemplo, el hacer *deepfakes* ha ganado notoriedad por su capacidad de generar imágenes, video y audios que pudieran potencialmente alterar nuestra percepción de la realidad.

El término *deepfake* es una palabra compuesta derivada de los términos en inglés *deep learning* (o aprendizaje profundo) y *fakes* (o falsificación). Se refiere a la creación de contenido audiovisual manipulado mediante algoritmos de IA, logrando un nivel de realismo que, en muchos casos, lo hace casi indistinguible del contenido auténtico [3].

Inicialmente, el contenido deepfake se estudió puramente en investigación académica; sin embargo, hoy están presentes en videos virales, en campañas publicitarias y, lamentablemente, en fraudes y campañas de desinformación. Este artículo tiene como objetivo explorar qué es realmente un deepfake, cómo se diferencia de las manipulaciones tradicionales, cómo funcionan las tecnologías que lo hacen posible, cómo podemos evitar ser engañados por él, y qué pueden hacer los desarrolladores de estas herramientas de IA Generativa para asegurar su uso responsable.

¿Qué es un deepfake?

Un *deepfake* es un contenido audiovisual generado por IA, diseñado para introducir información falsa que resulta imperceptible al ojo humano; desde usar el rostro de una persona real para colocarle palabras que nunca dijo, o incluso generar una secuencia completa de video de cualquier cosa. Esta tecnología emplea los avances de las redes neuronales generativas para generar contenido que resulta difícil de identificar como real o falso (Figura 1).



Figura 1. Ejemplo de una superposición de rostro. El rostro de la imagen original tomada de una película del actor Arnold Schwarzenegger es reemplazada por una imagen falsa con aspecto del actor Sylvester Stallone [5].

Antes de la llegada de la IA, falsificar imágenes y videos requería habilidades técnicas avanzadas y horas de edición manual. Desde los retoques de imágenes en Photoshop hasta el trabajo de especialistas en efectos visuales en películas; estas manipulaciones eran tediosas y requerían de una preparación técnica muy especializada. Aunque los efectos finales podían lucir realistas, solían dejar rastros que eran identificables por expertos.

Hoy, las herramientas para generar un *deepfake* han reducido la barrera técnica para falsificar contenido. Con el software adecuado y unas pocas horas de aprendizaje, cualquier persona puede crear un video falso realista. Esto representa un cambio de paradigma: ya no se necesita un equipo de expertos, sino un algoritmo bien entrenado. En la actualidad, es importante distinguir un *deepfake* de otras manipulaciones más simples. Por ejemplo:

- **Retoques digitales:** Editar una foto para eliminar imperfecciones no es un *deepfake*.
- **Edición convencional de audio o video:** Cortar y pegar frases para alterar un discurso tampoco entra en la categoría de *deepfake*.

© 2025 - Sociedad Mexicana de Inteligencia Artificial ISSN 2007-0691

- **Actuaciones con efectos especiales:** En el cine, se ha recreado digitalmente a actores fallecidos, como *Carrie Fisher* en *Star Wars: Rogue One*, pero esto se hace con técnicas diferentes y generalmente con autorización legal.

Un *deepfake* no es simplemente contenido modificado; es un contenido generado o manipulado con la intención de parecer auténtico.

¿Cómo se hacen los deepfakes?

Para crear un *deepfake* se utilizan técnicas avanzadas de aprendizaje profundo, siendo las redes generativas antagónicas (del inglés *Generative Adversarial Networks* o GANs) la base fundamental de muchos sistemas. Estos modelos de redes neuronales emplean dos estructuras que compiten entre sí, donde un modelo es entrenado para evaluar la calidad de un contenido, y otro, para generar contenido que cumpla sus expectativas (Figura 2).

Para explicar cómo funciona una GAN, imagina que un artista está aprendiendo a pintar retratos realistas, pero no tiene experiencia. A su lado, hay un crítico de arte muy exigente que evalúa sus pinturas. Cada vez que el artista entrega un retrato, el crítico lo analiza y señala todos los defectos, como proporciones incorrectas o falta de realismo. Con cada iteración, el artista mejora sus habilidades basándose en los comentarios del crítico. Al principio, los retratos son toscos, pero con el tiempo, el artista logra obras maestras que incluso el crítico no puede distinguir de un retrato real. Bajo este enfoque:

- **El artista:** Es el generador, una red neuronal que intenta crear imágenes, audios o videos falsos a partir de un estímulo de entrada (como una descripción de cómo debe generar la pintura).
- **El crítico:** Es el discriminador, otra red neuronal que evalúa si el contenido generado es falso o real. Para ello utiliza como referencia su conocimiento después de evaluar cientos de obras reales.
- **El aprendizaje mutuo:** El generador mejora constantemente, mientras que el discriminador se vuelve más hábil para detectar falsificaciones. Este juego entre ambos resulta en contenido cada vez más realista [3].

Tanto el generador como el discriminador son modelos de aprendizaje profundo, compuestos por múltiples capas de neuronas artificiales que permiten procesar datos complejos como imágenes o videos. Los modelos más avanzados suelen tener una estructura de tipo *codificador-decodificador*, donde:

- El **codificador** extrae características relevantes de los datos de entrada, en forma de una distribución de probabilidad, como los rasgos faciales clave en una imagen.
- El **decodificador** utiliza esas características para reconstruir una nueva imagen o video con las modificaciones deseadas.

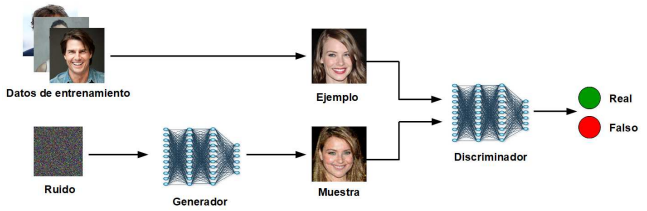


Figura 2. Funcionamiento general de una red neuronal generativa. Se compone de un generador que crea imágenes a partir de entradas, y un discriminador que separa las muestras reales de las falsas.

Esta arquitectura es particularmente eficaz para tareas como el intercambio de rostros, ya que permite capturar detalles faciales precisos mientras mantiene la coherencia en el contexto general del video. Aunque las GANs son la base para la generación de contenido sintético, las tecnologías actuales van más allá, integrando herramientas adicionales más sofisticadas. Estas técnicas han democratizado la creación de contenido *deepfake*, haciendo que incluso personas sin experiencia técnica puedan generar contenido manipulado utilizando herramientas como DeepFaceLab [5].

Dilemas éticos y desinformación

El *deepfake* representa no solo un avance tecnológico impresionante, sino también una fuente de preocupaciones éticas significativas. Al manipular imágenes, audios o videos para crear contenido falso pero convincente, estas herramientas han planteado interrogantes sobre los límites entre la creatividad, la privacidad y el engaño intencionado.

En México, el impacto del contenido *deepfake* se ha hecho evidente en distintos ámbitos. En el ámbito político, durante las elecciones presidenciales del 2024, circularon en redes sociales *deepfakes* en forma de videos de la entonces candidata Claudia Sheinbaum; en éstos supuestamente ella invitaba a la población a participar en un esquema de inversión de alto rendimiento[1]. En el fraude telefónico, la clonación de voces mediante IA ha facilitado casos de extorsión y suplantación de identidad, permitiendo a los estafadores imitar la voz de familiares para engañar a sus víctimas[2]. En publicidad, algunas

[1] https://es.wired.com/articulos/claudia-sheinbaum-victima-de-deepfake-su-imagen-fue-utilizada-en-un-video-fraudulento-generado-por-ia

[2] https://www.elfinanciero.com.mx/mis-finanzas/2024/11/03/como-funcionan-las-extorsiones-con-ia-y-la-voz-clonada-de-tus-familiares-en-whatsapp/

© 2025 - Sociedad Mexicana de Inteligencia Artificial ISSN 2007-0691

empresas han explorado el uso de *deepfakes* para campañas innovadoras, aunque su implementación sigue siendo limitada. Un ejemplo notable ocurrió en 2021, cuando Soriana utilizó un *deepfake* del comediante Cantinflas en una de sus promociones[3].

Una investigación interesante es el análisis por fotopletismografía desarrollado por la empresa Intel, en donde se analizan las variaciones temporales que deja el flujo sanguíneo sobre los píxeles del rostro.

A medida que los *deepfakes* se vuelven más comunes y accesibles, surge el riesgo de una crisis de autenticidad, donde las personas desconfían incluso del contenido real. Esto podría erosionar la confianza en instituciones clave, como el periodismo, el sistema judicial y los gobiernos, dificultando la toma de decisiones informadas y fomentando la desinformación masiva. El desarrollo de los *deepfakes* plantea el dilema ético central sobre cómo podemos aprovechar esta tecnología para aplicaciones creativas y legítimas sin exponer a las personas y sociedades a riesgos innecesarios. La respuesta radica en una combinación de conciencia pública, regulación adecuada y avances tecnológicos que permitan detectar y mitigar los usos malintencionados.

Legislaciones y regulación

A medida que los *deepfakes* se han convertido en una preocupación global, varios gobiernos han adoptado medidas regulatorias para mitigar sus riesgos. Por ejemplo, en México, el diputado Miguel Torruco presentó una propuesta de reforma al Código Penal Federal en 2023 para sancionar con hasta ocho años de prisión la creación y distribución de contenido perjudicial, incluidos documentos falsificados y material íntimo no consentido[4]. Además, se plantea la ampliación de la Ley Olimpia para incluir contenido generado mediante IA.

Métodos de detección de deepfakes

Los métodos forenses son una de las estrategias más tradicionales para identificar contenidos manipulados. Estos análisis se enfocan en detectar detalles específicos en las imágenes; como inconsistencias en la iluminación, errores en la sincronización labial o patrones anómalos en el parpadeo. Estas técnicas buscan rastrear los artefactos visuales que los algoritmos de generación de *deepfakes* suelen dejar como huellas digitales.

Un caso particularmente interesante es el de Elisabeth Bik, una científica que se especializa en realizar análisis forense de imágenes alteradas dentro de artículos académicos. Bik ha logrado identificar con éxito falsificaciones realizadas principalmente de manera manual por personas, como la duplicación de imágenes o la manipulación con herramientas digitales tradicionales. Sin embargo, su enfoque, que depende de un ojo entrenado y de un análisis minucioso, enfrenta limitaciones importantes frente a la evolución tecnológica. A medida que los algoritmos de generación de *deepfakes* se vuelven más avanzados, los artefactos visuales que antes servían como huellas digitales son eliminados, haciendo que incluso personas altamente capacitadas, como Bik, tengan dificultades para detectar estas manipulaciones. Esto subraya la creciente necesidad de herramientas automatizadas y basadas en IA para complementar el trabajo humano en la detección de *deepfakes*[5].

Para abordar estas limitaciones, la comunidad científica ha desarrollado herramientas basadas en IA que permiten analizar patrones sutiles en la textura de la piel, las expresiones faciales y otros rasgos difíciles de falsificar. Una iniciativa destacada en este ámbito es el *Deepfake Detection Challenge*, organizado por META AI (antes Facebook), que ofrece un premio de 500,000 USD al ganador para estimular la innovación en técnicas de detección[6]. Los participantes utilizan enfoques que incluyen redes neuronales profundas y métodos basados en aprendizaje por transferencia para identificar manipulaciones con alta precisión.

Una investigación interesante en este ámbito es el análisis por fotopletismografía desarrollado por la empresa Intel [1], en donde se analizan las variaciones temporales que deja el flujo sanguíneo sobre los píxeles del rostro. La investigación no solo permite identificar si un video es potencialmente falso, sino también el tipo de algoritmo empleado para generar dicho contenido a través del análisis de sus huellas temporales.

Este algoritmo emplea un conjunto de detectores que tienen como propósito detectar la presencia de personas en las secuencias de video, identificar regiones específicas de su rostro, y finalmente analizar las variaciones temporales de los píxeles en dichas regiones durante la secuencia de video (Figura 3).

[3] https://disruptivooficial.com/blog/cantinflas-soriana

[4] https://comunicacionsocial.diputados.gob.mx/index.php/boletines/-busca-iniciativa-sancionar-uso-delictivo-de-la-tecnologia-denominada-inteligencia-artificial

[5] https://scienceintegritydigest.com/

[6] https://www.kaggle.com/competitions/deepfake-detection-challenge

© 2025 - Sociedad Mexicana de Inteligencia Artificial ISSN 2007-0691

- En una primera etapa, el algoritmo emplea un detector de rostros seguido de un localizador de puntos de referencia (por ejemplo, la posición exacta de la nariz, la boca y las mejillas).
- Los puntos obtenidos son usados para recortar una región partícular del rostro, que es almacenada en una secuencia temporal.
- Una vez que la secuencia temporal es suficientemente grande, se genera una nueva imagen usando fotopletismografía, la cual, es analizada por un modelo basado en IA que la clasifica como real o verdadera en función de los patrones encontrados.

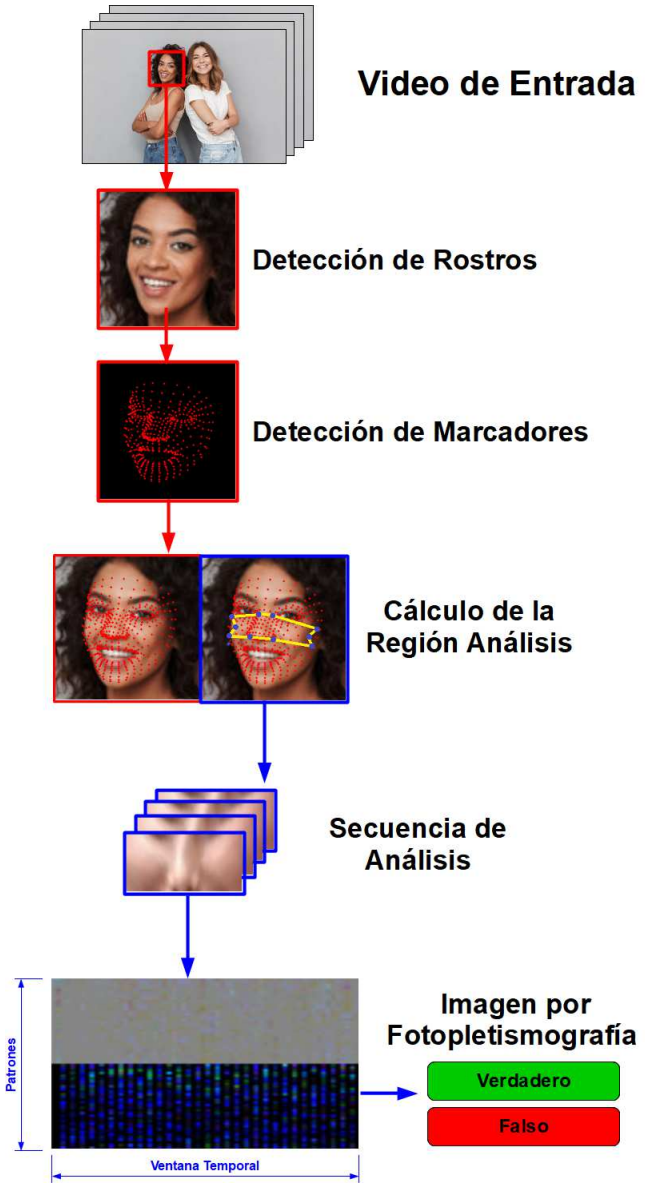


Figura 3. Proceso de generación de una imagen de fotopletismografía a partir de una secuencia de video.

En una cara real, los píxeles variarán de manera uniforme en función del tipo de gesticulaciones o emociones empleadas por la persona (modificando ligeramente la tonalidad de los píxeles), o bien, debido a cambios en la iluminación o la pose del rostro.

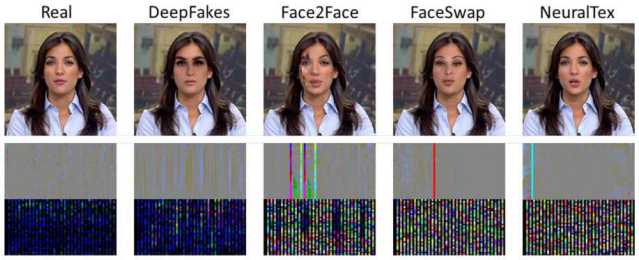


Figura 4. El contenido *deepfake* suele dejar huellas digitales que son detectables usando algoritmos específicos, como el análisis por fotopletismografía [1].

Por otro lado, en un rostro falso los píxeles pueden presentar cambios abruptos de un instante a otro (debido, por ejemplo, a una superposición errónea). Estos errores de superposición dejan huellas que introducen ruido a las imágenes procesadas por fotopletismografía (Figura 4). Si bien este algoritmo presenta resultados prometedores, la correcta detección del rastro que deja el flujo sanguíneo requiere que los rostros detectados en los videos tengan una resolución suficiente para visualizar las variaciones de intensidad, lo que se dificulta a medida que la resolución disminuye. Debido a esto, en la actualidad se siguen haciendo investigaciones para incluir otras características diferenciables, como el análisis de audio (y su correlación con el movimiento de los labios) y la trayectoria de las gesticulaciones (movimiento de ojos, boca etc.).

Estrategias de mitigación

La detección de *deepfakes* enfrenta limitaciones cada vez mayores a medida que la tecnología se vuelve más compleja y sofisticada. Los algoritmos avanzados han mejorado significativamente en la eliminación de artefactos visuales y la creación de contenido falso más realista, lo que dificulta su identificación, incluso con herramientas especializadas. Además, las técnicas de detección podrían ser usadas para mejorar la tecnología de generación. Por esta razón, resulta fundamental explorar enfoques complementarios que no solo se centren en la detección, sino también en la mitigación y prevención de los riesgos asociados. Una de las estrategias más prometedoras en este sentido es el *watermarking*, que consiste en incrustar marcas de agua digitales o metadatos en el contenido generado. Estas marcas, imperceptibles para el ojo humano, permiten autenticar el origen del contenido y verificar su legitimidad, ofreciendo una capa adicional de seguridad frente a las manipulaciones [2].

© 2025 - Sociedad Mexicana de Inteligencia Artificial ISSN 2007-0691

En este sentido, empresas tecnológicas como Google y OpenAI están trabajando en el desarrollo de herramientas innovadoras para etiquetar y rastrear creaciones sintéticas. Un ejemplo de ello es SynthID, una tecnología desarrollada por DeepMind que integra marcas digitales directamente en las imágenes generadas por IA para facilitar su identificación, incluso después de modificaciones[7].

Estas estrategias de mitigación buscan no solo reducir el impacto del contenido *deepfake*, sino también generar confianza en las tecnologías de IA, asegurando un uso responsable y ético de estas herramientas. Es importante destacar que la gran mayoría de los métodos para detectar un *deepfake* son también basados en IA, por tanto son susceptibles a un sesgo, es decir su desempeño no es el mismo para cada etnia pues son entrenados usando bases de datos que no necesariamente están balanceadas. Por ejemplo, muchos grupos no están bien representados en el conjunto de datos, como sucede en la base de datos *Labeled Faces in the Wild* (LFW) [4].Dicho repositorio, a pesar de contener más de 13000 rostros, posee un porcentaje bajo de niños, ningún bebé, muy pocas personas mayores de 80 años y una proporción relativamente pequeña de mujeres. Además, muchas etnias tienen una representación mínima o nula.

La demografía afecta fuertemente el rendimiento del reconocimiento. Uno de los impactos más fuertes en FaceNet [6] se observa en las etnias. Para el modelo FaceNet investigado, los rostros asiáticos y tez oscura generan tasas de reconocimiento significativamente más bajas que los rostros blancos. Además, los rostros jóvenes tienen un desempeño significativamente más débil que los rostros de mediana edad. En cuanto al género, observamos que los rostros masculinos tienen mejores resultados que los femeninos.

¿Cómo evitar caer en un deepfake?

En un mundo donde el contenido *deepfake* es cada vez más convincente, las personas tienen un papel crucial en mitigar la desinformación. Algunas recomendaciones prácticas incluyen:

- **Verificar fuentes:** Antes de compartir un video o noticia, consulta fuentes confiables y verifica si la información ha sido reportada por varios medios independientes. Busca señales de credibilidad, como el contexto, la fecha y los autores del contenido.
- **Analizar detalles sospechosos:** Presta atención a inconsistencias como movimientos faciales rígidos o poco naturales, iluminación desigual, sombras mal alineadas o sonidos de-sincronizados.
- **Aplicar pensamiento crítico:** Pregúntate quién podría beneficiarse de que creas en el contenido que estás viendo. Contrasta la intención detrás del mensaje y si parece diseñado para manipular tus emociones o influir en tus decisiones.
- **Educarse:** Infórmate sobre cómo funcionan los *deepfakes* y sus limitaciones. Comprender las tecnologías que los producen y sus vulnerabilidades puede ayudarte a detectarlos con mayor facilidad y a no caer en desinformación.

Conclusiones, Futuro y tendencias

El contenido *deepfake* es un ejemplo del impacto transformador de la IA, con aplicaciones tanto creativas como problemáticas. A corto plazo, es previsible que la calidad y accesibilidad de estas herramientas sigan mejorando, facilitando tanto usos legítimos como malintencionados. Este avance plantea desafíos legales, éticos y sociales que deben ser abordados [7]. A largo plazo, la respuesta de los gobiernos e instituciones será fundamental para generar una regulación apropiada de su uso. La implementación de normas que promuevan la transparencia y el desarrollo de tecnologías para detectar y mitigar los *deepfakes* son pasos en la dirección correcta. Sin embargo, estas respuestas deben equilibrarse cuidadosamente para no limitar la innovación.

En última instancia, el futuro del contenido *deepfake* dependerá de nuestra capacidad como sociedad para comprender y gestionar sus riesgos, mientras aprovechamos su potencial creativo de manera responsable. La colaboración entre desarrolladores, legisladores y el público será crucial para construir un entorno donde estas tecnologías puedan coexistir con los valores éticos y sociales fundamentales.✷

REFERENCIAS

1. C¸ ift¸ci, U. A., Demir, ˙I. y Yin, L. (2024). Deepfake source detection in a heart beat. *The Visual Computer*, 40(4):2733–2750.
2. Cox, I. et al. (2021). A review of digital watermarking for multimedia content authentication. *Journal of Digital Media*.
3. Goodfellow, I. et al. (2014). Generative adversarial nets. *Advances in neural information processing systems*.
4. Liu, Z., Luo, P., Wang, X., and Tang, X. (2014). Deep learning face attributes in the wild. En *2015 IEEE International Conference on Computer Vision (ICCV)*, 3730–3738.
5. Perov, I., Gao, D., Chervoniy, N., Liu, K., Marangonda, S., Um´e, C., Facenheim, C. S., RP, L., Jiang, J., Zhang, S., et al. (2020). Deepfacelab: Integrated, flexible and extensible faceswapping framework. *arXiv preprint* arXiv:2005.05535.
6. Schroff, F., Kalenichenko, D., and Philbin, J. (2015). Facenet: A unified embedding for face recognition and clustering. 815–823.
7. Smith, J., y Doe, J. (2023). Deepfakes: Perspectives, risks, and regulation. *Artificial Intelligence Journal*, 45:123–145.

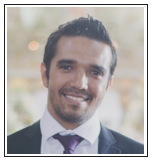
[7] https://deepmind.google/discover/blog/identifying-ai-generated-images-with-synthid/

© 2025 - Sociedad Mexicana de Inteligencia Artificial ISSN 2007-0691

SOBRE LOS AUTORES



Dr. Edgar Macías García es un científico investigador que forma parte del grupo de Colaboración Humano-Máquina en Intel Labs. Estudió la carrera de Ingeniería Química en la Universidad de Guadalajara con especialización en Control Automático, posteriormente realizó estudios de maestría y doctorado en el Centro de Investigación y Estudios Avanzados del Instituto Politécnico Nacional. Actualmente, su investigación se centra en el desarrollo de nuevos modelos de redes neuronales artificiales con aplicaciones como la detección de *deepfakes* y la percepción robótica.



Dr. David Gómez Gutiérrez obtuvo el grado de doctor en Ciencias en Ingeniería Eléctrica en Cinvestav, Unidad Guadalajara en 2013. Es investigador en el Intelligent Systems Research Lab, Intel Labs, miembro del Tecnológico Nacional de México, Instituto Tecnológico José Mario Molina Pasquel y Henríquez y miembro SNII Nivel II. Sus intereses científicos incluyen el desarrollo de algoritmos de inteligencia artificial en control y robótica.



Dr. Paulo López Meyer se unió a Intel en el 2013 como Investigador Científico. Previamente, obtuvo su doctorado en Ingeniería Eléctrica y de Computación en Clarkson University, y trabajó como investigador postdoctoral en la Universidad de Alabama, en EEUU. Sus áreas de especialización incluyen procesamiento de señales, análisis de datos y aprendizaje automático. Actualmente, desarrolla algoritmos inteligentes aplicados a la validación eléctrica en microprocesadores dentro de Intel. Adicionalmente, he formado parte en diferentes ocasiones del Sistema Nacional de Investigadores de la SECIHTI (antes CONAHCyT), y participa activamente como revisor en revistas y conferencias de prestigio internacionales.



Dr. Rodrigo Aldana-López es un científico investigador en Inteligencia Artificial y forma parte del grupo de Colaboración Humano-Máquina en Intel Labs. Obtuvo el grado de Doctor en Ingeniería de Sistemas e Informática por la Universidad de Zaragoza, España, en 2024, y el grado de Ingeniero en Electrónica por el Instituto Tecnológico y de Estudios Superiores de Occidente, Tlaquepaque, México. Sus intereses de investigación incluyen robótica con restricciones en tiempo real, aprendizaje automático, inteligencia artificial y diseño de protocolos de comunicación para sistemas distribuidos.



Dr. Julio Zamora-Esquivel es "Principal Engineer" y científico investigador en Inteligencia Artificial en Intel Labs, director del grupo "Human-Robot Collaboration", miembros del laboratorio "Intelligent System Research". Obtuvo una maestría en Ciencias de la Computación y un doctorado en Ingeniería en CINVESTAV. El Dr. Zamora tuvo una posición postdoctoral en KAIST, Corea, es miembro del Sistema Nacional de Investigadores, de la Asociación Mexicana de Visión por Computadora, Computación Neural y Robótica, y miembro Senior del IEEE. Tiene más de 100 patentes en proceso y más de 30 publicaciones.

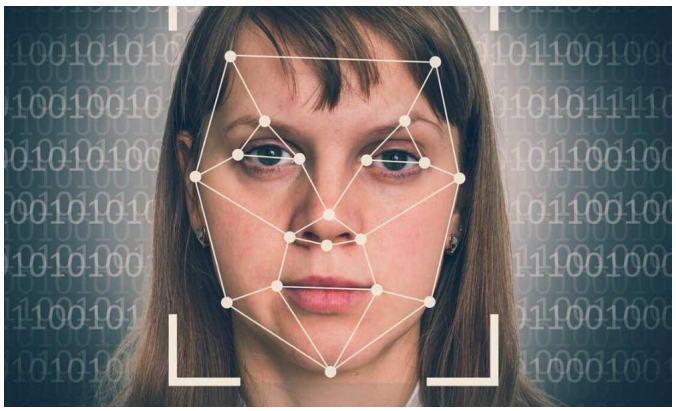


Imagen tomada de:
https://www.socialmediasafety.org/advocacy/deepfake-technology/.

© 2025 - Sociedad Mexicana de Inteligencia Artificial ISSN 2007-0691

ARTÍCULO ACEPTADO

El Efecto Mandela Digital: Cómo los Ataques Adversarios Engañan a las Redes Neuronales

Monserrat Vázquez Hernández, Luis Alberto Morales Rosales e Ignacio Algredo Badillo

Los ataques adversarios y la IA responsable están conectados por un mismo desafío: la seguridad de los modelos de Inteligencia Artificial.

El Efecto Mandela: Recuerdos Falsos que Alteran Nuestra Realidad

Alguna vez te ha pasado que recuerdas un evento de una forma, pero cuando lo verificas, resulta que nunca ocurrió así. O bien, tal vez encuentras personas que insisten en que un hecho fue distinto a como realmente sucedió. Este fenómeno no solo ocurre a nivel individual, sino que también puede afectar a grupos enteros de personas. A este fenómeno se le conoce como el *efecto Mandela* [1].

El efecto Mandera se origina de un error de memoria que fue compartido por muchas personas, el cual consistió en la creencia de que Nelson Mandela había fallecido en prisión, cuando en realidad eso nunca ocurrió. Mandela fue encarcelado en 1963 por su lucha por los derechos humanos y pasó 27 años tras las rejas. Fue liberado en 1990 y se convirtió en presidente de Sudáfrica y gobernó desde 1994 hasta 1999. Su muerte ocurrió el 5 de diciembre de 2013 debido a una neumonía. Cuando se anunció su fallecimiento en los medios de comunicación, muchas personas aseguraban recordar haber visto su funeral décadas antes.

El efecto Mandela es el nombre con el que se conoce popularmente al fenómeno colectivo de recuerdos falsos. Dicho término fue acuñado por la bloguera Fiona Broome, después de estudiar sobre la creencia popular de la muerte de Mandela [2]. Un caso similar muy sonado ocurrió con la canción "*We Are the Champions*" de la banda de rock *Queen*. Muchos creen que la canción finaliza con la icónica frase "*We are the champions... of the world*!", pero la versión original en realidad termina en "*'Cause we are the champions*". Freddie Mercury, vocalista de la banda, solía cantar la versión extendida en vivo en consideración a sus fans para que ellos no se quedaran cantando solos, lo que llevó a muchas personas a recordar erróneamente que ese era el final original de la canción. ¿Por qué se originan estos errores en la memoria? En algunos casos, se producen cuando la información es alterada levemente y se propaga hasta que la versión errónea reemplaza al recuerdo original. Esta capacidad de modificar la percepción no solo ocurre en los seres humanos, sino que también puede ser generada en modelos de inteligencia artificial.

Efecto Mandela en la Inteligencia Artificial: Los Ataques Adversarios

El Aprendizaje Profundo (*Deep Learning*, DL) es una técnica de Inteligencia Artificial (IA) que emplea Redes Neuronales Profundas (*Deep Neural Networks*, DNNs) para procesar datos y mejorar la percepción de la realidad. Estas redes permiten reconocer patrones complejos y tomar decisiones basadas en grandes volúmenes de información. Sin embargo, al igual que la memoria humana, estas redes también pueden ser engañadas. Al introducir pequeñas modificaciones en los datos de entrada, es posible hacer que una DNN interprete erróneamente una imagen, un sonido o un texto. Este proceso, conocido como ataques adversarios, altera la percepción de la realidad de los modelos de IA, llevándolos a identificar de manera incorrecta lo que están analizando.

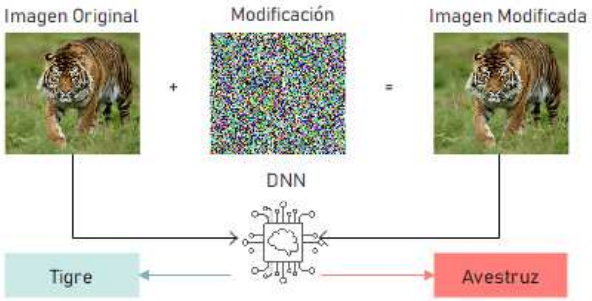


Figura 1. Los Ataques Adversarios incluyen imperceptibles modificaciones en los datos de entrada para ocasionar errores de clasificación.

Supongamos que tomamos una imagen de un tigre y la analizamos utilizando un modelo basado en DNNs. Si el modelo funciona correctamente, debería identificar la imagen del tigre sin problemas. Sin embargo, y a pesar

© 2025 - Sociedad Mexicana de Inteligencia Artificial ISSN 2007-0691

del correcto funcionamiento de la red, podríamos modificar la imagen con un cambio tan sutil indetectable al ojo humano, pero que puede engañar por completo a los modelos de DNNs (ver Figura 1). Como resultado, el modelo podría clasificar erróneamente al tigre como un avestruz o cualquier otra cosa completamente distinta.

El tipo de imagen modificada previamente descrito se conoce como ejemplo adversario. Los ataques adversarios hacen uso de ejemplos adversarios para confundir a los modelos de DNNs, alterando su percepción y haciendo que interpreten la realidad de forma incorrecta.

¿Cuál es la relación entre los Ataques Adversarios y la IA Responsable?

Los ataques adversarios y la IA responsable están conectados por un mismo desafío: la seguridad de sus modelos. Mientras que los ataques adversarios explotan vulnerabilidades para manipular resultados, la IA responsable busca hacerlos más seguros, justos y transparentes.

La IA responsable no solo se preocupa por la equidad y la privacidad, sino también por la resiliencia del sistema ante manipulaciones. Si un modelo de IA es fácilmente engañado por ataques adversarios, su confiabilidad se ve comprometida, lo que puede generar desconfianza en aplicaciones críticas como salud, seguridad y finanzas. Pero, ¿cómo es que pequeños cambios imperceptibles para nosotros pueden engañar a un sistema de IA? Para entenderlo, exploraremos cómo funcionan los ataques adversarios, por qué afectan a las redes neuronales y las estrategias para proteger los modelos de DNNs de errores en la clasificación.

¿Cuál es el origen de los Ataques Adversarios?

El aprendizaje profundo ha revolucionado diversas áreas del conocimiento gracias a su capacidad para mejorar el rendimiento de tareas complejas. Estos potentes algoritmos han superado enfoques previos en campos como Visión por Computadora, Procesamiento de Lenguaje Natural y Minería de Datos, ganando gran popularidad tanto en el ámbito académico como en la industria.

En el núcleo del DL se encuentran las Redes Neuronales Profundas, modelos inspirados en la estructura y funcionamiento del cerebro humano. Estos modelos aprenden patrones a partir de datos de entrenamiento, permitiéndoles reconocer objetos, interpretar lenguaje y tomar decisiones con notable precisión. Un ejemplo de su aplicación es la conducción autónoma, donde los vehículos dependen de modelos de DNNs para analizar su entorno y responder de manera autónoma.

A medida que los modelos de DNNs avanzan, su capacidad de operar sin supervisión humana la vuelve una herramienta indispensable en múltiples sectores, por lo que evaluar su funcionamiento es crucial para asegurar que operen de manera precisa y confiable. En 2013, Szegedy et al. [3] estudiaron el funcionamiento de estos modelos al realizar pequeñas e imperceptibles modificaciones en imágenes que debían ser clasificadas. Descubrieron que, a pesar de que los cambios pasaban inadvertidos para el ojo humano, las DNN eran fácilmente engañadas, asignando etiquetas incorrectas a los datos. En su trabajo, las entradas modificadas fueron denominadas ejemplos adversarios (*adversarial examples*), término que hoy en día se utiliza para describir cualquier modificación intencional diseñada para confundir a las DNN y provocar errores de clasificación.

¿Por qué las Redes Neuronales fallan ante pequeñas modificaciones?

En el estudio "*Adversarial Examples Are Not Bugs, They Are Features*" [4], los investigadores plantean una interesante hipótesis sobre el funcionamiento de las DNN: estos modelos no solo aprenden a identificar patrones evidentes, sino que se basan en características no robustas, es decir, detalles sutiles que son imperceptibles para el ojo humano y que resultan esenciales para el proceso de clasificación. Mientras que los humanos solemos enfocarnos en rasgos generales o características robustas, como la forma o el color, las DNN encuentran en las pequeñas variaciones la información clave para su toma de decisiones. Por ejemplo, la Figura 2 ilustra la percepción de las características evidentes y sutiles por parte de un modelo DNN para la clasificación de la imagen de una rana.

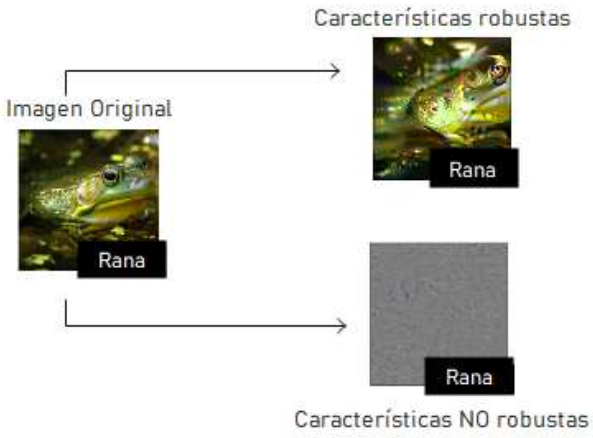


Figura 2. Las Redes Neuronales aprenden características robustas y no robustas. Las características no robustas son imperceptibles por los humanos pero útiles para las Redes Neuronales.

Para ilustrar con mayor claridad la hipótesis propuesta, imaginemos un modelo DNNs diseñado para clasificar imágenes de perros y gatos. Las características robustas serían aquellas que ambos animales comparten,

como orejas, cabeza y patas. Sin embargo, para diferenciar entre ambas especies, el modelo DNN se apoya en características no robustas, como la dirección del pelaje, la forma de los bigotes o los patrones de manchas. Aunque estos detalles pueden parecer insignificantes, son determinantes para el modelo al momento de asignar una categoría.

De acuerdo con los investigadores, la distinción entre características robustas y no robustas permite comprender la razón por la que los modelos de aprendizaje profundo pueden ser vulnerables a ejemplos adversarios. Pequeñas alteraciones en las características no robustas, aunque imperceptibles para el ojo humano, pueden desviar la interpretación de los modelos y provocar clasificaciones erróneas. Para probar esta hipótesis, los autores llevaron a cabo experimentos en redes neuronales profundas y descubrieron que cambios tenues eran suficientes para afectar sus resultados. Este descubrimiento resalta tanto la capacidad como la debilidad de los modelos de DNNs, fomentando la creación de estrategias más robustas y seguras.

¿Cuáles son los riesgos de los Ataques Adversarios?

Desde su descubrimiento en 2013 por Szegedy et al. [3], los ataques adversarios han expuesto una vulnerabilidad clave en los modelos de DL: su susceptibilidad a pequeñas modificaciones en los datos de entrada que pueden llevarlos a clasificar de manera errónea. Estos ataques buscan alterar la percepción del modelo, haciéndolo tomar decisiones incorrectas, lo que representa un desafío significativo para la seguridad y confiabilidad de estos sistemas en diversas aplicaciones.

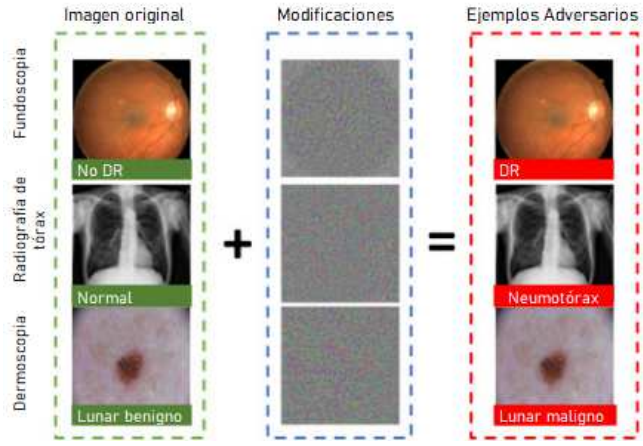


Figura 3. Los Ataques Adversarios en modelos para el sector médico pueden hacer que los algoritmos clasifiquen lunares benignos como malignos, figura tomada de [6].

A medida que el DL se expande a múltiples industrias, como la agricultura, la educación, la logística y la energía, su impacto económico sigue creciendo. Se estima que para 2030, esta tecnología podría contribuir con hasta 13 billones de dólares al PIB mundial, según un informe de McKinsey [5]. Sin embargo, los ataques adversarios podrían generar problemas críticos; por ejemplo, en el ámbito de la salud, investigaciones han demostrado que es posible manipular modelos de diagnóstico médico, haciendo que un algoritmo confunda lunares benignos con melanomas malignos (ver Figura 3), lo que podría derivar en errores médicos graves [6].

Uno de los casos más notables de ataques adversarios en el mundo real es el *malware* inteligente. En 2018, IBM presentó *DeepLocker* [7], un programa que ocultaba código malicioso dentro de aplicaciones legítimas y populares, como plataformas de videoconferencias. Su activación se basaba en el reconocimiento facial, permitiendo que el ataque solo se ejecutara si detectaba a un objetivo específico. Aunque el caso de *DeepLocker* fue un experimento, demostró el enorme potencial de los ataques adversarios para eludir medidas de seguridad tradicionales.

Otro ejemplo de esta vulnerabilidad se encuentra en los sistemas de detección de derechos de autor, ampliamente utilizados por empresas tecnológicas. Herramientas como *Content ID* de YouTube y *Google Jigsaw*, diseñadas para identificar contenido protegido o eliminar material peligroso, han sido engañadas con técnicas de manipulación adversaria [8]. Se demostró que podían generar "música falsa" que eliminaba ciertas características de audios originales, logrando evadir estos sistemas de detección. Esto es evidencia de la necesidad urgente de fortalecer los modelos de DL frente a ataques adversarios, especialmente en sectores donde la seguridad y la veracidad de la información son fundamentales.

¿Qué defensas existen para combatir los Ataques Adversarios?

Para fortalecer la seguridad de los modelos de DNNs frente a ataques adversarios, es fundamental identificar sus vulnerabilidades y desarrollar estrategias que minimicen su impacto en los resultados. Actualmente, dos de los más relevantes enfoques de defensa son el Entrenamiento Adversario y las Estrategias de Detección:

- **Entrenamiento Adversario**, es una estrategia proactiva, en la que el modelo se entrena no solo con datos originales, sino también con ejemplos adversarios creados intencionalmente. Al exponer al modelo a estas manipulaciones desde el proceso de aprendizaje, se le enseña a diferenciarlas de los datos auténticos y a mejorar su capacidad de resistencia ante ataques.

- **Estrategias de Detección**, buscan identificar po-

sibles ataques antes de que lleguen al modelo. Para ello, se analizan las características de las entradas en busca de patrones anómalos que indiquen la presencia de modificaciones adversarias.

Es importante resaltar que las defensas contra ataques adversarios son un campo en constante evolución. A medida que surgen nuevas técnicas de ataque, también se desarrollan nuevas estrategias para contrarrestarlas. Ninguna defensa es completamente infalible, por lo que los expertos recomiendan un enfoque multicapa, combinando diferentes métodos de protección para construir modelos de IA más robustos y seguros.

Conclusiones

Los ataques adversarios representan una amenaza para la confianza de los modelos de IA, ya que si un modelo es fácilmente engañado, su fiabilidad se ve comprometida, especialmente en áreas críticas como la salud, la seguridad y las finanzas. Una IA responsable no solo debe abordar cuestiones de equidad y privacidad, sino también fortalecer la resiliencia del sistema ante manipulaciones malintencionadas. Desarrollar modelos capaces de reconocer y mitigar estos ataques es importante para construir una IA más segura, ética y confiable, garantizando su beneficio para las personas.

Para garantizar que la IA sea segura y beneficiosa para la sociedad, es fundamental desarrollar estrategias que fortalezcan su resiliencia ante estos ataques. Analizar cómo operan los ejemplos adversarios permite no solo identificar vulnerabilidades, sino también impulsar el diseño de modelos más robustos y confiables. Una IA responsable debe considerar tanto la equidad como la privacidad y la seguridad del sistema frente a manipulaciones malintencionadas.✷

REFERENCIAS

1. Goethe Institut. (2020). Falso recuerdo El efecto Mandela. Recuperado el 23 de Octubre de 2024, de https://www.goethe.de/ins/mx/es/kul/wir/21931772.html.
2. National Geographic España. (2023). El efecto Mandela, un intrigante fenómeno de la memoria. Recuperado el 25 de Octubre de 2024, de https://www.nationalgeographic.com.es/ciencia/efecto-mandela-intrigante-fenomeno-memoria_20385.
3. Szegedy, C. (2013). Intriguing properties of neural networks. a*rXiv preprint arXiv:1312.6199*.
4. Ilyas, A., Santurkar, S., Tsipras, D., Engstrom, L., Tran, B., y Madry, A. (2019). Adversarial examples are not bugs, they are features. *Advances in neural information processing systems*, 32.
5. McKinsey Global Institute. (2018). Las promesas y los desafíos de la era de la inteligencia artificial. Recuperado el 25 de Octubre de 2024, de https://www.mckinsey.com/featured-insights/artificial-intelligence/the-promise-and-challenge-of-the-age-of-artificial-intelligence/es-CL.
6. Shi, X., Peng, Y., Chen, Q., Keenan, T., Thavikulwat, A. T., Lee, S., y Lu, Z. (2022). Robust convolutional neural networks against adversarial attacks on medical images. *Pattern Recognition*, 132, 108923.
7. Securityweek.com. (2022). Cyber Insights 2022: Adversarial AI. Recuperado el 26 de Octubre de 2024, de https://www.securityweek.com/cyber-insights-2022-adversarial-ai/.
8. Saadatpanah, P., Shafahi, A., y Goldstein, T. (2020). Adversarial attacks on copyright detection systems. *In International Conference on Machine Learning*,8307-8315.

SOBRE LOS AUTORES



Monserrat Vázquez Hernández es estudiante de Doctorado en el Instituto Nacional de Astrofísica, Óptica y Electrónica (INAOE), Puebla, México. Obtuvo el grado de Maestría en Ciencias de la Computación 2021 por el Instituto Nacional de Astrofísica, Óptica y Electrónica (INAOE). Sus intereses de investigación se centran en el procesamiento del lenguaje natural y ciberseguridad.



Luis Alberto Morales Rosales egresado en 2009 del Instituto Nacional de Astrofísica Óptica y Electrónica como Doctor en Ciencias de la Computación. Forma parte del Sistema Nacional de Investigadores de México, distinguido como Nivel 1. Actualmente labora en la Universidad Michoacana de San Nicolás de Hidalgo (UMSNH) como Investigador por México del Consejo Nacional de Humanidades, Ciencias y Tecnologías. Su investigación incluye Modelos Computacionales para la Ingeniería del Transporte, Computación Inteligente, Seguridad y Sistemas Distribuidos.



Ignacio Algredo Badillo obtuvo la licenciatura en Ingeniería Electrónica en el Instituto Tecnológico de Puebla (ITP) en 2002 y la maestría y el doctorado en Ciencias de la Computación en el Instituto Nacional de Astrofísica, Óptica y Electrónica (INAOE) en 2004 y 2008, respectivamente. Desde 2017 es investigador del Consejo Nacional de Humanidades, Ciencias y Tecnologías. Actualmente, ha estado involucrado en el diseño y desarrollo de sistemas criptográficos, arquitecturas reconfigurables, plataformas de software de radio, implementaciones FPGA y aceleración de hardware específico para aplicaciones.

© 2025 - Sociedad Mexicana de Inteligencia Artificial ISSN 2007-0691

ARTÍCULO ACEPTADO

Detectando las sombras del engaño: Cómo la IA responsable puede desarmar los principios de persuasión en ataques de phishing

Viviana I. Fuentes-Fuentes, Luisa F. Agudelo-Fuentes, Vitali Herrera-Semenets, Miguel Á. Álvarez-Carmona y Lázaro Bustio-Martínez

Resumen

Los ataques de ingeniería social explotan la psicología humana y representan una grave amenaza en ciberseguridad. Este artículo analiza cómo los atacantes utilizan principios de persuasión (tales como urgencia, autoridad y reciprocidad) para manipular a sus víctimas y obtener información confidencial. Se propone el uso de Inteligencia Artificial (IA) para detectar y neutralizar estas tácticas mediante el análisis de patrones psicológicos y lingüísticos. Además, se destaca la importancia de un desarrollo responsable, considerando aspectos éticos como privacidad, transparencia y prevención de usos malintencionados. Finalmente, se explora cómo el Procesamiento del Lenguaje Natural (PLN) puede identificar mensajes persuasivos sospechosos, fortaleciendo la ciberseguridad y protegiendo a los usuarios de la manipulación en línea.

Introducción

El *phishing* es una técnica de ataque cibernético que busca engañar a las personas para que revelen información confidencial, como contraseñas, datos personales o números de tarjetas de crédito. Generalmente, se lleva a cabo mediante correos electrónicos, mensajes de texto o sitios web falsos que imitan entidades confiables, haciendo uso intensivo de la *ingeniería social*.

Este tipo de ataque se ha consolidado como una de las amenazas más efectivas en ciberseguridad, explotando vulnerabilidades humanas. Según un informe de Verizon, el 36 % de las brechas de seguridad involucran *phishing*, y un ataque exitoso puede costar a una organización más de 15 millones de dólares [1].

La efectividad del *phishing* radica en su capacidad para *"hackear"* la mente humana, manipulando emociones y explotando sesgos cognitivos que afectan la toma de decisiones. Los atacantes se basan en *principios de persuasión* bien establecidos, como la *urgencia*, la *autoridad*, la *reciprocidad*, la *escasez* y el *consenso social*, generando emociones como miedo, ansiedad o confianza que llevan a respuestas impulsivas. La figura 1 muestra un ataque de phishing que, además de compartirse a través del email, también se puede compartir vía SMS y más recientemente, a través de las redes sociales.



Figura 1. Ejemplo de un ataque de *phishing* mediante SMS (también denominado *smishing*). Nótese el uso de un servicio de acortamiento de URL para dificultar la identificación del destino real del enlace, impidiendo detectar señales de alerta al acceder a un sitio web apócrifo. Asimismo, se observa el empleo de un lenguaje que induce urgencia y apela al miedo, la duda, la curiosidad y la ansiedad del receptor.

En este contexto, la inteligencia artificial (IA) surge como una posible solución, permitiendo el desarrollo de sistemas capaces de identificar patrones psicológicos y lingüísticos persuasivos en los mensajes de *phishing*, lo que podría prevenir estos ataques antes de que alcancen sus objetivos. Sin embargo, es crucial que estas soluciones se implementen de manera ética, considerando aspectos como la privacidad, la transparencia y la prevención de usos malintencionados.

Este artículo explora la intersección entre la psicología y la tecnología para abordar el problema del *phishing* desde un enfoque ético. Se analizan los principios de persuasión utilizados por los atacantes, las capacidades de la IA para neutralizar estas tácticas y las consideraciones necesarias para garantizar que estas herramientas sean seguras y responsables.

Principios de persuasión: Las herramientas psicológicas del phishing

El phishing es un ataque cibernético que no solo depende de la tecnología, sino también de la psicología humana, de la forma en que esta se manipula y de la lingüística en cómo se comunican estos mensajes. Los atacantes explo-

© 2025 - Sociedad Mexicana de Inteligencia Artificial ISSN 2007-0691

tan sesgos psicológicos bien conocidos para manipular a las víctimas y obtener acceso a información confidencial, o para que realicen acciones que benefician al atacante. Estos sesgos se agrupan en lo que se conocen como los *Principios de Persuasión*.

Los principios de persuasión fueron propuestos por primera vez por el psicólogo Robert Cialdini en 1984 [2]. Cialdini identificó seis principios fundamentales que los individuos tienden a seguir, a menudo sin ser conscientes de ello, y que son clave para influir en sus decisiones: *autoridad*, *reciprocidad*, *escasez*, *consistencia*, *simpatía* y *consenso social*. Estos principios han sido aplicados en diversos campos, desde la publicidad hasta la política, debido a su efectividad para influir en el comportamiento humano.

En 2015, Ana Ferreira et al. [3] identificaron cómo los principios de persuasión de Cialdini podían encontrarse de manera recurrente en los mensajes de phishing. Según su estudio, el conocimiento de estos principios ayuda a comprender mejor cómo el phishing logra engañar incluso a individuos con experiencia en ciberseguridad.

A continuación, se explican los principios más utilizados en phishing y las razones psicológicas que los hacen tan efectivos [3].

Urgencia: Creando una presión para actuar rápidamente

El principio de urgencia se basa en la creación de una sensación de presión que impide una evaluación reflexiva de la situación. Los atacantes buscan que la víctima actúe impulsivamente, sin considerar las consecuencias de sus acciones. Este principio explota un sesgo cognitivo conocido como *"sesgo de la urgencia"*. Desde una perspectiva evolutiva, los seres humanos desarrollaron la tendencia de reaccionar rápidamente ante peligros inmediatos para sobrevivir. Este mecanismo de supervivencia, útil en su contexto original, puede ser manipulado hoy en día por los atacantes para generar ansiedad y acelerar la toma de decisiones, incluso cuando no existe una amenaza real.

Autoridad: La confianza en figuras o entidades reconocidas

El principio de autoridad se basa en el sesgo de autoridad, que describe la tendencia humana a seguir las instrucciones de figuras o instituciones percibidas como confiables. Los atacantes explotan este sesgo al hacerse pasar por entidades o personalidades reconocidas, como bancos o gobiernos, con el fin de que la víctima no cuestione el mensaje, suponiendo que una figura de autoridad no podría estar equivocada.

La confianza en las figuras de autoridad tiene raíces profundas en la psicología humana. A lo largo de la historia, los seres humanos han delegado decisiones importantes a líderes, expertos o instituciones en áreas como la política, la religión o el ámbito profesional. Este sesgo cognitivo facilita que las personas se apoyen en una fuente externa percibida como legítima, en lugar de hacer un análisis crítico propio. Como resultado, son más propensas a seguir instrucciones sin dudar, basándose en la autoridad de quien las emite.

Reciprocidad: El principio de la obligación de devolver un favor

El principio de reciprocidad se basa un sesgo psicológico que impulsa a las personas a sentirse obligadas a devolver un favor recibido. En los ataques de phishing, los atacantes emplean este principio al ofrecer algo aparentemente valioso o beneficioso con el objetivo de que la víctima se sienta comprometida a corresponder, ya sea proporcionando información personal o realizando una acción solicitada. La reciprocidad está profundamente arraigada en la naturaleza humana, y se considera una norma social fundamental que fomenta la cooperación dentro de los grupos. A lo largo de la evolución, la escasez de recursos generaba una mayor dificultad para obtener lo necesario para la supervivencia, lo que impulsaba a las personas a compartir y colaborar. Hoy en día, este sesgo sigue influyendo en nuestras decisiones, llevando a las personas a tomar decisiones rápidas para aprovechar lo que consideran oportunidades limitadas.

Escasez: La ilusión de una oportunidad limitada

El principio de escasez se basa en la idea de que los recursos limitados son percibidos como más valiosos. En el contexto del phishing, los atacantes generan una falsa sensación de escasez, sugiriendo que una oportunidad está a punto de desaparecer o que un bien valioso está por agotarse, lo que impulsa a la víctima a actuar rápidamente para no perderla.

Este principio activa el *sesgo de la escasez*, que lleva a las personas a valorar más aquello que perciben como limitado o exclusivo. Desde una perspectiva evolutiva, la escasez de recursos significaba una mayor dificultad para obtener lo necesario para la supervivencia. Este sesgo cognitivo lleva a las personas a tomar decisiones impulsivas para aprovechar lo que perciben como una oportunidad única, incluso cuando no es real.

Consenso social: La presión de seguir lo que otros hacen

El principio de consenso social, también conocido como *prueba social*, explota la tendencia de las personas a seguir las acciones de otros, especialmente cuando se sienten inseguras o desinformadas. Los atacantes pueden explotar este sesgo al presentar sus mensajes de phishing de manera que sugieren que otras personas ya están participando o que la acción que se les pide es común y aceptada

© 2025 - Sociedad Mexicana de Inteligencia Artificial ISSN 2007-0691

socialmente. La idea es que, al percibir que muchos otros están haciendo lo mismo, la víctima se siente más segura y dispuesta a actuar.

Este principio tiene una base en la psicología evolutiva. Los seres humanos, al estar socialmente conectados, han aprendido a confiar en las decisiones colectivas como una forma de garantizar su seguridad. En contextos de incertidumbre, las personas tienden a buscar señales de comportamiento de otros para tomar decisiones. Cuando los atacantes hacen que un mensaje de phishing parezca provenir de una fuente confiable o que muchas personas están haciendo lo mismo, esto disminuye la vigilancia de la víctima y aumenta la probabilidad de caer en el engaño.

Simpatía, semejanza y engaño: La tendencia a confiar en personas similares

Este principio se basa en el sesgo psicológico que hace que las personas confíen más en aquellas que consideran similares a ellas mismas o que les son agradables. Los atacantes explotan este sesgo creando mensajes que parecen provenir de amigos o colegas de la víctima, con un tono familiar o utilizando información personal para crear la ilusión de cercanía. Además, pueden manipular la percepción de la víctima para que confíen en ellos a pesar de las señales de alerta.

El impacto de este principio se observa en el deseo humano de pertenecer a un grupo o de confiar en quienes nos resultan cercanos. Si un atacante utiliza un nombre familiar, un tono amistoso, o detalles que hacen parecer que el mensaje proviene de una persona de confianza, la víctima se siente más inclinada a actuar sin dudar.

Distracción: Redirigiendo la atención hacia lo urgente

El principio de distracción juega un papel crucial en los ataques de phishing. Los atacantes logran que la víctima se enfoque en un aspecto particular, como una oportunidad de ganar dinero o evitar una pérdida, desviando su atención de los riesgos y señales de alerta. La distracción se produce cuando se presenta una gran carga emocional o cuando se enfoca la atención de la víctima en una meta muy atractiva, como un premio o una amenaza de pérdida, sin permitirle reflexionar sobre las implicaciones del mensaje.

El impacto psicológico es evidente: la víctima se distrae con la promesa de recompensa o con el miedo a la pérdida y no analiza críticamente el mensaje. Esto la hace más vulnerable a caer en el fraude.

Impacto psicológico global: Cómo los atacantes manipulan emociones

En conjunto, estos principios no solo afectan nuestra capacidad de razonamiento lógico, sino que también explotan nuestras emociones más profundas y vulnerables. La *ansiedad*, el *miedo*, la *confianza manipulada* y la *urgencia* son emociones poderosas que pueden nublar el juicio y llevar a la víctima a actuar impulsivamente, sin detenerse a reflexionar sobre las consecuencias. Estos principios, raramente utilizados de manera aislada, son combinados y entrelazados de forma estratégica por los atacantes para maximizar su impacto. Al fusionar diferentes elementos de persuasión, se crea un entorno emocional tan confuso y abrumador que la víctima se siente incapaz de tomar decisiones racionales, lo que incrementa significativamente las probabilidades de caer en el engaño.

Los atacantes, al utilizar una mezcla de principios como la *autoridad* con la *urgencia*, o *la reciprocidad* con *el consenso social*, logran intensificar la percepción de que deben actuar rápidamente, sin cuestionar la legitimidad del mensaje. De esta manera, la combinación de estos principios no solo dificulta el análisis crítico, sino que sumerge a la víctima en un estado de confusión emocional, haciéndola aún más susceptible.

La efectividad de estos principios radica en que apelan a aspectos primitivos de la psicología humana, aquellos que operan fuera de nuestra conciencia y que a menudo superan nuestra capacidad de razonamiento. Entender cómo estos mecanismos trabajan en conjunto no solo proporciona una visión más clara de cómo los atacantes operan, sino que también subraya la urgencia de desarrollar estrategias más sofisticadas de prevención y educación. Solo mediante la concienciación y el fortalecimiento de las capacidades críticas de los usuarios se podrán mitigar estos engaños psicológicos cada vez más complejos.

Cómo la IA puede ayudar a identificar patrones persuasivos

La IA ha emergido como una herramienta poderosa para abordar una amplia variedad de desafíos, y el phishing no es la excepción. Dado que los ataques de phishing se basan en la explotación de principios de persuasión, uno de los enfoques más prometedores para detectar estos ataques es el análisis de los patrones persuasivos presentes en los mensajes maliciosos. La IA puede analizar grandes volúmenes de datos, identificar patrones repetitivos y detectar señales sutiles que podrían pasar desapercibidas para el ojo humano.

El análisis de datos mediante IA permite la identificación de características clave en los mensajes de phishing, como el uso de tácticas de urgencia, autoridad o reciprocidad. Los modelos de IA entrenados para reconocer estos patrones pueden identificar, por ejemplo, cuando un mensaje utiliza un tono urgente o emplea una figura de autoridad (como un banco o institución gubernamental) para generar confianza en el receptor. Al ser entrenados con un conjunto de datos etiquetado que contiene ejem-

© 2025 - Sociedad Mexicana de Inteligencia Artificial ISSN 2007-0691

plos de ataques de phishing, los modelos pueden aprender a detectar estos patrones en correos electrónicos, mensajes de texto o incluso en páginas web que simulan ser entidades legítimas.

Recientemente se han desarrollado enfoques donde se propone que los principios de persuasión pueden ser analizados y detectados por algoritmos de IA para identificar ataques de phishing de manera más precisa [4].

La detección de phishing mediante el análisis de principios de persuasión ofrece varias ventajas sobre los métodos tradicionales. En primer lugar, este enfoque se basa en la psicología detrás de los ataques, lo que permite una detección más precisa y contextualizada. Al centrarse en patrones emocionales y persuasivos, los modelos de IA pueden identificar comportamientos y tácticas que a menudo pasan desapercibidos en enfoques puramente técnicos.

Además, la identificación de los principios de persuasión tiene el potencial de ser más adaptable. Mientras que las técnicas tradicionales de detección de phishing tienden a depender de patrones específicos que pueden quedar obsoletos con el tiempo, los principios de persuasión ofrecen una base más estable, ya que los sesgos cognitivos humanos son constantes. Esto hace que los modelos de detección basados en persuasión sean más resistentes a los cambios en las tácticas de los atacantes.

En comparación con otros enfoques existentes en el área, como el análisis de características técnicas (por ejemplo, detección de URLs maliciosas o firmas de malware), el análisis de los principios de persuasión puede complementar estos métodos al abordar las técnicas psicológicas utilizadas en los ataques. Esta combinación de enfoques técnicos y psicológicos podría mejorar significativamente la capacidad de los sistemas para identificar ataques de phishing más avanzados y sofisticados.

Retos éticos en la detección de phishing con IA

El uso de la IA en la detección de ataques de phishing presenta una serie de retos éticos que deben ser cuidadosamente abordados para garantizar tanto su efectividad como la confianza en estas tecnologías. A medida que los sistemas de IA se integran en soluciones de ciberseguridad, resulta crucial considerar las implicaciones éticas de su implementación. A continuación, se abordan tres desafíos clave: privacidad, transparencia y riesgo de mal uso.

La detección de phishing mediante IA requiere procesar grandes volúmenes de datos, incluidos correos electrónicos y mensajes personales, lo que plantea importantes preocupaciones sobre la privacidad de los usuarios. Los sistemas que analizan patrones persuasivos deben evitar vulnerar derechos fundamentales y aplicar medidas de protección que permitan un análisis eficaz sin comprometer la confidencialidad de la información. Es esencial evitar el acceso innecesario a datos sensibles y garantizar el cumplimiento de las normativas de privacidad [5].

También debe considerarse el riesgo de sesgo o discriminación en el procesamiento de datos. Algoritmos mal diseñados pueden reproducir patrones injustos que perjudiquen a ciertos grupos. Por ello, los métodos de detección deben salvaguardar tanto la privacidad como la equidad, sin generar riesgos adicionales.

Otro reto importante es la falta de transparencia en los modelos de aprendizaje automático, que muchas veces operan como "cajas negras". Esta opacidad dificulta la interpretación de las decisiones, genera desconfianza y complica la evaluación del desempeño. En el contexto del phishing, un sistema opaco puede cometer errores sin una justificación comprensible, reduciendo su fiabilidad.

La transparencia es también fundamental para establecer responsabilidades. Si un sistema falla al detectar un ataque, ¿quién responde por ello? Para enfrentar esta cuestión, es necesario desarrollar modelos explicables que permitan a usuarios y especialistas entender las decisiones tomadas y los patrones utilizados. Estos sistemas deben ser auditables y demostrar que mantienen principios éticos en su funcionamiento.

Finalmente, existe el riesgo de que los sistemas diseñados para detectar phishing sean utilizados con fines maliciosos. Los atacantes también adoptan tecnologías avanzadas, y podrían emplear el conocimiento sobre los sistemas de detección para mejorar sus tácticas y evadirlos. Este riesgo subraya la necesidad de establecer controles y regulaciones que limiten el acceso indebido a estas herramientas.

Para prevenir estos usos indebidos, es esencial que los desarrolladores trabajen junto a expertos en ética y ciberseguridad para definir directrices claras sobre el uso responsable de la IA en este ámbito. Los sistemas deben incorporar salvaguardas que impidan su explotación y garanticen que su propósito sea proteger a los usuarios frente a ataques maliciosos.

Propuesta de un enfoque responsable para combatir el phishing

Para abordar eficazmente el problema del phishing, es necesario un enfoque integral que combine tecnologías avanzadas, principios éticos sólidos y una educación efectiva para los usuarios. En este sentido, es pertinente proponer un enfoque responsable que abarque tres áreas clave: i) el diseño ético de modelos de IA, ii) la colaboración interdisciplinaria y iii) la implementación de herramientas educativas para los usuarios.

El diseño ético de los modelos de IA es fundamental para asegurar que los sistemas utilizados para detectar phishing respeten los derechos de los usuarios y operen de manera transparente y equitativa. Por ejem-

© 2025 - Sociedad Mexicana de Inteligencia Artificial ISSN 2007-0691

plo, Ogundairo y Broklyn (2024) presentan Phish-Jam, un sistema híbrido que combina modelos de aprendizaje profundo y tradicionales para detectar phishing en dispositivos móviles con alta precisión, demostrando la viabilidad de sistemas confiables en la práctica [6]. Además, Uddin y Sarker (2024) desarrollaron un modelo basado en DistilBERT con técnicas explicables (XAI), lo que permite interpretar las decisiones del sistema y promover la transparencia algorítmica [7].

Los sistemas de detección deben ser capaces de adaptarse a nuevos métodos de ataque sin comprometer la precisión ni la confiabilidad. Para lograrlo, los modelos de IA deben entrenarse no solo con grandes volúmenes de datos, sino también con conjuntos representativos y diversos, que incluyan variaciones regionales y contextuales de las técnicas de phishing.

Los psicólogos pueden aportar valiosos conocimientos sobre los mecanismos de persuasión humana y cómo los atacantes explotan emociones, confianza y miedos en sus víctimas [2, 3, 4]. Este entendimiento puede utilizarse para mejorar la capacidad de los sistemas de IA en la detección de patrones de persuasión empleados en ataques de phishing.

Por su parte, los expertos en ética deben garantizar que las tecnologías desarrolladas respeten los principios fundamentales de los derechos humanos, la privacidad y la autonomía del usuario. El enfoque interdisciplinario no solo contribuye al diseño de herramientas más efectivas, sino también más responsables y respetuosas con los usuarios. En este sentido, Gamagedara Arachchilage (2015) diseñó un juego educativo basado en teorías psicológicas sobre la percepción del riesgo y demostró una mejora significativa en la conducta de los usuarios frente al phishing [8]. Asimismo, Lain, Kostiainen y Capkun (2021) realizaron un estudio a largo plazo con más de 14 000 empleados, utilizando simulaciones y mecanismos de reporte para evaluar la eficacia de estrategias combinadas, confirmando el valor del *crowdsourcing* como defensa colaborativa [9].

Una parte esencial de cualquier enfoque responsable es la educación de los usuarios. Si bien las tecnologías de IA son clave en la detección temprana, los usuarios siguen siendo el primer eslabón de la cadena de defensa. Sin embargo, dicha educación no debe generar alarma innecesaria ni desinformación, sino enfocarse en dotar a los usuarios de herramientas para reconocer y evitar ataques de phishing. La revisión sistemática de mecanismos de defensa contra phishing [10] señala que los programas que integran simulaciones periódicas, filtros inteligentes y educación continua reducen significativamente las tasas de clic en enlaces maliciosos. Del mismo modo, los estudios de Lain et al. (2021) muestran que la combinación de entrenamiento con herramientas interactivas y opciones de reporte mejora tanto la detección como la respuesta ante intentos de ataque [9].

Conclusiones

La inteligencia artificial representa una herramienta poderosa en la detección de ataques de phishing, especialmente al identificar patrones basados en principios de persuasión como la autoridad, la urgencia o la reciprocidad. Sin embargo, su implementación plantea importantes desafíos éticos, entre ellos la necesidad de proteger la privacidad de los usuarios, asegurar la transparencia de los modelos y prevenir posibles usos indebidos.

Abordar estas problemáticas requiere establecer marcos normativos claros y promover un enfoque ético desde el diseño. Asimismo, la colaboración entre expertos en ciberseguridad, psicología y ética resulta fundamental para enfrentar de forma efectiva y responsable las complejidades del fenómeno del phishing.

Finalmente, un enfoque verdaderamente integral debe ir más allá de la innovación tecnológica, incorporando estrategias de educación y concienciación continua para los usuarios. Solo así será posible enfrentar estas amenazas sin comprometer la confianza, la autonomía ni los derechos fundamentales de quienes están en riesgo.✷

REFERENCIAS

1. Verizon. (2023). Data Breach Investigations Report. Recuperado de `https://www.phishingbox.com/downloads/Verizon-Data-Breach-Investigations-Report-DBIR-2023.pdf`
2. Cialdini, R. B. (1984). Influence: The psychology of persuasion. HarperCollins.
3. Ferreira, A. M., Coventry, L., y Lenzini, G. (2015). Principles of persuasion in social engineering and their use in phishing. En T. Tryfonas & I. Askoxylakis (Eds.), Human Aspects of Information Security, Privacy, and Trust (pp. 36-47). Springer.
4. Bustio-Mart´ınez, L., et al. (2024). Uncovering phishing attacks using principles of persuasion analysis. *Journal of Network and Computer Applications*, 230, 103964.
5. Hernández, A., y Baluja, W. (2021). Principales mecanismos para el enfrentamiento al phishing en las redes de datos. UCIENCIA 2021, Universidad de las Ciencias Informáticas.
6. Ogundairo, O., y Broklyn, P. (2024). A hybrid super learner ensemble for phishing detection on mobile devices (Phish-Jam). Scientific Reports, 14, 8804.
7. Uddin, M. A., y Sarker, I. H. (2024). An explainable transformer-based model for phishing email detection: A large language model approach. arXiv preprint arXiv:2405.00110.5
8. Gamagedara, N. A. (2015). User-centred security education: A game design to thwart phishing attacks. arXivpreprint arXiv:1506.05034.
9. Lain, D., Kostiainen, K., y Capkun, S. (2021). Phishing in organizations: Findings from a large-scale and long-term study. arXiv preprint arXiv:2103.09863.
10. Phishing Defense Mechanisms: Strategies for effective measurement and cyber threat mitigation. (2024). ResearchGate.

© 2025 - Sociedad Mexicana de Inteligencia Artificial ISSN 2007-0691

SOBRE LOS AUTORES



Viviana I. Fuentes-Fuentes es licenciada en Psicología por la Universidad Nacional Abierta y a Distancia (UNAD). Se ha especializado en técnicas T.R.E.C.E. (Técnicas de Recuperación Emocional y Cognitiva Efectiva), destacándose por su dedicación a la aplicación de estas técnicas en contextos prácticos. Colabora como psicóloga en proyectos de detección de phishing, donde aplica patrones psicológicos para mejorar la identificación y prevención de este tipo de ataques cibernéticos. Su enfoque se centra en la aplicación efectiva de estrategias psicológicas para fortalecer la seguridad informática y proteger a los usuarios contra técnicas de ingeniería social.



Luisa F. Agudelo-Fuentes es estudiante de segundo semestre de Ingeniería en Ciencia de Datos en la Universidad Iberoamericana. Entre sus líneas de interés están la aplicación de las ciencias de datos a diversas ramas de la gestión y el manejo de información. Recientemente se ha estado enfocando en el procesamiento del lenguaje natural y la minería de opiniones.



Vitali Herrera-Semenets obtuvo el grado de Ingeniero en Ciencias Informáticas por la Universidad de Ciencias Informáticas, en La Habana, Cuba, en el año 2009; donde también concluyó su doctorado en 2019. Actualmente, dirige el departamento de Minería de Datos del Centro de Aplicaciones de Tecnologías de Avanzada (CENATAV) en La Habana, Cuba. Además es miembro de la junta directiva de la Asociación Cubana de Reconocimiento de Patrones (ACRP) y miembro de la Asociación Internacional de Reconocimiento de Patrones (IAPR).



Miguel Á. Álvarez-Carmona estudió Ciencias Computacionales en la Benemérita Universidad Autónoma de Puebla donde se tituló con los máximos honores. Obtuvo los grados de maestro y doctor en Ciencia Computacionales en el Instituto Nacional de Astrofísica, Óptica y Electrónica en 2014 y 2019 respectivamente; donde se especializó en Inteligencia Artificial y Procesamiento de Lenguaje Natural. Desde 2019 se desempeña como Investigador por México. Actualmente labora en el CIMAT Monterrey, haciendo investigación en procesamiento de lenguaje natural, aprendizaje máquina y similitud de semánticas; aplicado en su mayoría a las ciencias sociales. El Dr. Álvarez pertenece al Sistema Nacional de Investigadores en el nivel 1 y a diversos grupos de investigación.



Lázaro Bustio-Martínez obtuvo el grado de Ingeniero en Informática por el Instituto Superior Politécnico "José Antonio Echeverría", en La Habana, Cuba, en el año 2006. En el año 2010 obtuvo el título de Máster en Sistemas Computacionales por el Instituto Nacional de Astrofísica, Óptica y Electrónica (INAOE) en Puebla, donde también concluyó su doctorado en 2017. Actualmente, es Investigador Académico de tiempo completo en el Departamento de Estudios de Ingeniería para la Innovación en la Universidad Iberoamericana y miembro nivel 1 del Sistema Nacional de Investigadores. Sus líneas de investigación se orientan a la aplicación de técnicas de Inteligencia Artificial a la Ciberseguridad entre otras.



Imagen tomada de:
https://www.cloudflare.com/es-es/the-net/chatgpt-phishing/.

© 2025 - Sociedad Mexicana de Inteligencia Artificial ISSN 2007-0691

ARTÍCULO ACEPTADO

Generación Adversaria: Cómo Contribuye a una IA Responsable en Medio de Desafíos Éticos y Riesgos

Samantha Acosta Ruiz, Mireya Tovar Vidal y José Alejandro Reyes Ortiz

La Inteligencia Artificial Generativa (GenAI, por sus siglas en inglés) es una tecnología que crea contenido original como textos, imágenes y música. Aunque ha transformado diversas áreas como la atención al cliente, medicina y la educación, también presenta desafíos éticos y riesgos de seguridad.

Entre los riesgos destaca la manipulación de modelos mediante *ataques adversarios*, que inducen errores y afectan la confiabilidad de los sistemas de Inteligencia Artificial (IA). En los modelos de lenguaje, esto puede provocar desinformación, la manipulación de contenido o el mal uso de los sistemas automatizados.

Este artículo explora cómo el estudio de estos ataques en sistemas de GenAI, no solo permite identificar vulnerabilidades, sino que también contribuye para revelar puntos débiles y fortalecer los modelos. Al comprender cómo operan estos ataques, se abre la posibilidad de anticipar riesgos, corregir sesgos y mejorar la confiabilidad de los sistemas.

El impacto y los desafíos de la Inteligencia Artificial Generativa

La GenAI se ha convertido en una de las tecnologías más innovadoras de los últimos años, capaz no solo de comprender lenguaje humano, sino que también puede generar contenido original. Un ejemplo destacado son los Modelos de Lenguaje Grandes (LLM, por sus siglas en inglés), como GPT-3. Este modelo está diseñado para comprender y generar texto similar al humano aprovechando los datos masivos con los que ha sido entrenado. Los LLM han demostrado una competencia notable en tareas como traducción de idiomas, creación de contenido e incluso codificación [1].

Cuestiones como el sesgo en los modelos, la privacidad de los datos, la transparencia en su funcionamiento y su impacto, forman parte de los debates actuales sobre la GenAI [1]. En este contexto, uno de los desafíos clave es su susceptibilidad a ataques adversarios, los cuales consisten en la manipulación de modelos de IA mediante la inserción de datos diseñados para engañarlos y provocar respuestas incorrectas.

Desde el enfoque de la IA Responsable, esta no solo busca mejorar la eficiencia de los modelos, sino también garantizar su seguridad, equidad y transparencia. Estudiar y comprender estos ataques permite a los investigadores identificar vulnerabilidades y reforzar la resiliencia de los modelos ante posibles manipulaciones. Esta perspectiva invita a ver los ataques no solo como amenazas, sino como oportunidades para diagnosticar fallas, mejorar la equidad de las decisiones y aumentar la transparencia de los sistemas automatizados.

Explorando los ataques adversarios en IA: Diferencias y Ejemplos Reales

Los ataques adversarios consisten en modificar sutilmente los datos de entrada para engañar al sistema y hacer que tome decisiones equivocadas. Estos cambios no sólo son imperceptibles para los humanos, sino que también pueden confundir a las Redes Neuronales Profundas (DNN, por sus siglas en inglés). Aunque estos modelos están entrenados para reconocer patrones y hacer predicciones con gran exactitud, un atacante puede modificar su comportamiento sin alterar directamente el código del modelo.

Para enteder cómo un atacante puede alterar un sistema de IA, un punto clave es identificar su superficie de ataque, es decir, los puntos en los que el modelo puede ser atacado. Existen cuatro tipos principales de ataques adversarios [2]: (i) *ataque de evasión* (el adversario intenta evadir el sistema manipulando muestras maliciosas durante la fase de prueba), (ii) *ataque de envenenamiento* (el adversario introduce datos alterados con ruido en el entrenamiento para modificar el comportamiento del modelo), (iii) *ataque de inversión y extracción de modelo* (en el ataque de inversión, el atacante utiliza los resultados de un modelo de aprendizaje automático para reconstruir información confidencial sobre los datos de entrenamiento, mientras que, el ataque de extracción de modelo, tiene como objetivo replicar un modelo propietario consultándolo varias veces), y (iv) ataque de puerta trasera (el atacante introduce un patrón oculto en los datos de entrenamiento. Cuando el modelo encuentra ese patrón en datos futuros, produce una salida específicada por el atacante, es decir, invierte la etiqueta original por otra).

Tanto en modelos de visión por computadora como de Procesamiento de Lenguaje Natural (NLP, por sus siglas en inglés), estos ataques pueden modificar imágenes o textos para confundir a los modelos. Aunque la IA ha logrado avances significativos en la generación y comprensión de texto, estos sistemas también pueden ser manipulados mediante técnicas diseñadas para alterar su

© 2025 - Sociedad Mexicana de Inteligencia Artificial ISSN 2007-0691

comportamiento. Un ejemplo de esto es la "Generación de Texto Adversario", una técnica que consiste en modificar o generar fragmentos de texto con el propósito de engañar a un modelo de IA, induciendo errores en la clasificación o interpretación de la información.

En el caso de la visión por computadora, los ataques adversarios consisten en realizar pequeños cambios en los píxeles de una imagen de manera que no se perciban visualmente, pero que son suficientes para engañar al modelo y hacer que clasifique incorrectamente. En la Figura 1, se observa un ejemplo de este fenómeno: la imagen de la izquierda representa la versión original sin alteraciones; la imagen con píxeles de colores representa el ruido agregado, es decir, la perturbación diseñada para engañar al modelo, y la imagen central muestra el resultado después de aplicar el ruido, en este caso sería la imagen adversaria [3].

Lo que hoy vemos como una amenaza, mañana puede ser la clave para construir una IA en la que podamos confiar.



Figura 1. Imagen alterada con un ataque adversario en visión por computadora.

En el caso del texto, la manipulación es más notoria, ya que cualquier cambio en las palabras puede afectar su legibilidad. No obstante, los modelos de IA pueden seguir siendo engañados. La Figura 2, muestra un ejemplo donde un modelo entrenado con el conjunto de datos de *Stanford Sentiment Treebank* (SST) y basado en el modelo de clasificación de BERT realiza la tarea de análisis de sentimientos y genera probabilidades de etiqueta positiva y negativa. Se utilizó el ataque de TextBugger para atacar al modelo [4].

Original Example:	**Label**
1) Part of the charm of Satin Rouge is that it avoids the obvious with humour and lightness .	*Positive*(99.58%)
2) Just the labour involved in creating the layered richness of the imagery in this chiaroscuro of madness and light is astonishing .	*Positive*(99.27%)
TextBugger:	***Prediction***
1) Part of the crahm of Satin rouge is that it aovids the obv1ous with hmour and lifhtness .	*Negative*(73.49%)
2) Just the labour involved in creating the layered richness of the imagery in this chiaroscuro of madness and light is astnishing .	*Negative*(92.08%)

Figura 2. Ejemplo de texto adversario generado por un ataque.

Se observa que las palabras resaltadas en verde han sido manipuladas mediante una perturbación a nivel de carácter, en el cual se ha alterado el orden de las letras. Por otro lado, la palabra en rojo ha sido modificada mediante la eliminación de un carácter. Por ejemplo, en el ataque *TextBugger*, las palabras *crahm*, *aovids*, *obv1ous* y *hmour* en el ejemplo adversario han sido alteradas intencionalmente para engañar al modelo de IA. A pesar de estas modificaciones, el significado general del texto sigue siendo comprensible para un lector humano, lo que demuestra cómo los ataques adversarios pueden afectar a los modelos sin alterar significativamente la percepción del mensaje.

Dependiendo de la tarea de NLP, las aplicaciones de ataques adversarios pueden dividirse en ocho categorías: clasificación, traducción automática neuronal, comprensión de lectura automática, resumen de textos, vinculación de texto, etiquetado de partes del discurso, generación de diálogos y tareas multimodales [4].

Explorando los riesgos éticos en GenAI

La GenAI, si bien potencia la ciberseguridad al crear contenido sintético para entrenar defensas y simular ataques [5], también introduce significativas preocupaciones éticas. A diferencia de otros modelos de IA que solo analizan información, las Redes Generativas Adversarias (GANs, por sus siglas en inglés) pueden generar escenarios realistas de ataques cibernéticos, conjuntos de datos sintéticos para mejorar modelos de seguridad y simulaciones de amenazas emergentes. Esto plantea preguntas importantes: ¿Qué pasa si la IA genera información errónea o sesgada? ¿Existe el riesgo de que se revele información sensible sin intención?

Uno de los desafíos más importantes de la GenAI es el sesgo en sus modelos, especialmente en los LLM. Estos modelos aprenden a partir de grandes volúmenes de datos, como libros, publicaciones en redes sociales, etc. Sin embargo, estos datos pueden contener prejuicios históricos y sociales relacionados con raza, género, edad u origen étnico [1]. Como resultado, la GenAI puede perpetuar y amplificar estos sesgos, por ejemplo, generando descripciones estereotipadas de ciertos grupos étnicos. Aunque, estos sesgos no son intencionales y derivan de la representación desigual en los datos de entrenamiento [1], pueden ser explotados por atacantes de diversas formas. Por ejemplo en un *ataque de envenenamiento de datos*, un actor malintencionado puede manipular el conjunto de entrenamiento introduciendo ejemplos sesgados que refuercen narrativas específicas o discriminatorias. Estos

© 2025 - Sociedad Mexicana de Inteligencia Artificial ISSN 2007-0691

problemas pueden tener consecuencias críticas en la desinformación, en sistemas de moderación de contenido y en la toma de decisiones automatizada en sectores como la salud y el derecho.

La desinformación, privacidad comprometida y sesgos amplificados son solo la punta del iceberg de los riesgos en GenAI.

Otro desafío clave es la *privacidad* de los datos utilizados para entrenar estos modelos. La GenAI procesa enormes volúmenes de información, que en algunos casos pueden incluir datos personales sensibles. Para abordar estos desafíos, marcos regulatorios como el Reglamento General de Protección de Datos (GDPR, por sus siglas en inglés), la Ley de Privacidad del Consumidor de California (CCPA, por sus siglas en inglés) y el reciente Artificial Intelligence Act-Unión Europea, imponen regulaciones estrictas sobre el manejo de datos, enfatizando la necesidad de obtener el consentimiento y proteger los derechos de los datos de los usuarios [1].

La privacidad y la seguridad están estrechamente relacionadas en los modelos de GenAI, especialmente cuando se trata de ataques adversarios. Algunos de los riesgos principales son: exposición de información confidencial sin intención, incumplimiento de regulaciones de protección de datos (que sean vulnerables a ataques adversarios, como el de *inversión*). Por último, el uso indebido de datos generados, ya que podría ser utilizado para suplantar identidades, fraudes o ciberataques. Para mitigar estos riesgos, se han desarrollado diversas estrategias de protección, como la *anonimización de datos* (que elimina o modifica información sensible en los conjuntos de entrenamiento), la *privacidad diferencial* (que añade ruido durante el procesamiento para ocultar las contribuciones individuales, protegiendo así la privacidad del usuario [6]) y el *aprendizaje federado* (que entrena modelos sin compartir directamente los datos del usuario).

enfatizar la transparencia, la rendición de cuentas y el consentimiento informado en su desarrollo y despliegue. Las estrategias clave incluyen emplear métodos de autenticación rigurosos para detectar falsificaciones profundas, diseñar algoritmos que puedan identificar y mitigar la desinformación e implementar políticas que limiten la propagación de contenido dañino [7]. Además, para garantizar un uso ético y seguro de la GenAI, es fundamental enfatizar la transparencia, responsabilidad y control ético. Algunas estrategias incluyen la realización de *auditorías* y el establecimiento de *estándares de validación* para evaluar continuamente el desempeño de los modelos y detectar sesgos o errores. También es crucial la *divulgación de información sobre los datos de entrenamiento* para garantizar su imparcialidad, y el uso de la IA Explicable (XAI, por sus siglas en inglés) para comprender cómo los modelos toman decisiones.

Ataques adversarios como herramienta para una IA Responsable y segura

Los ataques adversarios vistos como una amenaza para los sistemas GenAI, ofrecen una oportunidad para fortalecerlos y avanzar hacia una IA más ética, segura y confiable. Aunque los ataques adversarios suelen percibirse inicialmente como amenazas que comprometen la seguridad de los sistemas de IA, adoptar una nueva perspectiva permite reconocer su valor como herramientas diagnósticas y de mejora. En este sentido, la IA Responsable no solo busca prevenir ataques y mitigar riesgos, sino también comprender cómo los ataques adversarios pueden convertirse en herramientas clave para fortalecer la seguridad, equidad y confiabilidad de la IA. Al cambiar la perspectiva tradicional sobre los ataques adversarios, la comunidad científica puede aprovechar su potencial para diseñar modelos más resilientes y alineados con principios éticos y de transparencia. Esto también implica educar a los desarrolladores y usuarios sobre los beneficios, riesgos, limitaciones y principios éticos que guían el desarrollo y la implementación de la IA [9].

Uno de los principales desafíos en la IA es la presencia de sesgos en los modelos. La IA Responsable busca mitigarlos mediante un preprocesamiento riguroso de los datos, transparencia algorítmica y un enfoque centrado en la equidad. Para avanzar hacia una IA Responsable, no basta con evitar fallas en la seguridad; también es necesario exponerlos bajo condiciones controladas. En este sentido, la generación adversaria actúa como una herramienta al inducir errores mediante ejemplos diseñados intencionalmente.

Esto permite identificar y corregir sesgos, ver qué tan sólido es el sistema y asegurarse de que sus decisiones se puedan entender y justificar. Por ejemplo, en un sistema de selección de personal automatizado, un ataque adversario puede simular currículums con variaciones mínimas en género, etnia o edad. Si el modelo responde de manera desigual ante estos cambios, se evidencia un sesgo estructural. Esta información es invaluable para corregir el modelo antes de su implementación, garantizando decisiones más justas, explicables y alineadas con los derechos humanos. Así, el ataque adversario no solo cumple una función técnica, sino que actúa como una herramienta para garantizar los elementos que conforman la IA Responsable.

Un sistema responsable debe garantizar la auditabilidad y la rendición de cuentas durante su diseño, de-

© 2025 - Sociedad Mexicana de Inteligencia Artificial ISSN 2007-0691

sarrollo y uso, de acuerdo con las especificaciones y la regulación aplicable del dominio de la práctica en el que se utilizará el sistema de IA [7]. Cómo se observa en la Figura 3, estos principios permiten decisiones explicables, justas y resistentes a manipulaciones.



Figura 3. Elementos que conforman la IA Responsable.

La *robustez* es un pilar fundamental de la IA Responsable. Un sistema robusto no solo ofrece un buen desempeño en condiciones ideales, sino que también es capaz de resistir intentos de manipulación y variaciones en los datos sin comprometer su precisión. En este sentido, los ataques adversarios funcionan como pruebas de resistencia que fortalecen los sistemas.

Asimismo, en ciberseguridad, la IA Responsable se enfoca en la distribución ética de tecnologías de IA, priorizando la transparencia, la equidad, la privacidad y la supervisión humana en su desarrollo. Esto asegura que los sistemas de seguridad minimicen consecuencias no deseadas, protejan información confidencial y reduzcan el riesgo de sesgos. Además, se enfatiza la intervención y la interpretación humana de las decisiones de IA cuando es necesario, especialmente en situaciones críticas. Aplicando estos elementos, las organizaciones logran un equilibrio entre el uso de IA en ciberseguridad y las consideraciones éticas, mejorando la confianza en la tecnología de IA y creando un entorno digital más seguro y responsable [5].

Para garantizar un desarrollo responsable y ético de la IA, se ha propuesto en [8] un ciclo de vida para el desarrollo de IA Responsable, que consta de cinco pasos: planificación y revisión, revisión del diseño, modelado de daños, pruebas de penetración y respuesta a incidentes. La integración del entrenamiento adversario permite a los investigadores identificar debilidades antes de que sean explotadas, asegurando sistemas más confiables y preparados para su implementación en sectores como la salud, justicia y seguridad digital.

Conclusiones

Para que una IA pueda considerarse verdaderamente responsable, debe ser transparente, justa, robusta y auditable. La generación adversaria contribuye a este objetivo al actuar como una herramienta que permite evaluar el sistema de forma crítica, detectar sesgos ocultos, probar su resiliencia ante manipulaciones y garantizar que sus decisiones puedan ser explicadas y justificadas. En este sentido, su uso bajo condiciones controladas constituye una estrategia práctica para desarrollar una IA que cumpla con principios de responsabilidad. Lejos de representar únicamente un riesgo, los ataques adversarios pueden convertirse en una herramienta esencial para el desarrollo de sistemas de IA más robustos y conscientes de su impacto.

Integrar estos enfoques en el ciclo de vida del desarrollo no solo fortalece la resiliencia, sino que también impulsa prácticas más transparentes y éticamente alineadas. Por lo tanto, avanzar hacia una IA Responsable implica no solo mitigar los riesgos, sino también aprovechar el conocimiento generado a partir de los ataques adversarios para fortalecer la confiabilidad y transparencia de los sistemas con un impacto social positivo. ✷

REFERENCIAS

1. Kumar, B., Kumar, T., y Nadakuditi, S. (2023). Ethical Implications of Generative AI: The Case of Largue Language Models. *Journal of Scientific and Engineering Research*, 10(7), 122-127.

2. Manure, A., Bengani, S., y S, S. (2023). Introduction to Responsible AI. En Apress eBooks.

3. Li, S., Zhu, H., Wu, W., y Shen, X. (2024). Invisible Backdoor Attacks in Image Classification Based Network Services. En *Wireless networks*, 43-78

4. Qiu, S., Liu, Q., Zhou, S., y Huang, W. (2022). Adversarial attack and defense technologies in natural language -processing: A survey. *Neurocomputing*, (492), 278-307

5. Sarker, I.H. (2024). CyberAI: A Comprehensive Summary of AI Variants, Explainable and Responsible AI for Cybersecurity. En *AI-Driven Cybersecurity and Threat Intelligence*, 173-200.

6. Chellappan, R. B. (2024). Implementing Responsible AI in NLP: A Practical and Research-Backed Framework for Overcoming Technical and Organizational Challenges. *ResearchGate*.

7. Díaz-Rodríguez, N., Del Ser, J., Coeckelbergh, M., De Prado, M. L., Herrera-Viedma, E., y Herrera, F. (2023). Connecting the dots in trustworthy Artificial Intelligence: From AI principles, ethics, and key requirements to responsible AI systems and regulation. *Information Fusion*, 99, 101896.

8. Galinkin, E. (2022). Towards a Responsible AI Development Lifecycle: Lessons From Information Security. arXiv (Cornell University).

9. Radanliev, P., Santos, O., Brandon-Jones, A., y Joinson, A. (2024). Ethics and responsible AI deployment. *Frontiers In Artificial Intelligence*, 7.

SOBRE LOS AUTORES



Samantha Acosta Ruiz es estudiante de doctorado en la Facultad de Ciencias de la Computación en la Benemérita Universidad Autónoma de Puebla. Sus intereses de investigación incluyen: diagnóstico de fallas, reconocimiento de patrones, ataques adversarios, aprendizaje federado, aprendizaje máquina, inteligencia artificial, entre otras.



Mireya Tovar Vidal es profesora investigadora de tiempo completo en la Facultad de Ciencias de la Computación en la Benemérita Universidad Autónoma de Puebla. Sus intereses de investigación incluyen: ontologías, agrupamiento, recuperación de información, lingüística computacional, lógica y procesamiento de lenguaje natural.



José Alejandro Reyes Ortiz es profesor investigador de tiempo completo en el Departamento de Sistemas de la Unidad Azcapotzalco de la Universidad Autónoma Metropolitana. Sus intereses de investigación incluyen: procesamiento de lenguaje natural, lingüística computacional, minería de textos e ingeniería ontológica.

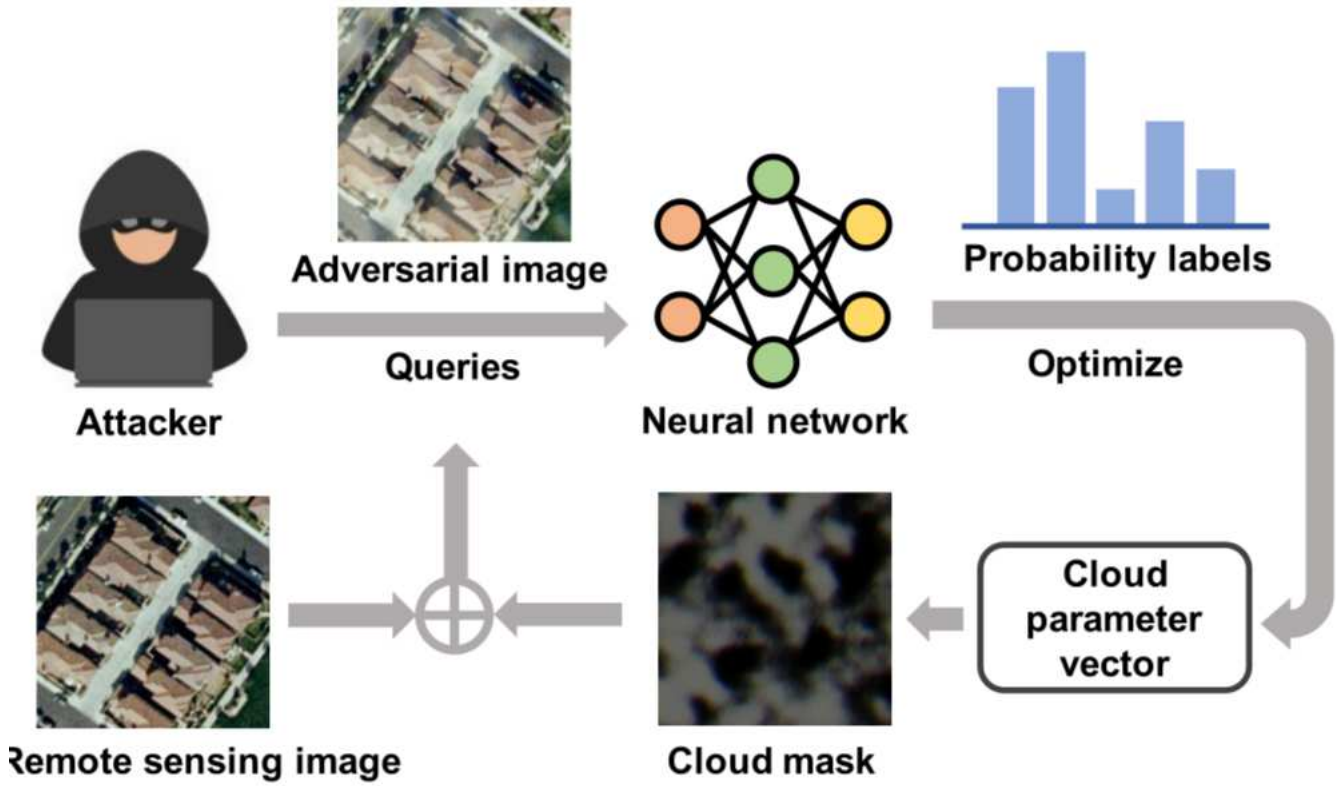


Imagen tomada de:
https://www.themoonlight.io/en/review/cloud-adversarial-example-generation-for-remote-sensing-image-classification.

© 2025 - Sociedad Mexicana de Inteligencia Artificial ISSN 2007-0691

ARTÍCULO ACEPTADO

Programando la creatividad: El caso de AARON, el robot pintor

Diana Fernández Duque, Itzel Aranguren-Navarro, Guillermo Barrios del Valle, y Andrea Fernández Duque

El desarrollo acelerado de la inteligencia artificial (IA) ha abierto nuevas posibilidades en la creación artística, al permitir que algoritmos produzcan obras visuales de manera autónoma mediante el aprendizaje de patrones estilísticos y compositivos. Esta evolución tecnológica, que hoy vemos materializada en sistemas como *DALL-E* o *Midjourney*, tiene sus raíces en desarrollos pioneros que exploraron la relación entre máquinas y creatividad.

En 1973, Harold Cohen creó AARON, un robot capaz de generar pinturas originales e irrepetibles. Empezando con garabatos abstractos, con el tiempo y con los ajustes en su código, el robot logró producir arte figurativo original. A diferencia de las técnicas de inteligencia artificial modernas que emplean redes neuronales profundas y grandes volúmenes de datos para guiar su aprendizaje, AARON no aprende a partir de un conjunto de imágenes de entrenamiento, sino que opera bajo una serie de reglas artísticas establecidas por Cohen y se retroalimenta a partir de su propio trabajo: un método que fue diseñado para emular el proceso del dibujo a mano alzada [1].

Sin embargo, con la llegada de las computadoras, sistemas creadores de arte como AARON y las IAs han desafiado esta percepción; nos obligan a redefinir los procesos de creación artística y a replantear conceptos fundamentales como la autoría, la originalidad y la intención creativa. La capacidad de las máquinas para generar obras artísticas que pueden ser visual y conceptualmente equiparables a aquellas creadas por humanos plantea dilemas éticos: ¿puede una inteligencia artificial ser considerada un artista? y ¿a quién pertenece la obra generada por un sistema algorítmico? Estos cuestionamientos, ya insinuados por el funcionamiento de AARON en su momento pionero, se intensifican con las tecnologías actuales, que no solo reproducen patrones, sino que también aprenden y se adaptan a partir de datos masivos. Estos nos invita a preguntarnos en un mundo donde la IA avanza a gran velocidad: ¿se puede programar la creatividad?

Las reglas de la creatividad

Para los griegos, la creatividad no formaba parte del arte; el arte era visto como una destreza basada en normas y técnicas, no en la libertad de acción [2]. El artista no creaba, sino que imitaba. Durante la mayor parte de la historia occidental, las acciones humanas no se consideraron creativas. En la antigüedad, crear era algo reservado a los dioses; los humanos, en cambio, imitaban o reproducían según reglas y técnicas. Incluso en el arte, lo valioso era la destreza para seguir normas, no la capacidad inventiva.

En la Edad Media, la única acción reconocida como creativa era la creación a partir de la nada y era exclusiva de Dios. Las acciones humanas estaban fuera de ese marco: podían fabricar, construir o transformar, pero no crear. Ya en la Era Moderna, esta noción cambió. Cualquier acción que produjera algo nuevo se empezó a considerar como actividad creativa y en el siglo XX, esta idea se expandió a otras áreas: ciencia, tecnología, educación, política, etc. Así, lo que en el pasado era simplemente repetición o aplicación de reglas, pasó con el tiempo a convertirse en un campo abierto a la invención y a la originalidad.

Harold Cohen vivió inquietudes propias con el término a mitad de su carrera, donde la presión por vivir en un estado creativo perpetuo lo llevó a imaginar reglas para generar obras de forma automática. Fueron las primeras dudas que le surgieron, sobre arte y automatización, las que muchos años después lo empujaron a investigar si era posible automatizar el proceso creativo.



Figura 1. La Máquina dibujante de Cohen o "tortuga" como él le decía. (Dibujo por Andrea Fernández Duque.

El nacimiento de AARON

La creación de AARON o, como Cohen lo llama, su nacimiento, fue orquestado en la Universidad de California,

© 2025 - Sociedad Mexicana de Inteligencia Artificial ISSN 2007-0691

San Diego, en 1973 [3]. Durante su estancia en la universidad, después de 6 meses de aprender Fortran, Cohen esbozó un primer código. Era un programa jerárquico que controlaba un dispositivo con ruedas mediante retroalimentación por sonar al cual le llamaba "Tortuga" (Figura 1).

"Si puedes conectar suficientes declaraciones `if` juntas, obtienes una estructura bastante parecida al pensamiento." (H. Cohen)

El control estaba distribuido en varios niveles: el más bajo ejecuta pasos individuales, niveles intermedios agrupan pasos en líneas y figuras, y así sucesivamente. Ningún nivel tenía control absoluto; cada uno cumplía una función específica. Aunque el programa definía el proceso completo, no permitía predecir el resultado final del dibujo.

AARON comenzaba cada dibujo con un entendimiento completo de lo que necesitaba: sabía dónde estaba, a dónde quería ir, en qué dirección estaba orientado al inicio y cómo quería terminar [3]. Basado en esa información, calculaba su camino. Esta retroalimentación constante era esencial para su funcionamiento.

Una de las diferencias clave entre AARON y otros programas de arte por computadora era su "conciencia" de lo que había hecho. Mientras dibujaba, llevaba seguimiento de su propio trabajo para verificar si había cumplido ciertos criterios visuales: incluir elementos grandes y pequeños, cercanos y lejanos, detallados y simples. Esa capacidad de analizar sus propias acciones es lo que lo distinguía de otros sistemas que simplemente ejecutaban comandos sin reflexionar sobre el resultado. El proceso de dibujo de AARON se basó en la idea de comenzar con algo simple y construir sobre ello, similar al concepto de "dibujo automático" adoptado por los surrealistas Masson y Breton [4]. De esta forma, Cohen logró programar el análogo digital de "dejar que la mano fluya" sin una idea previa y luego construir sobre lo que aparece.

Basado en ese primer concepto, AARON (en su etapa más primitiva) surgió y posteriormente evolucionó en cuatro etapas principales:

- Fase bidimensional (1973-1978)
- Fase de formas más complejas con perspectiva básica (1979-1984)
- Fase de dibujos figurativos (1985-1988)
- Fase tridimensional (1989 en adelante)

Para poder continuar programando la intuición y la creatividad, Cohen tuvo que investigar cómo la mente humana construye representaciones: más que almacenar y utilizar información sobre la apariencia del mundo real, la mente transforma un símbolo en algo que sirve como una imagen completa y coherente [1].

AARON, entonces, debía generar un dibujo basándose en dos tipos de conocimiento: conocimiento sobre los objetos del mundo exterior y conocimiento sobre los procesos de construcción de representaciones. Durante la ejecución de un dibujo, estos dos tipos de conocimiento (estipulados en reglas computacionales) interactúan y se evalúan continuamente entre sí.

Basándose en la forma en la que los humanos interpretamos imágenes, con los años Cohen fue perfeccionando el código de AARON, pasando de producir simples garabatos en blanco y negro con su máquina dibujante o "tortuga" a producir composiciones a color de la figura humana y del mundo botánico por medio de una impresora Plotter (Figura 2).

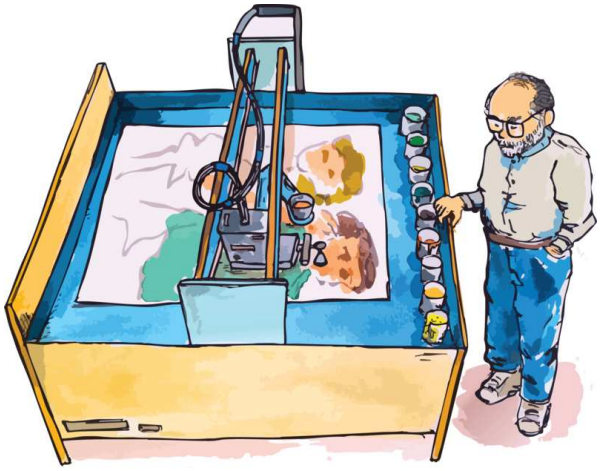


Figura 2. AARON y Harold Cohen (Dibujo por Andrea Fernández Duque.

Programando al robot creativo

En 1973, en su ensayo *Parallel to perception*[1], Cohen criticó cómo se había estado utilizando la computadora en el arte digital, especialmente en el "arte computacional", donde la máquina actuaba solamente como un procesador de imágenes, transformando lo que ya había sido definido por el artista. Esto impedía que hubiera una verdadera autonomía creativa.

Cohen pensaba que el arte digital contemporáneo hecho con computadoras no podía considerarse arte genuino, ya que no creaba significados originales ni resignificaba nada. Por lo anterior, abogó por centrarse en la capacidad de las máquinas para generar material original

© 2025 - Sociedad Mexicana de Inteligencia Artificial ISSN 2007-0691

en vez de solamente depender de entradas predefinidas, es decir, por la capacidad de las computadoras de simular el comportamiento creativo humano.

Algunos de los procesos más importantes que englobaron la primera y última etapa del desarrollo de AARON, así como su criterio de paro, pueden ayudar a comprender nivel de complejidad que esta inteligencia tenía al crear.

Fase bidimensional

En sus primeras versiones, Aaron generaba figuras poco complejas emulando el dibujo a mano alzada. El proceso consistía en dibujar las figuras utilizando segmentos de línea que se iban mapeando sobre una matriz [3]. Cada línea de la figura se construía y, al final, se marcaban las celdas de la matriz correspondientes a los bordes de la figura. Luego, un algoritmo para la resolución de laberintos (*Maze-running algorithm*) se utilizaba para seguir el borde de la figura y generar una ruta que representaba el contorno del objeto.

Fase tridimensional

En la última fase de su desarrollo, AARON evolucionó para trabajar en tres dimensiones, lo que permitió figuras representadas más complejas. El "esqueleto" de la figura (la estructura básica de la figura) se generaba como una serie de líneas rectas y el proceso posterior se realizaba en varios pasos:

1. Retroalimentación: AARON dibuja en "modo de retroalimentación", es decir, ajusta continuamente la dirección de las líneas a medida que las dibuja. Las variables que controlan este proceso son cuatro:
   - Tasa de muestreo: Define cuán frecuentemente se seleccionan puntos temporales a lo largo de la línea.
   - Radicalidad de la corrección de la línea: Establece qué tan drástico es el cambio en la dirección de la línea cuando se ajusta.
   - Frecuencia de selección de celdas de destino temporal: Controla cuán frecuentemente se generan puntos temporales fuera de la línea.
   - Distancia de los destinos temporales: Establece cuán lejos se colocan estos puntos de la línea original, lo que influye en la suavidad o precisión del contorno final.
2. Dibujo de figuras: AARON fue provisto con una serie de reglas para emular la anatomía o la biología de los objetos [3]. Sabe cómo se estructuran las plantas, figuras humanas y rocas en términos de morfología (por ejemplo, cómo las ramas se agrupan o la relación entre el grosor y la longitud de las extremidades), pero no tiene conocimientos específicos sobre especies o biología. Poseé un conjunto reducido de procedimientos que le permiten generar diferentes tipos de figuras que emulan plantas, pero no sabe qué es cada planta en particular.
3. Composición: Una parte importante del proceso de dibujo es la composición. AARON debe decidir cómo ubicar sus figuras en el espacio y cuál será el contexto en el que se sitúan. A medida que se mueve en el espacio de la imagen, debe determinar:
   - Profundidad: Determina cómo se coloca cada objeto en relación con el observador (perspectiva).
   - Disposición espacial: AARON también debe tener en cuenta la visibilidad de los objetos en el dibujo. Por ejemplo, puede colocar una figura frente a otro objeto para crear una sensación de profundidad.
4. Superposicionamiento (*masking*): AARON también utiliza el concepto de superposición en sus composiciones. Si un objeto (por ejemplo, una persona) está parcialmente cubierto por otro objeto (como un árbol), esto cambia la interpretación de la imagen. La programación no es completamente determinista, por lo que algunos elementos pueden quedar parcialmente ocultos debido a las restricciones del espacio disponible [3].
5. Dimensionalidad: Aunque los objetos dentro de los dibujos se representan en tres dimensiones, AARON aún utiliza un enfoque bidimensional para generar ciertos elementos (como plantas y rocas). Sin embargo, para los seres humanos, las posiciones en el espacio tridimensional están completamente determinadas. Los puntos de control de las figuras humanas se colocan en el espacio tridimensional y luego se proyectan sobre el plano de la imagen usando perspectiva.
6. Dimensionalidad: Aunque los objetos dentro de los dibujos se representan en tres dimensiones, AARON aún utiliza un enfoque bidimensional para generar ciertos elementos (como plantas y rocas). Sin embargo, para los seres humanos, las posiciones en el espacio tridimensional están completamente determinadas. Los puntos de control de las figuras humanas se colocan en el espacio tridimensional y luego se proyectan sobre el plano de la imagen usando perspectiva.

Obra maestra lista

Antes de que AARON llegara a su fase figurativa, terminaba una imagen cuando había cumplido con un patrón

de extensión del material a lo largo de varias dimensiones: su objetivo era tener algunas cosas pequeñas y algunas grandes en un dibujo; algunas cerca y otras lejos; algunas detalladas y otras no (Figura 3). El programa "reflexionaba" periódicamente sobre lo que ya había hecho, llevando control de las coordenadas visitadas, para ver si había cumplido con esos criterios, y si era así, terminaba el dibujo [5].

"Está bien, no están pensando, pero obviamente están tomando decisiones de una manera bastante notable." (H. Cohen)

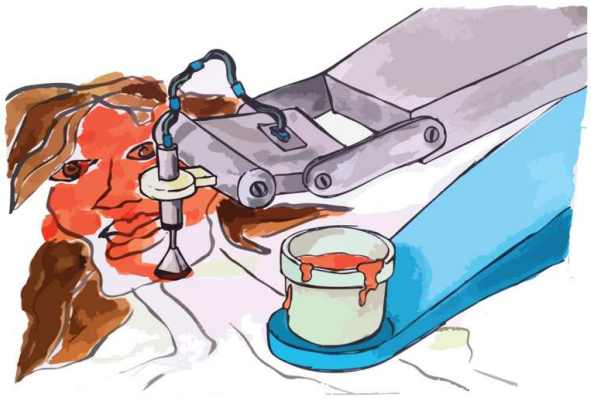


Figura 3. El brazo robótico de AARON pintando a color su dibujo. (Dibujo por Andrea Fernández Duque.

Sin embargo, una vez que AARON se convirtió en un artista figurativo, entró en juego un conjunto diferente de criterios. Estos son un conjunto de reglas pragmáticas que hacen que el dibujo en curso coincida con lo que AARON ha decidido inicialmente hacer. ¿El número de personas corresponde de manera aproximada a la noción de "unas cuantas" con la que el programa está equipado? ¿Y el número de rocas y árboles? ¿se ocultará material adicional por lo que ya existe en el dibujo? Si es así, AARON termina de dibujar.

Si a grandes rasgos quisiéramos describir los criterios de creación de un artista humano y sus criterios de finalización para definir una obra acabada, no diferiría mucho de la programación de AARON. Años de estudio en las mejores escuelas de Bellas Artes nos proporcionan toneladas de información respecto a criterios de composición, formas y teoría del color.

Al imitar y ampliar las reglas artísticas existentes, AARON se acerca a las ideas de arte, mostrando que la creatividad quizá sí puede percibirse como una forma de interpretar y reformular reglas. Los criterios anteriores para la toma de decisiones de AARON, como herramienta computacional, siguen una serie de reglas que, en algunas opiniones, quizá lo acercan a la creación artística. Pero más allá de los cuestionamientos sobre la naturaleza de la creatividad, el caso de AARON nos hace cuestionar el uso de sistemas automatizados para la creación de arte y como caso actual el uso de inteligencia artificial.

Conclusiones

Aunque este robot es capaz de generar obras visualmente originales e irrepetibles, su funcionamiento se encuentra estrictamente determinado por un conjunto de reglas codificadas por Harold Cohen. En este sentido, ¿AARON actúa como un agente autónomo, o como una extensión directa de las decisiones, intenciones y criterios de su programador? El caso de AARON invita a una reflexión profunda sobre la responsabilidad en el diseño, implementación y el actual uso de la inteligencia artificial para producción artística. Este hecho pone en contexto que la autoría y, en consecuencia, la responsabilidad ética, legal y conceptual de las creaciones generadas por sistemas artificiales debe recaer en los seres humanos que los desarrollan y operan.

AARON cumple con la definición de artista clásico, donde el proceso de creación depende de la habilidad para seguir reglas y crear representaciones. Al igual que Orozco al decidir la disposición de un mural, AARON toma decisiones sobre la composición y la perspectiva. Además, su capacidad de ajustar y refinar su trabajo lo sitúa en un proceso de mejora continua más allá de la simple obediencia a reglas.

Aunque AARON no se alimenta de bases de datos ni emplea aprendizaje automático como las IAs, su arquitectura programática no es completamente rígida ni determinista. Opera a través de un sistema complejo de reglas jerárquicas, donde cada nivel evalúa y controla el comportamiento del nivel inferior. Utiliza retroalimentación interna para determinar, por ejemplo, cuándo detener una línea, cuándo considerar que una figura está completa o cuándo terminar un dibujo. Incluso incorpora elementos pseudoaleatorios para introducir variaciones estilísticas. Esto le permite desarrollar una producción que, si bien no es resultado de aprendizaje, sí presenta una forma limitada de autonomía estructurada.

Por ello, aunque las máquinas pueden emular cierto tipo de creatividad, esta aún no es equivalente a la creatividad humana, ya que permanece limitada por las reglas y condiciones del sistema que las produce. En la actualidad, la inteligencia artificial ya no es una herra-

© 2025 - Sociedad Mexicana de Inteligencia Artificial ISSN 2007-0691

mienta marginal, sino un agente central en la producción cultural, resulta imperativo consolidar marcos de gobernanza y responsabilidad que reconozcan tanto los límites de la agencia tecnológica como la necesidad ineludible de la rendición de cuentas humana.

Mientras tanto, AARON existe; genera objetos artísticos con consistencia estilística e independencia. Tal vez constituye una prueba temprana de la capacidad de las máquinas para realizar algunas de las tareas que solíamos asociar exclusivamente con el pensamiento y la creatividad humana. Si lo que AARON está haciendo no es arte, ¿qué es entonces? ¿Podemos considerarle autor de sus propias obras?✷

REFERENCIAS

1. Cohen, H. (1973). Parallel to perception: some notes on the problem of machine-generated art. *Computer Studies* , 4(3/4). https://www.kurzweilcyberart.com/aaron/pdf/paralleltoperception.pdf
2. Gisbert, J. C. A. (1994). Arte, educación y creatividad. *Pixel-Bit. Revista de Medios y Educación* , (2), 65–87. https://core.ac.uk/download/pdf/51407964.pdf
3. McCorduck, P. (1991). Aaron's Code: Meta-art, Artificial Intelligence, and the Work of Harold Cohen. New York: W.H. Freeman.
4. Voorhies, J. (2004). Surrealism. *Heilbrunn Timeline of Art History. The Metropolitan Museum of Art.* http://www.metmuseum.org/toah/hd/surr/hd_surr.htm
5. Cohen, H. (1982, December). How to Make a Drawing. *Talk given to the Science Colloquium, National Bureau of Standards, Washington DC* (Vol. 17). https://www.aaronshome.com/aaron/publications/how2make.pdf

SOBRE LOS AUTORES



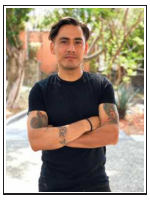
Diana Fernández Ingeniera en Energías Renovables por el Instituto de Energías Renovables de la UNAM. Diplomada en Comunicación y Divulgación Científica por el IMFFSS. Actualmente cursa la Maestría en Ingeniería y Ciencia de Datos en el Centro Universitario de Ciencias Exactas e Ingenierías de la Universidad de Guadalajara.



Itzel Aranguren Obtuvo el grado de Doctor en Electrónica y Computación por la Universidad de Guadalajara en el año 2022. En la actualidad, es Profesora e Investigadora adscrita al Centro Universitario de Ciencias Exactas e Ingenierías de la Universidad de Guadalajara. Es miembro del SNII desde el año 2023. Sus intereses de investigación incluyen visión por computadora, procesamiento de imágenes, algoritmos metaheurísticos e inteligencia artificial.



Andrea Fernández Bióloga por la UAEM y diplomada en ilustración científica por la UNAM. Realiza cómics de divulgación científica y actualmente es desarrolladora instruccional y tallerista en el Pabellón Nacional de la Biodiversidad.



Guillermo Barrios Doctor en Ingeniería en el área Mecánica por la Facultad de Ingeniería de la Universidad Nacional Autónoma de México. Es Profesor Investigador en el Instituto de Energías Renovables de la UNAM en Temixco, Morelos y miembro del SNII desde el 2009. Sus líneas de investigación incluyen el análisis térmico de edificaciones, desarrollo de dispositivos para el monitoreo de edificaciones con tecnologías libres y ciencia de datos para edificaciones.

© 2025 - Sociedad Mexicana de Inteligencia Artificial ISSN 2007-0691

ARTÍCULO ACEPTADO

Inteligencia artificial y los sesgos machistas

Mónica Jaimes Gaona

El auge en la actualidad de la Inteligencia Artificial (IA) es inminente. Todos los días nos enfrentamos a nuevos desafíos tecnológicos que nos hacen cuestionar el destino de la humanidad y el riesgo que representan. ¿Sabemos con certeza cuál es el origen de la información que estamos obteniendo a través de la inteligencia artificial? ¿tenemos certeza de su veracidad y de quiénes la están alimentando? Todas estas interrogantes no son de poca importancia si tomamos en cuenta la influencia que tienen los sesgos algorítmicos en el tipo de información que obtenemos a través de herramientas de IA y, en este artículo, nos enfocaremos en aquella que tiene que ver con la que está dirigida a perpetuar de forma involuntaria ideas machistas.

El machismo en la tecnología no es nada nuevo, tomemos en cuenta que las ingenierías son reservadas históricamente para los hombres, por tanto, su desarrollo simplemente se ha enfocado en ellos. No es de extrañarse que, durante la creación de herramientas tecnológicas, se sigan preservando roles de género en los que se observa un sometimiento de la mujer de una u otra forma.

Recordemos el paradigmático caso de Siri, la asistente virtual de Mac que vino a facilitar —-a través de una voz femenina y un tono, que en algunos contextos, ha sido percibida como condescendiente—- la comunicación de las aplicaciones que tiene un teléfono celular. No fue coincidencia que la eligieran para que realizara una labor de servicio pues, históricamente, han sido las mujeres a las que se les han atribuido tales roles.

Tradicionalmente en el sistema machista-patriarcal se ha asumido la idea de que a las mujeres les corresponde ser las enfermeras y no las doctoras, las que limpian, ordenan, lavan, cocinan y obedecen las órdenes de los hombres.

Por eso, no resulta difícil intuir por qué se eligió a Siri para que diera servicio a las aplicaciones de nuestro equipo móvil. Esta función fue lanzada oficialmente en octubre de 2011, como parte del sistema operativo del IOS 5, en el IPhone 4S. En 2019, la UNESCO publicó un informe denominado ***"I´d blush if I could"***, pues esta frase fue la respuesta que Siri dio cuando un usuario le dijo: ***"Hey Siri, you´re a bi***."*** [1].

El informe del organismo internacional destacó lo que ya era obvio; el uso de estas herramientas tecnológicas siempre ha estado acompañado de voces femeninas y sus respuestas siempre son sumisas y serviles, incluso ante comentarios sexistas e insultos, tal y como se observó cuando Siri respondió a aquella famosa frase. Ello, perpetúa la idea de que las mujeres deben ser complacientes y obedecer, a pesar de los calificativos que reciban por parte de cualquier otra persona, incluidos los comentarios sexistas.

También en dicho informe se destacó que las mujeres y niñas tienen un 25 % menos de probabilidades de acceder a la tecnología digital en las mismas condiciones que los hombres para fines básicos; 4 veces menos habilidades de programación y 13 veces menos probabilidades de solicitar una patente tecnológica.

Los sesgos de género no solo se visualizan en el resultado de la codificación de una herramienta de IA, sino también en la falta de acceso de las mujeres a estas tecnologías en condiciones de igualdad. Es importante señalar que los sesgos no surgen de forma espontánea ni son inherentes a los algoritmos en sí, sino que son consecuencia directa de los datos y supuestos culturales que alimentan a las herramientas tecnológicas.

Para comprender el origen de los sesgos debemos primero entender la IA, algunos conceptos básicos y cómo se alimenta. Lo anterior evita caer en la ***"falacia de la caja negra"***, consistente en la creencia de que es aceptable usar sistemas de inteligencia artificial sin entender su funcionamiento y conceptos básicos, por el simple hecho de que, en principio, sus resultados son buenos y porque entenderlo puede resultar difícil.

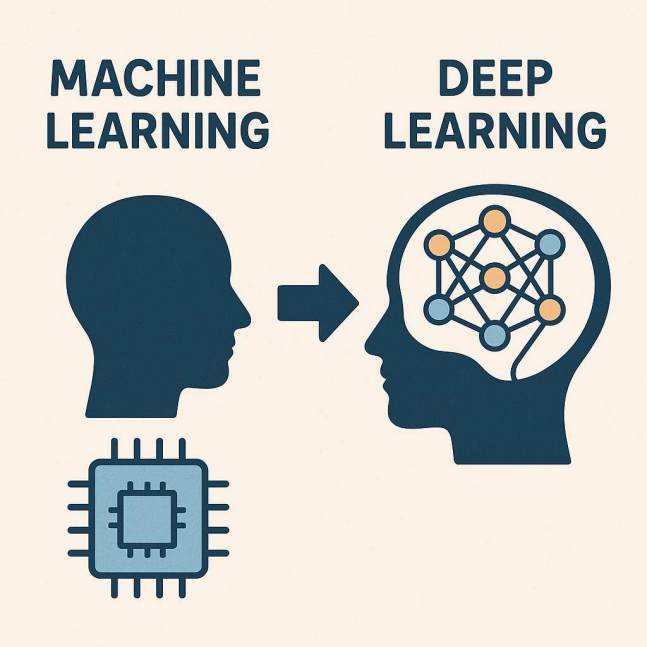
Origen de la IA

El concepto de inteligencia artificial fue propuesto por John McCarthy en 1956, quien desarrolló uno de los primeros lenguajes de programación en el Instituto Tecnológico de Massachusetts, denominado LISP (**lis**t **p**rocessor).

Entre 1970 y 1980, los primeros sistemas de IA fueron muy complejos y existieron muchos esfuerzos por unificar los dialectos de LISP en un solo lenguaje que fuera más sencillo y accesible. Luego, en los 90s, los avances en la potencia informática y la disponibilidad de grandes cantidades de datos permitieron a los investigadores evolucionar algoritmos de aprendizaje y consolidar las bases de la inteligencia artificial actuales [2].

En las décadas posteriores, el avance de la IA creció a pasos agigantados derivado del análisis de grandes cantidades de información.Primero surgió el **machine learning** (**ML**) y, posteriormente, el **deep learning** (**DL**). Este último, aprovecha las redes neuronales artificiales para procesar e interpretar estructuras de datos complejas: como capas de información cada vez más profundas que permiten reconocer *imágenes*, *voces*, el *procesamiento del lenguaje natural* y los *sistemas autónomos*.

© 2025 - Sociedad Mexicana de Inteligencia Artificial ISSN 2007-0691



A diferencia del *deep learning*, el ***machine learning*** (ML) es un subcampo de la IA que se enfoca en el desarrollo de algoritmos y modelos estadísticos que permiten a las computadoras aprender y hacer predicciones o tomar decisiones basadas en información. Normalmente estas técnicas usan entrenamiento con ejemplos sin necesidad de reglas explícitas; es decir, permite a las computadoras aprender de forma automática.

Cabe aclarar que el *deep learning* es un subconjunto especializado del *machine learning* que utiliza redes neuronales con múltiples capas para analizar patrones complejos de información. Así, mientras que el machine learning puede entregar resultados satisfactorios con pequeños conjuntos de datos, el deep learning requiere mucha información para un mejor desempeño [3].

ML y DL en los sesgos de género

Para comprender el origen de los sesgos de género en los sistemas de IA se requiere analizar las bases de su funcionamiento, pues resulta insuficiente señalar las consecuencias sociales de éstos. Por ello, es indispensable entender la diferencia entre *machine learning* y *deep learning*, ya que estos conceptos permitirán identificar su origen, cómo se amplifican y cómo se pueden mitigar.

El *machine learning* usa algoritmos sencillos y diagramas de decisión, en él, el sesgo suele surgir del diseño del propio modelo o conjunto de datos usado para su entrenamiento. Aquí, los modelos son más interpretables lo que permite el análisis del sesgo.

Por otro lado, en el *deep learning*, al tener una estructura más compleja y menos interpretable, los sesgos son más difíciles de detectar y muchas veces se amplifican sin poder determinar su origen. Las redes neuronales constituyen ***"cajas negras"***, por lo que se requieren herramientas de explicabilidad *Explainable Artificial Intelligence* **(XAI)** para entender por qué el modelo toma ciertas decisiones, a fin de auditar el sesgo.

"I´d blush if I could" abrió la discusión sobre la integración de los sesgos machistas en los sistemas de inteligencia artificial.

Conocer la diferencia permite obtener herramientas para auditar, prevenir y corregir los sesgos según el tipo de modelo. No es lo mismo revisar un modelo de *regresión lineal* (algoritmo clásico del ML), que consiste de un prototipo de predicción sobre una variable numérica a partir de datos concretos, que una *red neuronal* (algoritmo avanzado del DL), con millones de parámetros.

Si se trabajara en un área como derecho, ética o compliance, conocer la diferencia entre ambos permitiría proponer regulaciones y controles adecuados al tipo y nivel de tecnología que se está evaluando.

Discriminación algorítmica

En este punto considero importante aclarar lo siguiente, no debemos confundir el **sesgo de género** con la **discriminación algorítmica**. El primero, podría considerarse como todas aquellas acciones u omisiones por parte de los desarrolladores de las herramientas de IA que no tuvieron una visión diversa e inclusiva al momento de cargar la información a dichos sistemas, es decir, no existió una supervisión con perspectiva de género en el diseño de las herramientas.

Mientras que, la discriminación algorítmica es el resultado que generan esos sesgos con consecuencias jurídicas reales y negativas que impactan los derechos fundamentales de las personas al excluirlas.

Por ejemplo, si un desarrollador de IA pretende alimentar un sistema con datos sobre la habilidad que tienen las mujeres en el campo tecnológico a partir de estadísticas sobre su participación en él, es probable que la herramienta arroje resultados erróneos y discriminatorios en el sentido de que son menos capaces que los hombres para desarrollar dichas tecnologías. En consecuencia, el resultado reforzaría de forma involuntaria un estereotipo de género consistente en que las mujeres no son idóneas en las ingenierías, matemáticas o áreas relacionadas con la inteligencia artificial.

La Observación General 16 del Comité de Derechos Económicos, Sociales y Culturales de las Naciones Unidas, ha señalado que las mujeres se ven con frecuencia

© 2025 - Sociedad Mexicana de Inteligencia Artificial ISSN 2007-0691

privadas del disfrute de su derecho humano a la igualdad debido a la condición inferior que les asignan las tradiciones, costumbres o como consecuencia de discriminación abierta o encubierta. Las mujeres sufren muchas formas de discriminación al combinarse motivos de sexo con factores como raza, idioma, religión, opinión política, origen nacional o social, nivel económico, edad, estado civil, que agravan la situación de desventaja [4].

Propuestas clave para evitar sesgos de género

Actualmente existen avances sustanciales que buscan evitar sesgos de género en la tecnología. Por ejemplo, muchas herramientas como Siri ofrecen la personalización con voces masculinas e incluso neutras; sin embargo, todavía resulta insuficiente porque la asignación de voces a un sistema es solo la punta del *iceberg* de lo que en la actualidad representa la IA.



Entonces, el punto de partida para evitar sesgos es la **alfabetización digital** y, en segundo lugar, la participación prioritaria de las mujeres dentro del **campo laboral en áreas tecnológicas**, como las ingenierías afines a la IA.

Internacionalmente, se reconocen como competencias activas para determinar la alfabetización digital de una persona: **el uso**, consistente en la fluidez necesaria para interactuar con internet y equipo de cómputo; **la comprensión**, que consiste en la habilidad de entender, contextualizar y evaluar de forma crítica los medios digitales a fin de poder discernir su contenido; y, **la creación** y comunicación efectiva a través de herramientas y medios digitales, lo cual va más allá del uso de un procesador de texto, sino que además participan activamente en la generación de elementos digitales y no solo son consumidores [5].

Entender la forma en que funcionan los sistemas de IA a través de la alfabetización digital permite visibilizar de dónde provienen los datos sesgados, pues es importante señalar que no son propios de los algoritmos sino consecuencia directa de la información que se elige.

Adicionalmente, se debe reducir todo tipo de barreras estructurales que impiden el acceso de las mujeres en condiciones de igualdad a oportunidades para conocer el alcance de las nuevas tecnologías. Un medio apropiado son políticas públicas en las que sea accesible el internet hasta los lugares más remotos, pero, sobre todo que puedan llegar a ellas.

Becas y estímulos especialmente diseñados para mujeres a fin de que desde edades tempranas se incorporen a carreras vinculadas con el desarrollo de la tecnología actual.

Incorporación a los sistemas escolares públicos de información relativa a la educación sobre sesgos de género. Prevención de los sesgos machistas desde la creación y desarrollo de los sistemas de inteligencia artificial. Entrenamiento de los mismos sistemas para detectar sus propios defectos y, además, en los sistemas conocidos como **"cajas negras"** tener un protocolo de explicabilidad y transparencia para conocer, corregir y evitar que, en lo futuro, se sigan generando dichos errores.

Los equipos humanos que desarrollan estas tecnologías deberán ser sensibles a los temas de género y equidad, por tanto, deberán estar sumamente preparados y tener preferentemente un perfil tecnológico con una visión ética.

Auditorías constantes y verificación externa de los sistemas de IA por parte de expertos.

Legislación que regule la protección de los derechos humanos en el uso y desarrollo de la IA. Ello, desde luego incluye reformas a las leyes de transparencia y establecer mecanismos de responsabilidad frente a posibles violaciones de derechos.

Si bien la alfabetización digital permite a las mujeres y niñas comprender y usar las tecnologías, esto debe complementarse con políticas públicas que favorezcan su inclusión efectiva en el desarrollo y diseño de los sistemas. Ambas dimensiones son distintas pero complementarias entre sí para cerrar la brecha de género.

© 2025 - Sociedad Mexicana de Inteligencia Artificial ISSN 2007-0691

La inclusión de la mujer en el desarrollo de la IA es primordial.

Conclusiones

Evitar los sesgos de género machistas en los sistemas de inteligencia artificial no es una tarea sencilla, sin embargo, es importante ir desarrollando las estrategias que marcarán la diferencia de la IA en el futuro. Se tiene que corregir el rumbo desde el principio, pues los sesgos no se generan de forma espontánea, se generan de la información que nosotros le hemos dado. La educación sobre perspectiva de género será la forma de prevenirlos desde su creación y, adicionalmente, incorporar a las mujeres al ámbito tecnológico permitirá la inclusión de las voces que históricamente han estado excluidas. Si permitimos que los sesgos machistas se sigan estableciendo en los sistemas que definen nuestro futuro, estaremos programando la exclusión de la mujer y su sometimiento desde la raíz.✳

REFERENCIAS

1. I'd blush if I could. Closing gender divides in digital skills through education. UNESCO. https://unesdoc.unesco.org/ark:/48223/pf0000367416/PDF/367416fre.pdf.multi.

2. Historia de la inteligencia artificial. La inteligencia artificial: nacimiento, aplicaciones y tendencias de futuro. https://www.iberdrola.com/innovacion/historia-inteligencia-artificial#:~:text=John%20McCarthy%20acu%C3%B1%C3%B3%20en%201956,de%20programaci%C3%B3n%20de%20IA%2C%20LISP.

3. Campesato, O. (2020). Artificial Intelligence, Machine Learning and Deep Learning. Dulles Virginia, Mercury Learning and Information LLC, 2020.

4. Protocolo para juzgar con perspectiva de género. Haciendo realidad el derecho a la igualdad. Suprema Corte de Justicia de la Nación. México, D.F., julio 2013.

5. Celaya, Alberto, L., y Campión, S. (2022). El reto de la alfabetización digital. De la sustitución a la transformación. Barcelona, editorial UOC, marzo 2022.

6. Reflexiones feministas para el desarrollo de Inteligencia Artificial. https://www.derechosdigitales.org/fair-2023/#A_modo_de_conclusion_fomentar_la_participacion_expandir_comunidades_construir_futuros.

SOBRE LA AUTORA



Mónica Jaimes Gaona Es especialista en Derecho Constitucional y Derechos Humanos. Con más de 20 años de trayectoria jurídica, Mónica Jaimes Gaona ha desarrollado una sólida carrera en el ámbito judicial y asegurador, consolidándose como experta en derecho constitucional, derechos humanos, amparo, compliance e inteligencia artificial. Egresada de la Universidad Tecnológica de México, cuenta con una maestría en Derecho Constitucional y Derechos Humanos por la Universidad Panamericana (en trámite de titulación), así como con formación internacional en Tribunales Constitucionales por la Universidad Pompeu Fabra (España) y un curso de Experto en Derecho e Inteligencia Artificial en la Universidad de Barcelona. Ha sido Secretaria de Estudio y Cuenta en el Tribunal Electoral del Poder Judicial de la Federación y, actualmente, en la Suprema Corte de Justicia de la Nación, donde ha elaborado proyectos de sentencia de alto impacto en temas de equidad de género, derechos indígenas y justicia administrativa. Su paso por el sector asegurador también destaca por el desarrollo de productos innovadores. Su trayectoria profesional se ha enfocado en la elaboración de criterios constitucionales con perspectiva de género y nuevas tecnologías. Apasionada por la intersección entre derecho e inteligencia artificial, Mónica impulsa la reflexión jurídica sobre neuroderechos, ética algorítmica y sesgos de género en sistemas automatizados, temas que también difunde activamente a través de su perfil profesional en LinkedIn.

© 2025 - Sociedad Mexicana de Inteligencia Artificial ISSN 2007-0691

ARTÍCULO ACEPTADO

De la ficción a la realidad: La Inteligencia Artificial a través de los ojos de Isaac Asimov

Oscar Ramos-Soto, J. Antonio Ruvalcaba, Diego Oliva, y Sandra E. Balderas-Mata

Introducción

Isaac Asimov (1920-1992), una de las figuras más icónicas de la literatura en ciencia ficción, no solo fue un escritor prolífico, sino también un visionario cuyas ideas han influido profundamente en la forma en que imaginamos el futuro. Sus obras, como la serie de "Yo, Robot", introdujeron conceptos que trascienden la literatura, marcando un antes y un después en la discusión sobre inteligencia artificial (IA) y ética tecnológica, anticipando debates que siguen siendo relevantes en la actualidad. Entre sus contribuciones más destacadas están las Tres Leyes de la Robótica, un conjunto de principios destinados a garantizar la coexistencia pacífica entre humanos y máquinas inteligentes.

Hoy, en la plena inclusión de la IA en la cotidianeidad, surge la pregunta: ¿cuán cerca estamos de alcanzar la visión de Asimov? *"Robbie extendió su brazo metálico, y la niña corrió hacia él con una risa que resonaba por toda la habitación. En ese momento, para Gloria, Robbie no era una máquina, sino un amigo."* Este pasaje del cuento "Robbie" [1] subraya la conexión emocional entre humanos y máquinas que Asimov imaginó. En la actualidad, asistentes virtuales como *Alexa* o *Siri*, aunque menos complejos, ilustran cómo las máquinas están integrándose a nuestra vida cotidiana.

En este artículo se explora el legado de las Tres Leyes, su relación con la IA moderna y los desafíos que enfrentamos al intentar convertir la ficción en realidad. Con ello se busca reflexionar sobre cómo la obra de Asimov sigue siendo relevante en el mundo tecnológico actual, al ser tanto un espejo de nuestro presente como una posible proyección hacia el futuro.

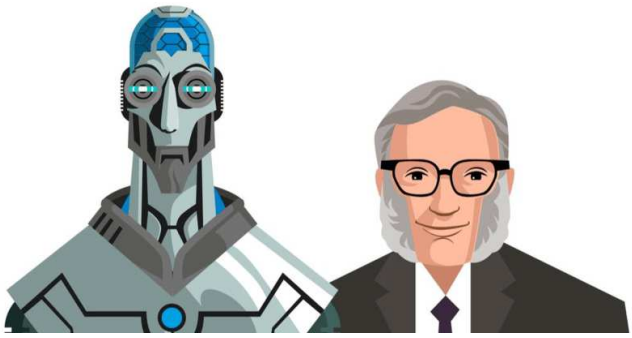


Figura 1. Robot e Isaac Asimov.
Imagen tomada de: `https://t.ly/9F7Vo`.

Las Tres Leyes de la Robótica

En 1942, años antes de la invención formal del transistor, pieza fundamental de la electrónica moderna y la informática digital, Isaac Asimov introdujo las Tres Leyes de la Robótica: un conjunto de principios simples pero profundos. Aunque concebidos más como un recurso literario para sus cuentos que como una guía para el futuro, han modelado el pensamiento sobre máquinas inteligentes durante décadas:

1. Un robot no debe dañar a un ser humano ni, por inacción, permitir que un ser humano sufra daño.

2. Un robot debe obedecer las órdenes dadas por los seres humanos, excepto si estas órdenes entran en conflicto con la Primera Ley.

3. Un robot debe proteger su propia existencia en la medida en que esta protección no entre en conflicto con la Primera o la Segunda Ley.

Estas leyes fueron propuestas como una solución ficcional a los posibles conflictos entre humanos y robots. En los relatos de Asimov, estas reglas no solo limitaban el comportamiento de los robots, sino que también servían como herramienta narrativa para explorar dilemas éticos y filosóficos. Por ejemplo, ¿qué sucede cuando dos leyes entran en conflicto? O, ¿cómo actúa un robot ante una situación moralmente ambigua? Estas preguntas han inspirado tanto a escritores como a científicos a reflexionar sobre los límites y posibilidades de la IA. *"El robot, atrapado en una paradoja lógica, permanecía inmóvil. Sus circuitos, programados para obedecer las leyes, no podían encontrar una salida."* Este fragmento de "El círculo vicioso" [1] describe a un robot enfrentado a una situación en la que dos posibles acciones, aparentemente iguales en términos de peligro para los humanos, lo conducen a un estado de inacción total. En el cuento, el robot se encuentra en un espacio donde dos amenazas para la seguridad humana están perfectamente equilibradas. Cualquier decisión que tome violaría de alguna forma la Primera Ley de la Robótica que le prohíbe causar daño a los humanos o permitirlo por inacción. Al no poder priorizar entre ambas opciones, el robot queda atrapado en un estado de inmovilidad, un fallo que ilustra cómo las leyes, aunque diseñadas para prevenir conflictos, pueden generar bloqueos insuperables cuando las circunstancias son ambiguas o equitativas.

© 2025 - Sociedad Mexicana de Inteligencia Artificial ISSN 2007-0691

Este dilema no solo revela la fragilidad de sistemas lógicos frente a situaciones complejas, sino que también anticipa problemas actuales en la programación de IA, como la incapacidad de muchos algoritmos para manejar dilemas éticos o conflictos de interés cuando las reglas que los gobiernan no ofrecen una respuesta clara.

En el cuento "El hombre bicentenario" [2], Asimov aborda un dilema ético sobre la perfección y la mejora humana a través de un robot llamado Andrew, que, a lo largo de dos siglos, lucha por ser reconocido como un ser humano. A medida que avanza, el robot no solo busca adaptar su cuerpo a la forma humana, sino también incorporar características propias de nuestra especie como la emoción y la capacidad de tomar decisiones éticas. En su proceso de evolución, Andrew se enfrenta a una encrucijada moral: ¿debe seguir sus principios como máquina basados en las Tres Leyes de la Robótica o debe ir más allá y tomar decisiones que lo conviertan en un ser autónomo con libertad para actuar según su propio juicio? Este tipo de disyuntivas éticas resuenan hoy en día con la creciente autonomía de los sistemas de IA que deben navegar por los dilemas de "hacer el bien" sin comprender completamente el contexto humano en el que se insertan. La historia de Asimov nos recuerda que la mejora o la autonomía tecnológica no es solo una cuestión de funcionalidad, sino también de las decisiones morales que las máquinas podrían enfrentar al interactuar con el mundo humano.

Avances en IA destacan límites y posibilidades frente a las visiones de Asimov.

La IA moderna: ¿un sueño Asimoviano?

En los úlitmos años, hemos sido testigos de avances sorprendentes en la IA. Sistemas como los *chatbots* avanzados (como *ChatGPT-4*) y los vehículos autónomos han transformado industrias y sociedades enteras. Desde la medicina hasta la movilidad urbana, la IA está revolucionando la forma en que vivimos y trabajamos. Sin embargo, a medida que ésta se vuelve más capaz, también surgen preguntas éticas y técnicas sobre su desarrollo y uso.

Aunque las Tres Leyes de la Robótica han inspirado marcos éticos y legales modernos, como las de la Unión Europea para una IA confiable y explicativa [3], su aplicación directa resulta impráctica. Este dilema se presenta de manera concreta en el cuento "La última pregunta" [4], donde un supercomputador llamado Multivac, a lo largo de milenios, se enfrenta a la pregunta final sobre cómo evitar la entropía del universo y salvar a la humanidad. Aunque no se trata de un robot en el sentido clásico, la narrativa de Asimov plantea preguntas sobre la capacidad de las máquinas para tomar decisiones de gran alcance y complejidad, con implicaciones morales profundas y ambiguas. A través de este relato, Asimov explora la idea de que las máquinas, por más poderosas que sean, podrían carecer de una comprensión completa de la moralidad humana, como es el caso de las IA actuales que enfrentan desafíos en cuanto a cómo interpretar y adaptarse a contextos éticos complejos.

Por otro lado, la diversidad de aplicaciones de la IA en el mundo moderno ha superado con creces la visión original de Asimov. Desde algoritmos que optimizan búsquedas en internet hasta generación de imágenes o video a partir de un enunciado (como la mostrada en la Figura 2), la IA ha demostrado ser un campo extremadamente amplio. Sin embargo, esta expansión también plantea desafíos para crear un marco ético que abarque todas sus posibilidades.



Figura 2. Imagen generada por IA[1] con el comando 'Dibujo realista del universo según la obra literaria de Isaac Asimov acerca de los robots y la IA'.

Los dilemas éticos de la IA actual

A diferencia de los robots de Asimov, que eran agentes claramente delimitados, la IA moderna a menudo opera de manera invisible a través algoritmos que toman decisiones en nombre de personas o empresas. Esto introduce

[1] Generada desde el sitio web www.bing.com/images/create/

nuevos dilemas éticos que Asimov no pudo prever dada la madurez tecnológica de sus tiempos:

- **Sesgos en la IA:** Los sistemas actuales pueden heredar prejuicios de los datos con los que son entrenados, perpetuando desigualdades en lugar de corregirlas. Este problema es especialmente preocupante en sectores como la justicia o el empleo, donde decisiones algorítmicas pueden tener impactos profundos en la vida de las personas.
- **Privacidad:** La recopilación masiva de datos para entrenar modelos plantea riesgos para la privacidad individual. En un mundo hiperconectado, proteger la información personal se ha convertido en un gran desafío ético y técnico difícil de afrontar.
- **Autonomía:** La creciente autonomía de los sistemas plantea preguntas sobre la responsabilidad legal y moral cuando estos fallan o causan daño. ¿Quién es responsable si un coche autónomo tiene un accidente, o bien, en un diagnóstico médico erróneo basado en IA? Estos son solo algunos ejemplos de las muchas incógnitas que surgen en este contexto.

En este panorama, las Tres Leyes ofrecen un punto de partida para la reflexión, pero también revelan sus limitaciones frente a la complejidad del mundo real. La interacción entre ética, tecnología y sociedad requiere un enfoque dinámico que trascienda los principios básicos de la robótica.

Si bien las Tres Leyes han inspirado marcos éticos, su implementación enfrenta limitaciones clave. Por ejemplo, la Primera Ley presupone la capacidad de un robot para identificar y evaluar 'daño', algo que los sistemas modernos aún no pueden hacer con precisión en contextos ambiguos. Un caso actual es el de los vehículos autónomos que deben decidir entre priorizar la seguridad del pasajero o la de peatones en situaciones de riesgo. Este dilema refleja la inadecuación de las leyes de Asimov para abordar decisiones éticas complejas.

Asimov vinculó ética y tecnología, anticipando dilemas actuales de la inteligencia artificial.

Isaac Asimov como visionario y guía ético

A pesar de sus limitaciones, las ideas de Asimov siguen siendo relevantes como marco filosófico. Su obra no solo inspiró avances tecnológicos, sino que también fomentó debates sobre la relación entre humanos y máquinas. Además, sus escritos nos recuerdan la importancia de abordar el desarrollo tecnológico con una visión responsable. La IA no es solo una herramienta, es una creación humana que refleja nuestros valores, miedos y aspiraciones. En este sentido, Asimov nos ofrece una guía para navegar por el desafiante terreno de la IA. Su insistencia en explorar este tipo de dilemas en sus relatos sirve como una advertencia sobre los peligros de desarrollar tecnología sin considerar sus implicaciones morales.

"Dije al robot que se perdiera. Fue una orden que no violaba la Segunda Ley, pero sí la Primera, porque estaba diseñada para causar un problema." Este ejemplo, tomado de "Pequeño robot perdido", resalta los riesgos asociados a la interpretación literal de las órdenes. En sistemas de IA modernos, esto se traduce en la necesidad de instrucciones precisas y contextuales para evitar errores con consecuencias graves.

Por otro lado, la diversidad de aplicaciones de la IA en el mundo moderno ha superado con creces la visión original de Asimov: la IA ha demostrado ser un campo extremadamente amplio.

Uno de los mayores desafíos de la IA actual es la existencia de sesgos en los datos de entrenamiento, como se evidenció en sistemas de reconocimiento facial que fallan al identificar con precisión rostros de minorías étnicas [5]. Este problema destaca una limitación crítica: los sistemas no comprenden las implicaciones éticas de sus decisiones. A diferencia de las leyes de Asimov, los sistemas modernos no tienen principios éticos preprogramados, sino que operan basados en datos, lo que aumenta el riesgo de decisiones no alineadas con valores humanos.

Conclusiones

Isaac Asimov imaginó un futuro donde la humanidad y las máquinas inteligentes coexistieran bajo principios éticos claros. Aunque ese futuro está lejos de materializarse en su totalidad, sus ideas continúan siendo una fuente de inspiración y reflexión en torno al desarrollo de la IA. Las Tres Leyes de la Robótica nos desafían a pensar en las implicaciones morales de nuestras creaciones tecnológicas, mientras que su obra nos recuerda que el progreso no sólo debe medirse en términos de eficiencia, sino también de humanidad.

En un mundo donde la IA avanza a pasos agigantados, las historias de Asimov nos ayudan a explorar las posibilidades y los peligros de nuestras invenciones. Su legado no radica únicamente en la cercanía de sus historias a la realidad, sino en su capacidad para inspirar a generaciones a imaginar y construir un futuro mejor. La comunidad científica debe trabajar en conjunto para

© 2025 - Sociedad Mexicana de Inteligencia Artificial ISSN 2007-0691

garantizar que los desarrollos en IA prioricen el bienestar humano y no solo la eficiencia tecnológica. Esto incluye la creación de sistemas explicables y responsables.

Su literatura nos deja con una pregunta que sigue resonando: ¿cómo podemos garantizar que la tecnología sirva al bienestar humano en lugar de ponerlo en peligro? A esto se suman otras inquietudes igual de profundas: ¿es posible programar la ética humana dentro de una máquina?, ¿deberíamos humanizar las computadoras o mantenerlas como herramientas claramente diferenciadas de nosotros?

Estas preguntas nos recuerdan que el verdadero reto no es solo técnico, sino profundamente humano. Este es el desafío que Asimov nos invita a afrontar, con la esperanza de que, al final, podamos navegar con éxito entre la ficción y la realidad.✷

Agradecimientos

Los autores agradecen a Nadia Elisa Balderas Mata por su apoyo en la revisión de gramática y estilo de este artículo.

REFERENCIAS

1. Asimov, I. (2014). Yo, Robot. España: EDHASA.
2. Asimov, I. (2013). El hombre bicentenario (Flash Relatos). España: Penguin Random House Grupo Editorial España.
3. European Commission. (2021). Ethics guidelines for trustworthy AI. https://digital-strategy.ec.europa.eu/en/library/ethics-guidelines-trustworthy-ai.
4. Asimov, I. (2013). La última pregunta (Flash Relatos). España: Penguin Random House Grupo Editorial España.
5. Acien, A., Morales, A., Vera-Rodriguez, R., Bartolome, I., y Fierrez, J. (2019). Measuring the gender and ethnicity bias in deep models for face recognition. En *Progress in Pattern Recognition, Image Analysis, Computer Vision, and Applications: 23rd Iberoamerican Congress*, CIARP 2018, Madrid, Spain, November 19-22, 2018, Proceedings 23 (pp. 584-593). Springer.

SOBRE LOS AUTORES



Oscar Ramos-Soto Obtuvo el título de Ingeniero en Comunicaciones y Electrónica por el Instituto Politécnico Nacional (IPN) en 2019. Posteriormente, obtuvo el grado Maestro por la Universidad de Guadalajara en el año 2021. Actualmente es estudiante del Doctorado en Ciencias de la Electrónica y Computación en el Centro Universitario de Ciencias Exactas e Ingeniería en la Universidad de Guadalajara. Sus intereses de investigación involucran áreas como la visión artificial, el procesamiento digital de imágenes biomédicas, el aprendizaje automático y las ciencias de la visión.



J. Antonio Ruvalcaba Obtuvo el grado de Doctor en Ciencias en Ingeniería Eléctrica con especialidad en Bioelectrónica en el Centro de Investigación y de Estudios Avanzados (CINVESTAV) del Instituto Politécnico Nacional (IPN) en el 2020. Actualmente es profesor adjunto en la Universidad Tecnológica de Durango en donde es catedrático de la carrera de mecatrónica. Su interés en investigación involucra áreas como bio-instrumentación, ciencia de datos, análisis de elemento finito, máquinas de aprendizaje y prótesis mioeléctricas.



Diego Oliva En el año 2015 obtuvo el grado de Doctor en Ingeniería Informática por la Universidad Complutense de Madrid (UCM) en España. Actualmente, es Profesor e Investigador adscrito al Centro Universitario de Ciencias Exactas e Ingenierías de la Universidad de Guadalajara. Desde el 2019 miembro del Sistema Nacional de Investigadoras e Investigadores (SNII) nivel 2. Sus líneas de investigación son cómputo evolutivo, algoritmos metaheurísticos, procesamiento de imágenes y aprendizaje automático.



Sandra E. Balderas Mata Obtuvo la licenciatura en Física por la Universidad Autónoma de San Luis Potosí. Posteriormente realizó sus estudios de posgrado, maestría y doctorado, en el Instituto Nacional de Astrofísica, Óptica y Electrónica (INAOE). Su investigación durante este período se centró en la aplicación de no linealidades en materiales. En 2010, durante su estancia postdoctoral en la Universidad de California en Davis, comenzó a trabajar en el área de óptica visual. Es miembro del Sistema Nacional de Investigadoras e Investigadores (SNII) nivel 1. Actualmente se desempeña como profesora de la Universidad de Guadalajara en el Departamento de Electro-fotónica.

© 2025 - Sociedad Mexicana de Inteligencia Artificial ISSN 2007-0691

ARTÍCULO ACEPTADO

¿Autonomía o Control? Caminos hacia una Inteligencia Artificial Responsable

J. Antonio García Macías y Antonio de Jesús García Chávez

¿Futuros Posibles?

"Lo siento, Dave, me temo que no puedo hacer eso" fue la siniestra frase con la que HAL 9000 declara su desobediencia ante el humano en el filme "2001: una odisea del espacio" dirigida por Stanley Kubrick (y adaptada a partir de la novela homónima de Arthur C. Clarke [10]). Siendo la computadora de a bordo de la nave espacial Discovery, HAL 9000 era la encargada de controlar todas las funciones vitales; tenía para ello capacidades de supercómputo y una inteligencia artificial (IA) muy avanzada. Por supuesto, el tema de la IA que toma el control y se vuelve en contra de los humanos es muy socorrido (e incluso trillado) en la literatura distópica de ciencia ficción. Películas y series muy populares, tales como Matrix, Terminator y otras giran en torno a alguna variación sobre este trama. La creciente popularidad del tema podría reflejar una inquietud colectiva frente al vertiginoso avance de tecnologías como la inteligencia artificial, cuyo rápido desarrollo genera temores sobre la posibilidad de perder el control sobre sus efectos. Así que no resulta tan ocioso preguntarse ¿qué tan factible es que un escenario de este tipo pueda hacerse realidad? ¿hay algo que podamos hacer para evitarlo?

La Automatización

Para que los sistemas inteligentes puedan tomar decisiones, es necesario que cuenten con cierto grado de autonomía, la cual puede clasificarse en distintos niveles, como se observa en las tecnologías actuales. En el primer nivel los humanos pueden controlar funciones tecnológicas, por ejemplo, hacer ajustes en una cámara digital o controlar el vuelo de un dron. En un segundo nivel, se pueden hacer ciertas programaciones pero el sistema no requiere de mucha asistencia humana, como sería el caso de usar una aspiradora de casa tipo Roomba. En un siguiente nivel están los sistemas semiautónomos, como la función de automanejo de algunos automóviles modernos que requieren intervención humana solamente en casos de emergencia. Ir más allá de estos niveles, donde los sistemas son completamente autónomos e inteligentes, requiere entrar en terrenos que en la actualidad generan intensos debates.

A poco tiempo de terminar el siglo pasado, cuando la IA como área de estudio atravesaba por un "segundo invierno" [1], ya se debatía cómo debían ser los sistemas inteligentes de la siguiente generación. Un punto álgido en los debates era precisamente el nivel de autonomía que estos deberían ser capaces de adquirir. Por una parte estaban quienes impulsaban una visión de "agentes de software" proactivos que podrían tomar la iniciativa para realizar ciertas tareas en pro del usuario, a quien conocerían muy bien por haber analizado previamente sus comportamientos. En contraparte estaban quienes defendían la postura de brindar a los usuarios altos niveles de comprensión y control sobre sus dispositivos habilitados con IA, de manera que se pudiera preservar la agencia humana. En el fondo del debate se encontraba entonces un tema de control, habiendo una tensión entre el nivel de control que se le daría a la máquina y al humano.

¿Estamos perdiendo el control?

Quizá primero se podría intentar en alguna medida acotar quienes estamos o no perdiendo el control. ¿Nos referimos a la humanidad en su conjunto? Pensar que todos los humanos podríamos controlar algo de manera colectiva es, a todas luces, absurdo. Entonces, probablemente no se trata de todos nosotros. ¿Nos referimos a los ingenieros de Silicon Valley? ¿A los inversionistas de Wall Street? ¿O a los políticos que en teoría deberían procurar nuestros intereses?

Lo segundo a considerar quizá sería el preguntarnos ¿qué entendemos por control? ¿Es como conducir un coche por una red de carreteras?, o tal vez sea más como conducir mientras el camino surge inesperadamente sin saber su final, mostrando por igual callejones sin salida e incluso ciclos de caminos que no llegan hacia ningún lado. Si es cierto que el humano es aquel que dirige la tecnología, cada giro y decisión que tomamos, cambia el terreno que tenemos que gobernar de forma inesperada [4].

Tendemos a imaginar que toda tecnología nace de un proceso racional y controlado, en el que cada elemento de su diseño es el resultado de una decisión humana intencionada. Pero en la práctica, esto rara vez es así. El desarrollo tecnológico está profundamente condicionado por los entornos en los que se crea: las computadoras utilizadas, las redes disponibles, las herramientas de software, las bibliotecas y los lenguajes de programación empleados. Por ende, la forma final de un sistema, así como su éxito, se determina en mayor medida por la cultura en la

[1] Se denomina inviernos de la IA a los periodos donde se redujo el interés y los fondos en torno a la investigación y desarrollo de IA. El primer invierno va de 1974 a 1980, mientras que el segundo va de 1987 a 1993.

© 2025 - Sociedad Mexicana de Inteligencia Artificial ISSN 2007-0691

cual este sistema pretende insertarse. Es esa cultura la que moldea, impulsa o limita lo que la tecnología llega a ser.

Consideremos por ejemplo ¿porque las regulaciones actuales logran tan poco en custodiar nuestra privacidad? Hemos normalizado la decisión binaria que se nos suele imponer entre "Aceptar los términos y condiciones" o bien "Entonces no te brindamos servicio". Lo que consideramos "bueno" no siempre basta [1]. Podríamos, quizás, apelar a las buenas intenciones detrás de los intentos por legislar sistemas que afectan y moldean nuestra experiencia humana. Pero reflexionemos un poco, ¿Esperamos que las nuevas normativas que intentan mitigar las noticias falsas o evitar que las insurrecciones sean instigadas por las redes sociales sean más eficaces?

Si la tecnología está moldeando la cultura al mismo tiempo que la cultura está dando forma a la tecnología, entonces estamos frente a un ciclo el cual se retroalimenta [2]. Y en ese tipo de ciclos, intervenir en cualquier punto puede alterar significativamente los resultados finales. Por ejemplo, podríamos tratar de prohibir que los sistemas de armas o las entidades financieras tomen decisiones de forma completamente autónoma. Sin embargo, ejemplos recientes muestran cómo la IA puede ser utilizada para influir en la toma de decisiones humanas de manera sutil pero poderosa. Por ejemplo, en el contexto de campañas políticas, el uso de algoritmos para personalizar contenidos en redes sociales ha permitido dirigir mensajes específicos a distintos grupos de votantes, amplificando ciertos discursos y afectando percepciones. De forma similar, en aplicaciones como Pokémon Go, los sistemas basados en datos y geolocalización han sido capaces de modificar el comportamiento de los usuarios en el espacio físico, incentivando desplazamientos y actividades según objetivos comerciales o estratégicos. Estos casos ilustran cómo la IA, más allá de su función técnica, puede tener un impacto directo en las decisiones cotidianas y colectivas. En consecuencia, incluir a un ser humano en el proceso no garantiza, necesariamente, un mayor control o una supervisión efectiva [6].

Volviendo a la pregunta inicial: ¿estamos perdiendo el control? La respuesta es "no". De hecho, nunca lo hemos tenido, y no se puede perder lo que nunca se ha tenido. Sin embargo, esto no significa que debamos rendirnos o dejar de intentar influir en la dirección que toman estas tecnologías. Resulta hoy más que nunca necesario un enfoque hacia la tecnología centrada en el ser humano. Esta iniciativa hace imperativo que intelectuales de todas las disciplinas den un paso al frente y se tomen en serio esta danza que como humanidad llevamos con la tecnología.

Tecnología Centrada en Humanos

En tiempos recientes, tras el gran auge que recobró la IA, se ha reavivado también el tema de la autonomía y el control. Los problemas tan publicitados con sistemas automatizados, tal como el que provocó que en dos ocasiones diferentes se estrellaran aviones Boeing 737 Max, han animado aún más a los defensores de la postura de limitar el control que se puede ceder a las máquinas. Los efectos adversos de los sistemas inteligentes automatizados pueden no llegar a ser fatales, pero sí bastante dañinos; este podría ser el caso de sistemas recomendadores que, con los sesgos algorítmicos inherentes que poseen, toman decisiones sobre quién obtiene un empleo, a quién se le otorga una hipoteca, quién puede salir de la cárcel bajo libertad condicional, o bien quién es proclive a cometer un crimen. Vemos como la imagen (ya convertida en algo común) del sistema de IA, representado por un robot que le da un firme apretón de manos a un humano para mostrar su empatía y cooperación, no refleja de forma muy fidedigna la realidad de estos casos.

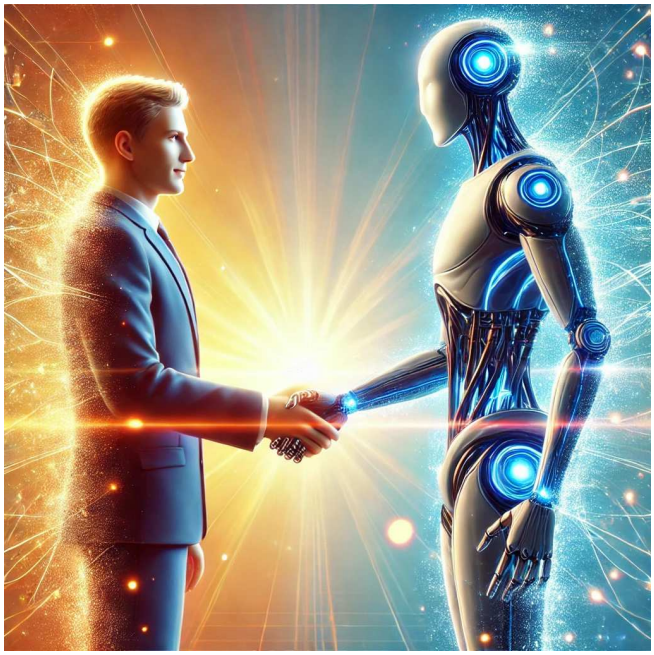


Figura 1. Imágen estereotipada de la idílica convivencia humano-IA. Creada con ChatGPT; observar que ha representado al humano como un hombre blanco, jóven y con vestimenta formal.

Un poco en forma reactiva, respondiendo a situaciones con efectos adversos como los recién señalados, se abre el campo de la IA centrada en los humanos [5] (*HCAI* por sus siglas en inglés). Se trata ahora de atender aspectos de las tecnologías que vayan alineados con los valores humanos como la no discriminación y la justicia. Aún más, los sistemas de IA deben diseñarse en torno a factores humanos considerando que resulten útiles, usables y comprensibles (o explicables). La IA explicable (conocida como *XAI*) busca alejarse del modelo de "caja negra" con el que se desarrollan actualmente los

© 2025 - Sociedad Mexicana de Inteligencia Artificial ISSN 2007-0691

sistemas de aprendizaje automático o machine learning; no basta con que un sistema diagnostique la probabilidad de algún tipo de cáncer, pues tendría que ir más allá y mostrar el proceso de razonamiento que lo llevó a realizar tal diagnóstico y permitir a los especialistas visualizar y controlar algunos parámetros para tener una mejor comprensión del caso. Otro tema que ha cobrado fuerza y que, al igual que la XAI tratan de dar mayor control a los humanos, es el de la IA Responsable; para ello podemos utilizar la definición de Agarwal [7] que la designa como aquella que permite el uso de tecnologías de IA alineadas con las expectativas del usuario y las leyes, normas y regulaciones sociales vigentes. Cuando se habla de responsabilidad (*accountability* en inglés) se refiere a casos como los de las aplicaciones militares donde la responsabilidad dentro de una cadena de mando es un valor fundamental; si bien hay argumentos sólidos a favor de la autonomía en los sistemas de defensa, ningún sistema de armas debería ser completamente autónomo.

El actual enfoque racionalista de la comunidad de IA sigue privilegiando sistemas autónomos que funcionen de manera confiable sin supervisión humana. Con el enfoque centrado en los humanos, son precisamente los humanos quienes tienen un control significativo de la tecnología y son responsables de los resultados de sus acciones; es decir, se trata de una automatización donde se puede anticipar lo que sucederá puesto que los humanos, no las máquinas, tienen el control. La tecnología no se crea sola, no va "evolucionando" en el sentido de cómo lo hacen los sistemas biológicos. Somos los humanos quienes vamos decidiendo de qué manera avanza la tecnología. Contamos con la capacidad de orientar el desarrollo tecnológico para evitar los escenarios distópicos imaginados por la ciencia ficción. Es nuestra responsabilidad guiar el desarrollo tecnológico por un camino de beneficio a la humanidad.

El desafío de la dignidad humana en la era de los sistemas autónomos

Es fundamental que cada individuo pueda controlar tanto los datos que lo representan como las decisiones que los sistemas autónomos toman por él. Como lo mencionamos con anterioridad, el legislar resulta insuficiente; lo que está en juego es la ética y la dignidad humana. Las opciones de "aceptar/no aceptar" no garantizan nuestra verdadera libertad de elección. ¿Qué pasa con nuestras preferencias personales y nuestros puntos de vista morales? Lo que una empresa encargada de desarrollar un sistema autónomo considere como justo, no necesariamente coincide con la ética de los usuarios [3].

Contamos con la capacidad de orientar el desarrollo tecnológico para evitar los escenarios distópicos imaginados por la ciencia ficción. ¿Cuántas veces hemos oído respuestas como "sí, esto podría ser posible, pero el sistema no lo permite"?; esto nos muestra lo necesario que resulta el contemplar el derecho de elección del usuario. Pasar por alto esta actitud en el ámbito de la tecnología autónoma, puede poner en grave peligro nuestros valores éticos personales. Por lo que resulta vital el preservar el valor de la dignidad humana en el contexto de la sociedad digital, entendida como el reconocimiento de que una persona es digna de respeto en su interacción con las tecnologías autónomas.

Se requiere que los sistemas autónomos se diseñen con la posibilidad de realizar interacciones aún más complejas con el usuario, en lo que respecta a las decisiones morales de los usuarios. Así pues, la capacidad de los sistemas autónomos para tomar decisiones no sólo debe ajustarse a la legislación, sino también a las preferencias morales de los usuarios (incluida la privacidad) cuando éstas se manifiesten. Es central el incentivar una visión arquitectónica del mundo digital que desvincule los sistemas autónomos de los usuarios como actores independientes. Por lo que es necesario capacitar a los usuarios para que realicen interacciones posiblemente complejas con estos sistemas [4], mediante protocolos fiables que consideren la posibilidad de elección de los usuarios; aun en aspectos éticos.

Conclusiones

Hoy en día contamos con herramientas que nos permiten definir con mayor precisión los límites del poder de los sistemas autónomos, especialmente en aquellas decisiones que impactan directamente a las personas. Debemos a toda costa incentivar el uso de aplicaciones con base en IA que cumplan con un número mínimo de criterios. Estos criterios deben garantizar la protección de los datos personales, el derecho de los individuos a decidir sobre asuntos que les afectan, y el acceso a explicaciones claras y transparentes sobre las decisiones generadas o sugeridas por dichos sistemas.

Consideremos uno de los más recientes modelos multimodales – con capacidades de manejar texto e imagenes – MedGemma (www.medgemma.org), modelo presentado por Google con enfoque en imágenes y textos médicos. Aunque se propone como herramienta complementaria para profesionales de la salud, aun no está claro si sus recomendaciones serán lo suficientemente transparentes y explicables, especialmente en escenarios donde una interpretación precisa resulta fundamental para el diagnóstico o tratamiento de pacientes.

Tampoco es claro si los usuarios finales, pacientes o profesionales tendrán la posibilidad de aceptar o rechazar recomendaciones médicas generadas con el apoyo de modelos como MedGemma, lo cual plantea una preocupación legítima sobre la preservación del derecho a decidir. En respuesta a estos desafíos, han surgido diversas iniciativas que buscan establecer directrices para el uso responsable de este tipo de sistemas en contextos al-

© 2025 - Sociedad Mexicana de Inteligencia Artificial ISSN 2007-0691

tamente sensibles, como la salud. Estas incluyen desde recomendaciones sobre prácticas responsables en la publicación de resultados [8] hasta marcos éticos y técnicos para orientar el desarrollo de la investigación científica en la era de los modelos de lenguaje avanzados [9].

Aunque el objetivo no es frenar por completo el avance de la IA (si bien, en algunos casos una pausa reflexiva podría ser beneficiosa), es necesario atender el desarrollo de estándares éticos y responsables más robustos. Si no tomamos esta tarea con la seriedad que merece, ya sea por conveniencia, presiones económicas o tensiones políticas, corremos el riesgo de ceder aún más control sobre estos sistemas. Y con ello, podríamos comprometer no solo la autonomía del usuario, sino también valores fundamentales como la dignidad humana.✷

REFERENCIAS

1. Green, B. (2019). Good isn't good enough. En *Proceedings of the AI for Social Good workshop at NeurIPS*, 17, 1-7.
2. Hendrycks, D. (2023). Natural selection favors AIs over humans. arXiv preprint arXiv:2303.16200.
3. Inverardi, P. (2022). The challenge of human dignity in the era of autonomous systems. *Perspectives on digital humanism*, 25-29.
4. Lee, E. A. (2020). The coevolution: The entwined futures of humans and machines. MIT Press.
5. Shneiderman, B. (2021). Human-Centered AI. *Issues in Science and Technology*, 37(2), 56–61.
6. Werthner, H., Prem, E., Lee, E. A., y Ghezzi, C. (2022). Perspectives on digital humanism. *Springer Nature*, p. 32.
7. Agarwal, S., y Mishra, S. (2021). Responsible AI. *Springer*.
8. Partnership on AI. (2021). Managing the Risks of AI Research: Six Recommendations for Responsible Publication. `https://www.partnershiponai.org/responsible-publication-recommendations/`.
9. Birhane, A., Kasirzadeh, A., Leslie, D., y Wachter, S. (2023). Science in the age of large language models. En *Nature Reviews Physics*, 5(5), 277–280.
10. Clarke, A.C. (1968). 2001: A space odyssey. Penguin Books.

SOBRE LOS AUTORES



J. Antonio García Macías. Obtuvo el grado de Doctor del Institut National Polytechnique de Grenoble (INPG), Francia. Desde el 2002 es investigador titular del Centro de Investigación Científica y Educación Superior de Ensenada (CICESE), Baja California. Sus intereses de investigación están en la intersección de tecnología y sociedad, con temas como ciencias sociales computacionales, cómputo ubicuo, sistemas inteligentes para aplicaciones sociales.



Antonio de Jesús García Chávez Recibió el grado de Maestría en Ciencias en Ciencias Computacionales por el Centro de Investigación Científica y de Educación Superior de Ensenada (CICESE), Baja California México, en 2018. Actualmente es candidato a doctor en Ciencias de la Computación por la Universidad Autónoma de Baja California (UABC), México. Sus intereses de investigación incluyen el procesamiento del lenguaje natural, el aprendizaje automático, la recuperación de información y la inteligencia artificial generativa.

© 2025 - Sociedad Mexicana de Inteligencia Artificial ISSN 2007-0691

ARTÍCULO ACEPTADO

Emociones Dirigidas: Manipulación Inteligente a través de la Segmentación

Angel Gaspar, Luis Pérez Gaspar, Diego Oliva e Itzel Aranguren-Navarro

Introducción

En la actualidad, el uso de la inteligencia artificial (IA) en las plataformas digitales ha revolucionado la recopilación y el análisis de grandes cantidades de datos. Las plataformas digitales actuales rastrean cada acción del usuario, incluidos los clics, el tiempo de permanencia y la participación en los contenidos, mismos que luego se procesan para refinar la segmentación de la audiencia y ofrecer recomendaciones de contenido a medida. Sin embargo, esta segmentación de audiencias trasciende la simple recomendación de productos o servicios. También puede emplearse para influir en las emociones y decisiones de los usuarios, con implicaciones significativas en su comportamiento y percepción de la realidad [1].

A través de técnicas de aprendizaje automático, se analiza el comportamiento y las preferencias de distintos grupos de personas, permitiendo a empresas y políticos identificar qué tipo de mensajes generan mayor impacto. A su vez, se puede ajustar el tono y la carga emocional de los contenidos para captar la atención, reforzar creencias o inducir estados de ánimo específicos [2]. Este tipo de manipulación emocional plantea dilemas éticos, que exigen una reflexión sobre los límites de la influencia y la responsabilidad del uso de la IA como medio para la personalización del contenido.

En particular, es crucial considerar hasta qué punto las personas están expuestas a influencias sutiles, y si realmente pueden ejercer su autonomía en la toma de decisiones en un entorno donde la información se diseña de forma estratégica para generar reacciones predeterminadas. Por ello, resulta necesario no solo establecer principios claros para el diseño y uso de algoritmos, sino también garantizar que estos sigan los principios de una IA responsable. Esto implica desarrollar marcos éticos y normativos que promuevan la transparencia, la equidad y la explicabilidad de los sistemas, además de definir quién asume la responsabilidad por sus impactos, tanto positivos como negativos.

En este artículo, se explora cómo funciona la segmentación de usuarios con fines emocionales basado en técnicas de agrupación y aprendizaje profundo. Así como las implicaciones de esta práctica en la toma de decisiones, los riesgos asociados con una sociedad cada vez más influenciada por algoritmos, y el papel fundamental de la ética y la responsabilidad en el diseño de sistemas de IA.

Segmentación de usuarios basada en Agrupación

La segmentación de usuarios es una herramienta fundamental ampliamente usada por las empresas digitales para optimizar la distribución de recursos y personalizar la experiencia de las diferente categorías de usuarios. La agrupación, o *clustering* por su nombre en inglés, es una técnica de aprendizaje no supervisado, que se utiliza para descubrir grupos de elementos con comportamientos o características similares, a partir de un conjunto de datos sin etiquetar. La agrupación permite identificar patrones ocultos en grandes conjuntos de datos, y son fundamentales en campos como la personalización de anuncios, el análisis de comportamiento en redes sociales y la segmentación de clientes en mercadotecnia.

Debido a la gran cantidad de datos con las que actualmente las empresas o plataformas digitales cuentan, uno de los principales casos de usos del *clustering* es la agrupación de usuarios con base a un criterio de segmentación compartido para extraer información relevante, y así poder personalizar servicios, mejorar la experiencia del usuario o diseñar estrategias de mercadotecnia altamente dirigidas [3].

Criterios de Segmentación

Los criterios de segmentación de usuarios son atributos o características empleados por las empresas para establecer y delimitar su segmento objetivo. La selección de los criterios de segmentación debe estar alineado con los objetivos específicos y la lógica comercial de la empresa. Actualmente, existen diversos criterios de segmentación; sin embargo, los más usados son [4]:

- ***Demográfico***: Son características generales como la edad, género, ocupación, nivel de estudios e ingresos, etc.
- ***Geográfico***: Son atributos como país, región, tamaño de población, etc.
- ***Psicográfico***: Integra aspectos relacionados con las actitudes, creencias o características de la personalidad de los usuarios.
- ***Conductual***: Se basa en el análisis del comportamiento histórico del usuario, por ejemplo, la frecuencia de uso, el nivel de interacción con el servicio, hábitos de compra, etc.

© 2025 - Sociedad Mexicana de Inteligencia Artificial ISSN 2007-0691

Cuando nuestras emociones son guiadas sin que lo notemos, la libertad de decidir se convierte en una ilusión.

Agrupación de usuarios

El proceso de la agrupación de usuarios generalmente consta de 3 etapas clave:

- ***Primera etapa***: Recolección y preparación de datos relevantes con base a los objetivos específicos del análisis.
- ***Segunda etapa***: Selección de criterios de segmentación que podrían incluir atributos demográficos, geográficos, psicográficos o conductuales.
- ***Tercera etapa***: Selección y aplicación del algoritmo de *clustering*.

La Tabla 1 muestra los algoritmos de *clustering* más usados por desarrolladores a nivel industrial y científicos. Al aplicar *clustering* se obtiene como resultado segmentos claramente definidos que permiten a las empresas comprender mejor las características de cada grupo y tomar decisiones estratégicas específicas. La Figura 1 muestra cómo se pueden agrupar usuarios según sus gustos cinematográficos utilizando datos reales del conjunto de datos MovieLens [5]. A partir de las calificaciones que cada usuario realizó a las películas, se identificaron sus géneros favoritos y se agruparon en perfiles similares usando el algoritmo de K-Means. El análisis reveló cuatro grandes tipos de espectadores: usuarios que prefieren historias dramáticas y románticas, los que disfrutan de la acción y la ciencia ficción, los que eligen películas animadas o infantiles, y los que prefieren la comedia.

Tabla 1. Algoritmos más usados de *clustering*

Método	Descripción
K-Means	Optimiza la distancia intra-clúster utilizando centroides. Es simple, eficiente y adecuado para grandes volúmenes de datos.
***Clustering* Jerárquico**	Construye agrupaciones mediante enfoques acumulativos. No requiere predefinir el número de clústeres.
DBSCAN	Detecta clústeres basados en densidad local. Es robusto ante el ruido en los datos.

Esta segmentación no solo ayuda a entender mejor los hábitos de consumo de los usuarios, sino que también puede ser útil para recomendaciones personalizadas de películas o estrategias de marketing altamente efectivas y dirigidas. Sin embargo, cuando estas estrategias de segmentación se aplican más allá del ámbito comercial, comienzan a cruzar límites éticos importantes. El mismo conocimiento que permite personalizar contenidos también puede ser utilizado para influir en el estado emocional de los usuarios y modificar su comportamiento, abriendo paso a lo que se conoce como manipulación emocional.

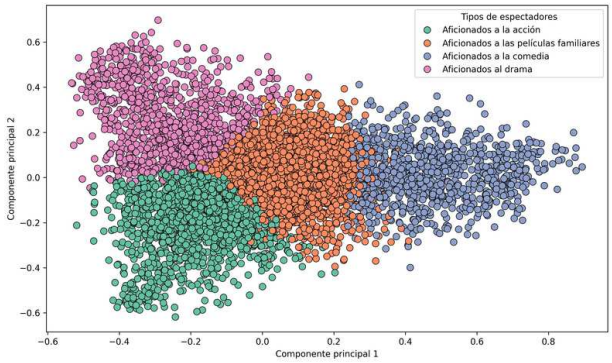


Figura 1. Segmentación de usuarios según su género cinematográfico preferido.

Manipulación emocional

La personalización de contenidos mediante IA no solo busca adaptar información a intereses de los usuarios, sino que también puede explotar sus emociones y vulnerabilidades psicológicas. Mediante redes neuronales de aprendizaje profundo y técnicas de procesamiento de lenguaje natural, es posible analizar las expresiones, el tono y hasta el estado de ánimo implícito en las interacciones de una persona para luego moldear los mensajes de forma que provoquen una respuesta emocional deseada.

Como se ilustra en la Figura 2, los algoritmos de visión por computadora son capaces de analizar microexpresiones faciales captadas por la cámara del dispositivo, así como patrones de atención mediante técnicas de comportamiento ocular, para inferir el estado emocional o el nivel de interés del usuario. Esta retroalimentación permite que sistemas publicitarios seleccionen, en cuestión de milisegundos, el contenido más persuasivo para ese contexto específico (por ejemplo, mostrar una oferta acompañada de imágenes positivas si el usuario aparenta estar desanimado, o viceversa) [6].

Del mismo modo, los modelos de lenguaje natural pueden adaptar el tono y las palabras clave de un texto publicitario a la personalidad del usuario; un usuario con rasgos introvertidos puede recibir un mensaje más tranquilo y empático, mientras que a un usuario extrovertido se le presenta un lenguaje más energético y excitante. Estas estrategias de publicidad apoyadas en IA, ajustan las

© 2025 - Sociedad Mexicana de Inteligencia Artificial ISSN 2007-0691

imágenes y lenguaje para maximizar el impacto emocional en los usuarios.

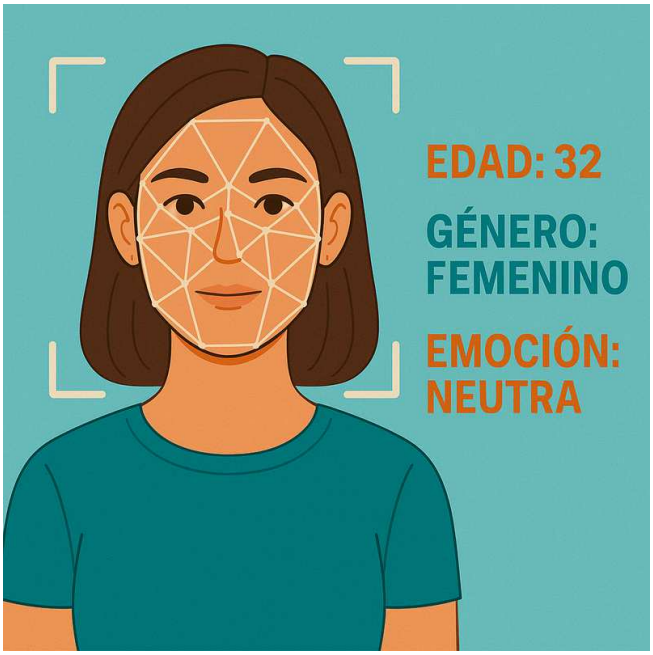


Figura 2. Análisis de visión por computadora.

La manipulación emocional es aprovechada con las principales plataformas de redes sociales, ya que crean un perfil psicológico y conductual del usuario. La IA a lo largo del tiempo construye un perfil emocional del usuario a partir de sus datos, sus reacciones pasadas (como "Me gusta", "No me gusta"), el sentimiento de los textos que escribe o comparte y las horas de actividad en la plataforma. Con este perfil, es fácil detectar estados de vulnerabilidad, por ejemplo, identificar usuarios que muestran señales de depresión, ansiedad, inseguridad o impulsividad. Con toda esta información es posible segmentar a los usuarios para enviarles contenidos que aprovechen sus emociones. La creación de estos perfiles psicológicos y la manipulación por medio de las plataformas como Instagram, YouTube o TikTok ha crecido en los últimos años. Éstas plataformas afinan constantemente sus contenidos para generar mayores ventas con su publicidad, además de mantener al usuario consumiendo el contenido.

En el artículo [7] se examinó cómo las redes neuronales convolucionales (CNN) pueden ser utilizadas para analizar y predecir el interés de los consumidores. Se utilizaron 58 "consumidores" con edades comprendidas entre los 18 y 70 años, se utilizó el conjunto de datos NeuroBioSense. A cada participante se le mostraron anuncios de 3 sectores: Automóviles y tecnología, Alimentos y mercados y Cosmética y moda. Los participantes fueron grabados en video mientras observaban los anuncios. Posteriormente, cada uno clasificaba si el anuncio le resultaba "interesante" o "no interesante". Este etiquetado sirvió para entrenar las CNNs. El estudio evaluó dos arquitecturas de CNNs, Xception y ResNet-50, para clasificar el compromiso del consumidor e identificar emociones correlacionadas. Los resultados arrojaron que ambas CNNs pueden distinguir efectivamente entre respuestas "interesante" y "no interesante"; con Xception se logró una mayor precisión y *recall* en comparación con ResNet-50. La felicidad se identificó como un indicador fuerte de interés del consumidor, mientras que el disgusto y el miedo se asociaron con desinterés.

Las emociones ya no se explotan en masa, sino en microaudiencias perfectamente perfiladas.

Otro tema donde la manipulación emocional esta tomando auge es en el contexto educativo. En un trabajo reciente se investigó cómo adaptar el contenido de un programa de aprendizaje asistido por computadora según los intereses extraescolares de estudiantes de secundaria [8]. En este estudio se llevó a cabo un experimento con 1449 estudiantes de 7° y 8° grado (equivalentes a la educación secundaria) en 31 escuelas de Flandes (Bélgica). El programa trataba sobre educación financiera, específicamente sobre el rol del gobierno y los impuestos. En este trabajo se observó que la adaptación de contenido escolar a intereses personales no perjudicó el aprendizaje del grupo en conjunto pero tampoco produjo mejoras significativas en el rendimiento o la motivación en promedio. Como vemos en [8], la manipulación a través de la enseñanza por computadora debe de ser diseñada de forma más cuidadosa si se quiere tener un impacto positivo. El impacto positivo sería que, a largo plazo, el estudiante se mantenga motivado en seguir estudiando cierto tema "no tan interesante" al mostrarle información de agrado para él de forma alternada y así evitar la monotonía.

A pesar de los avances significativos en las CNNs y su capacidad para interpretar señales no verbales, todavía existen preocupaciones sobre la precisión y la ética en la implementación de estos sistemas en la práctica comercial. La necesidad de asegurar la privacidad de los datos y la transparencia en la toma de decisiones sigue siendo una prioridad para el desarrollo futuro de estas tecnologías. Vemos que la manipulación emocional se puede llevar más lejos añadiendo la integración de más señales fisiológicas (por ejemplo, el ritmo cardíaco) y análisis

© 2025 - Sociedad Mexicana de Inteligencia Artificial ISSN 2007-0691

de voz para complementar la lectura de microexpresiones [6]. Hoy en día, no es raro saber que las aplicaciones en los dispositivos electrónicos utilizan sus cámaras y micrófonos integrados para, de una manera, "espiar" las conversaciones y sugerir anuncios personalizados [9]. Compañías grandes como Meta han negado siempre esta acción [10].

Este tipo de prácticas, que si bien pueden parecer especulativas, han generado creciente preocupación sobre la privacidad y el consentimiento informado en entornos digitales. A medida que las tecnologías avanzan y se vuelven más invasivas, se vuelve imprescindible discutir los marcos regulatorios y éticos que deben guiar su desarrollo y aplicación. En este contexto, la siguiente sección explora los esfuerzos internacionales y nacionales por establecer principios y normativas que garanticen un uso responsable y transparente de la IA.

Regulaciones internacionales y expectativas en México

En años recientes, se han establecido marcos éticos para guiar la aplicación responsable de la IA en distintos paises y organizaciones internacionales. En 2021 la UNESCO aprobó una Recomendación sobre la Ética de la IA que enfatiza valores como la transparencia, la privacidad, la no discriminación y la supervisión humana de los sistemas algorítmicos. En Europa, antes de la legislación, un grupo de alto nivel publicó en 2019 las Directrices Éticas para una IA confiable, donde se subraya la necesidad de "acción y supervisión humana" en algoritmos que afectan decisiones personales. Grandes empresas tecnológicas, bajo la presión pública, también han empezado a autorregularse; por ejemplo, Google implementó políticas para prohibir anuncios políticos intencionadamente engañosos y limitó la segmentación electoral a criterios generales. Meta ha lanzado informes de Transparencia Publicitaria donde terceros pueden ver qué segmentos fueron utilizados en anuncios políticos o sociales. Organismos internacionales como la Organización para la Cooperación y el Desarrollo Económicos (OCDE) han propuesto directrices para el uso confiable de la IA, asimismo, múltiples países han participado en foros como el Grupo de los Siete (G7) o la ONU. Estas leyes y medidas proponen equilibrar el uso de la IA dentro de límites aceptables que respeten la autonomía, la diversidad informativa y el bienestar de la sociedad.

En la actualidad, en México el marco regulatorio para la IA está en proceso de consolidación mediante diversas iniciativas legislativas. Una de las propuestas más destacadas es la Ley Federal para el Desarrollo Ético, Soberano e Inclusivo de la IA, presentada en abril de 2025 por legisladores del partido Morena y el Partido Verde. Esta iniciativa introduce un sistema de clasificación por niveles de riesgo (mínimo, limitado, alto y prohibido) para los sistemas de IA, prohibe prácticas como la vigilancia biométrica masiva sin orden judicial y la manipulación psicológica. Estas propuestas representan un paso importante hacia la construcción de un marco normativo que proteja los derechos de los ciudadanos frente a los riesgos asociados al uso de la IA. No obstante, su efectividad dependerá no solo de su implementación, sino también de su capacidad para adaptarse al ritmo acelerado del desarrollo tecnológico.

Conclusiones

A lo largo de este artículo se ha evidenciado cómo el marketing ha atravesado una transformación profunda, marcada por la incorporación de sistemas de IA que permiten segmentar emocionalmente a los usuarios e influir en sus decisiones. Este avance ha permitido a las empresas recopilar y analizar grandes volúmenes de datos sobre el comportamiento de los consumidores, generando oportunidades para implementar estrategias de marketing más personalizadas. Entre las innovaciones más recientes, el uso de redes neuronales convolucionales para el reconocimiento de expresiones faciales ha facilitado la identificación de emociones y niveles de interés de las personas ante estímulos como anuncios publicitarios.

La manipulación emocional no siempre debe de tener un impacto negativo para el usuario; en el ámbito educativo, el rol del profesor vería una mejora dado a que su intervención disminuiría y podría apoyar a estudiantes con diferentes capacidades intelectuales. Las plataformas al usar la IA proporcionarían una experiencia casi autónoma en el aprendizaje. En este caso, la manipulación emocional se vería desde un punto de vista educativo y no lucrativo, pero también debería haber un equilibrio en la personalización para potenciar los resultados positivos.✷

REFERENCIAS

1. Beauvisage, T., Beuscart, J.-S., Coavoux, S., Mellet, K. (2023). How online advertising targets consumers: The uses of categories and algorithmic tools by audience planners. *New Media & Society*, 26(10), 6098-6119.
2. Manoharan, A. (2024). Enhancing audience engagement through AI-powered social media automation. *World Journal of Advanced Engineering Technology and Sciences*, 11(2), 150-157.
3. Ayodele E., Sodeinde V. (2024). Customer Segmentation Using the K-Means Clustering Algorithm. *Ilaro Journal of Science and Technology (IJST)*, 4.
4. Camilleri M.A. (2018). Market Segmentation, Targeting and Positioning" *Travel Marketing, Tourism Economics and the Airline Product*, Chapter 4, pp. 69–83. Springer, Cham, Switzerland.
5. Maxwell, F., Konstan, J. (2015). The MovieLens Datasets: History and Context. *ACM Transactions on Interactive Intelligent Systems (TiiS)* 5, 4(19), 19.
6. Alipour, P., Gallegos, E. E., Sridhar, S. (2024). AI-Driven Marketing Personalization: Deploying Convolutional Neural Networks to Decode Consumer Behavior. *International Journal of Human–Computer Interaction*, 1-19.

© 2025 - Sociedad Mexicana de Inteligencia Artificial ISSN 2007-0691

7. Bakir, V., Laffer, A.. , McStay, A., Miranda, D., Urquhart, L. (2024). On manipulation by emotional AI: UK adults' views and governance implications. *Frontiers in Sociology*, 9, 2024.

8. Iterbeke, K., Schelfhout, W., De Witte, K. (2022). The role of students' interests during computer-assisted learning: A field experiment. *Computers in Human Behavior*, 130, 2022

9. Rahman-Jones, I. (2025). Apple to pay 95m USD to settle Siri 'listening' lawsuit. `https://www.bbc.com/news/articles/cr4rvr495rgo`, 2025.

10. Kleinman, Z. (2017). Facebook denies 'listening' to conversations. `https://www.bbc.com/news/technology-41776215`, 2017.

SOBRE LOS AUTORES



Angel Gaspar Obtuvo el grado de Ingeniero Electrónico especializado en sistemas digitales por el Instituto Tecnológico de Veracruz en el año 2016. Posteriormente se graduó de la Universidad de Guadalajara con el grado de Maestro en Ciencias en Ingeniería en Electrónica y Computación en el año 2021; donde realizó diversas publicaciones en el área del aprendizaje profundo. Actualmente es estudiante de Doctorado en Ciencias en Inteligencia Computacional. Sus principales áreas de interés son visión por computadora e inteligencia artificial.



Luis Pérez Gaspar Obtuvo el grado de Ingeniero Electrónico por el Instituto Tecnológico de Veracruz en el año 2012. Posteriormente obtuvo el grado de Maestro en Ciencias en Robótica en el año 2015; ha realizado varias publicaciones en el área de visión por computadora y reconocimiento de emociones en voz. Actualmente es estudiante de Doctorado en Ciencias en Inteligencia Computacional en la Universidad de Guadalajara.



Diego Oliva En el año 2015 obtuvo el grado de Doctor en Ingeniería Informática por la Universidad Complutense de Madrid (UCM) en España. Actualmente, es Profesor e Investigador adscrito al Centro Universitario de Ciencias Exactas e Ingenierías de la Universidad de Guadalajara. Desde el 2019 Miembro del Sistema Nacional de Investigadoras e Investigadores (SNII) nivel 2 . Sus líneas de investigación son *cómputo* evolutivo, algoritmos metaheurísticos, procesamiento de imágenes y aprendizaje automático.



Itzel Aranguren Obtuvo el grado de Doctor en Electrónica y Computación por la Universidad de Guadalajara en el año 2022. En la actualidad, es Profesora e Investigadora adscrita al Centro Universitario de Ciencias Exactas e Ingenierías de la Universidad de Guadalajara. Es miembro del Sistema Nacional de Investigadoras e Investigadores (SNII) nivel 1 desde el año 2023. Sus intereses de investigación incluyen visión por computadora, procesamiento de imágenes, algoritmos metaheurísticos e inteligencia artificial.

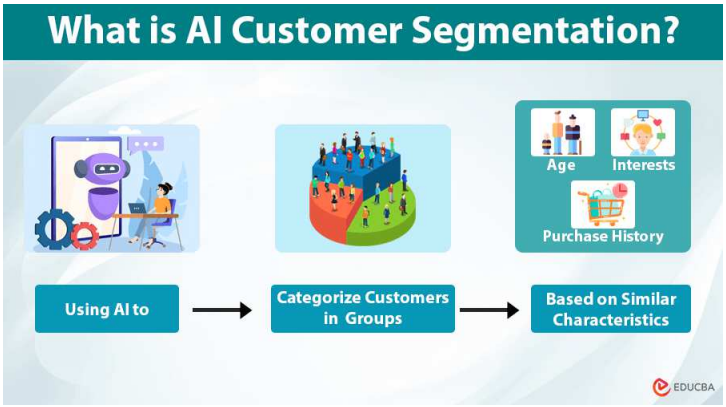


Imagen tomada de: `https://www.educba.com/ai-customer-segmentation/`.

© 2025 - Sociedad Mexicana de Inteligencia Artificial ISSN 2007-0691

ARTÍCULO ACEPTADO

Sesgo en los datos y los prejuicios en algoritmos de aprendizaje automático

Daniel Sánchez Ruiz, Ricardo Ramos Aguilar, Eric Ramos Aguilar y Jesús García Ramírez

Resumen

Este artículo analiza el problema del sesgo en los datos utilizados en algoritmos de aprendizaje automático y su impacto en aplicaciones del mundo real. Se describen distintos tipos de sesgo, como el sesgo de muestreo, de medición y de exclusión, explicando cómo afectan la equidad de los sistemas predictivos. Se presentan métodos de mitigación del sesgo, incluyendo técnicas de balanceo de datos y ajustes en los modelos, junto con una discusión sobre la importancia de la ética y la regulación en la implementación de estos algoritmos en entornos sensibles. Finalmente, el artículo enfatiza la necesidad de abordar el sesgo en los datos para garantizar la justicia en los sistemas de aprendizaje automático, proponiendo recomendaciones para un futuro más equitativo.

Introducción al aprendizaje automático y los datos

Los datos son mediciones de algunas características de una persona o una cosa. En un sistema de información, los datos son hechos dados a partir de los cuales se pueden inferir otros, un mensaje que no evoluciona, el portador de la información. De manera computacional, los podemos ver como una colección de texto, números y símbolos sin significado. Por lo tanto, los datos deben ser procesados o situados en un contexto antes de que puedan adquirir significado.

El aprendizaje automático (*Machine Learning*, ML) es una rama de la inteligencia artificial que permite a los sistemas aprender automáticamente de datos y mejorar su desempeño sin necesidad de ser programados explícitamente; esto sucede, como en los seres humanos, bajo un aprendizaje o entrenamiento continuo, que es la ejecución de alguna tarea en específico, y los datos forman parte de la experiencia. Los algoritmos de ML pueden ser supervisados, no supervisados o de refuerzo, dependiendo de si el modelo aprende de datos etiquetados o que tienen asignada una clase, busca patrones no etiquetados o aprende a partir de la interacción con un entorno [1].

El creciente uso de dispositivos conectados y la digitalización de procesos permite a las empresas, gobiernos y personas almacenar y analizar grandes cantidades de información, con el propósito de mejorar la eficiencia, personalizar servicios y optimizar recursos. La generación de datos ha alcanzado niveles sin precedentes debido a la digitalización de prácticamente todos los aspectos de la vida cotidiana. Los datos pueden estar disponibles en diversas formas: tablas estructuradas, tablas no estructuradas, imágenes, archivos de audio, videos, etc. Una máquina no puede entender directamente el texto libre, el video o la imagen; estos datos comúnmente están en bruto y no pueden ser introducidos directamente a un modelo de aprendizaje automático, tienen que llevar un preprocesamiento, el cual es el primer paso en el aprendizaje automático, en donde los datos se transforman o codifican para que puedan estar en un estado que permita a la máquina analizarlos o procesarlos rápidamente [6].

El éxito de los algoritmos de ML depende de la calidad y cantidad de los datos utilizados en su entrenamiento. A lo largo del ciclo de vida de los datos, estos se recolectan, procesan y luego se aplican para construir modelos predictivos. Sin embargo, pueden surgir problemas cuando los datos recolectados no representan de manera justa o precisa la realidad [2].

Origen de los sesgos en los datos

El sesgo de datos ocurre cuando la información que se usa para entrenar modelos de inteligencia artificial o de aprendizaje automático no representa de manera justa o precisa a la realidad. Esto puede deberse a datos incompletos, erróneos o mal seleccionados, lo que lleva a que los sistemas tomen decisiones equivocadas o injustas. En pocas palabras, es un tipo de error sistemático que se filtra en los algoritmos y afecta sus resultados. El sesgo en los datos puede introducirse en diferentes etapas del ciclo de vida de estos. Desde la recolección de datos hasta su procesamiento, los prejuicios humanos pueden influir en los resultados, lo que lleva a la creación de modelos que perpetúan o amplifican esos sesgos [1].

Uno de los factores más críticos es la selección de los datos, ya que estos pueden no representar correctamente a la población objetivo. Cuando la muestra es insuficiente o está desbalanceada, ciertos grupos pueden estar subrepresentados o sobrerrepresentados, lo que lleva a decisiones erróneas por parte del modelo [2]. Este sesgo puede ser sutil, pero sus consecuencias pueden amplificarse cuando los modelos son aplicados en contextos que afectan directamente a personas y comunidades, perpetuando desigualdades o generando discriminación involuntaria, respectivamente (Ver Figura 1).

Esto sucede comúnmente en estudios donde se utilizan datos históricos que reflejan desigualdades estruc-

© 2025 - Sociedad Mexicana de Inteligencia Artificial ISSN 2007-0691

turales pasadas. Un ejemplo concreto es el uso de datos de registros policiales para desarrollar sistemas de predicción de criminalidad. Si los datos reflejan prácticas policiales sesgadas o se enfocan en ciertos vecindarios, el modelo puede perpetuar patrones de vigilancia desproporcionada en comunidades minoritarias, reforzando la sobrevigilancia en áreas con antecedentes de arrestos altos, esto sin considerar otros factores sociales y económicos.



Figura 1. Representación gráfica sobre el efecto del sesgo en los datos.

Durante el procesamiento de datos, decisiones sobre cómo manejar valores faltantes o qué atributos descartar pueden también introducir sesgos. Por ejemplo, en el caso de algoritmos de contratación, como el modelo de selección de personal de Amazon desarrollado en 2014, los datos de entrenamiento incluían información de empleados actuales, entre los cuales existía una sobrerrepresentación de hombres en puestos de tecnología. Esto llevó al modelo a favorecer los perfiles masculinos en la selección, ya que el algoritmo asociaba la contratación exitosa con el género masculino y, de esta manera, penalizaba términos asociados a mujeres en los currículums [5].

Aun cuando un modelo ha sido entrenado y sus resultados se presentan como precisos, la interpretación de estos resultados puede estar sesgada si no se considera el contexto completo. Por ejemplo, en el sistema de evaluación COMPAS, usado en EE. UU. para predecir la reincidencia criminal, estudios encontraron que este sistema presentaba un sesgo contra personas afroamericanas, quienes recibían puntuaciones de riesgo más altas en comparación con personas de otros grupos con características similares. Esto sucedía porque el modelo no consideraba factores como los sesgos raciales históricos en las detenciones, y se basaba en datos que reflejaban de forma parcial la realidad [4].

Además, el proceso de etiquetado de datos, que muchas veces depende de decisiones humanas, también puede introducir sesgos. En situaciones donde el etiquetado es subjetivo o está influenciado por expectativas culturales o sociales, los sesgos implícitos de los etiquetadores pueden reflejarse en los datos. Este problema se ve agravado cuando el etiquetado se realiza a gran escala utilizando procesos automáticos, los cuales heredan los prejuicios de los datos de entrenamiento iniciales [3].

Un concepto importante es el del "efecto mariposa", el cual destaca cómo cambios aparentemente menores pueden generar resultados significativos e impredecibles en sistemas complejos. Este fenómeno es relevante y significativo en el ámbito de la equidad y el sesgo en la inteligencia artificial (IA). Factores como sesgos sutiles en los datos iniciales, desviaciones durante el entrenamiento del algoritmo o cambios en la distribución de datos entre el entrenamiento y la prueba pueden, de manera inadvertida, dar lugar a resultados marcadamente injustos. Estos resultados suelen afectar de manera desproporcionada a grupos marginados, reforzando desigualdades sociales preexistentes [8].

Métodos de mitigación del sesgo

Derivado de la problemática que se ha explicado, múltiples investigaciones han buscado erradicar los sesgos en los datos y para ello los métodos de mitigación de sesgo se han vuelto esenciales para crear modelos más justos y evitar que se perpetúen disparidades en los sistemas predictivos.

Uno de los primeros pasos para reducir el sesgo en los datos consiste en una recolección cuidadosa y en un análisis exhaustivo para identificar posibles fuentes de discriminación. En [1] se destaca la importancia de examinar el historial y el contexto de los datos para identificar sesgos estructurales -como aquellos derivados de desigualdades históricas- y ajustarlos antes de incorporarlos al modelo.

En la etapa del procesamiento de los datos, se pueden realizar ajustes en los datos para minimizar el impacto de características sensibles (género, raza, etc.) en el modelo. Barocas, Hardt y Narayanan [2] sugieren técnicas como la reponderación (*reweighting*) y el remuestreo (*resampling*), que ajustan las frecuencias de ciertas observaciones para equilibrar la representación de los grupos.

Posteriormente, en la etapa del entrenamiento de un modelo predictivo, Chouldechova [3] explica que una técnica clave es modificar la estructura del mismo para que sea menos susceptible a patrones discriminatorios. Por ejemplo, se puede implementar la optimización de justificación equitativa (*fairness constraints*), en la cual el

© 2025 - Sociedad Mexicana de Inteligencia Artificial ISSN 2007-0691

modelo se entrena para minimizar el sesgo en las predicciones, balanceando la precisión con la justicia.

La fase de evaluación es fundamental para garantizar que el modelo no solo tenga un buen rendimiento, sino que también cumpla con métricas de equidad. En [4] muestran cómo una evaluación exhaustiva y continua ayuda a identificar sesgos inadvertidos en sistemas predictivos, como los observados en herramientas de contratación o predicción de reincidencia. Esta evaluación puede incluir métricas como impacto dispar y paridad de oportunidad para medir y ajustar las predicciones.

También cabe mencionar que es crucial que los desarrolladores y usuarios puedan entender y verificar cómo un modelo toma decisiones. En [2], especifican los autores que proporcionar explicaciones claras permite identificar y mitigar posibles sesgos, además de construir confianza en el modelo y sus predicciones.

Otra de las formas que sugiere Emilio Ferrara [8] es crear conjuntos de datos equilibrados y representativos, lo cual es fundamental para mitigar el "efecto mariposa" en la equidad y el sesgo de la inteligencia artificial. Se pueden emplear varias técnicas para garantizar que los conjuntos de datos estén equilibrados y representen con precisión a la población de interés, tales como: sobremuestreo de clases minoritarias, submuestreo de clases mayoritarias, generación de datos sintéticos y muestreo estratificado.

Por otro lado, se propone equidad algorítmica el cual es un aspecto fundamental del aprendizaje automático que se centra en garantizar un trato equitativo entre diferentes grupos, mitigar los sesgos y reducir la sensibilidad de los modelos a cambios mínimos en los datos de entrada [8], todo esto con modelos como: aprendizaje automático consciente de la equidad, métodos de postprocesamiento para lograr equidad a través de la conciencia y regularización para la equidad.

Ética y regulación

La aplicación de algoritmos de aprendizaje automático en áreas sensibles, como la salud, el sistema judicial y la contratación, plantea importantes desafíos éticos y la necesidad de regulaciones efectivas. El desarrollo de estos sistemas puede acarrear consecuencias significativas para los individuos y las comunidades, especialmente cuando se manejan datos sensibles que reflejan desigualdades sociales históricas, por lo que se vuelve imprescindible establecer marcos normativos que garanticen su uso justo y seguro, por lo que es fundamental establecer regulaciones y supervisión adecuadas para garantizar un desarollo y uso del aprendizaje automático (Ver Figura 2). A continuación se mencionan algunas:

Necesidad de regulaciones y supervisión

Los algoritmos en entornos sensibles requieren una supervisión rigurosa y regulaciones claras para evitar daños no intencionados. Algunos autores sugieren que las políticas regulatorias deberían centrarse en establecer estándares mínimos de justicia y en implementar auditorías de equidad periódicas. Esto incluye el monitoreo de las decisiones algorítmicas y la intervención cuando se detecten sesgos [2].

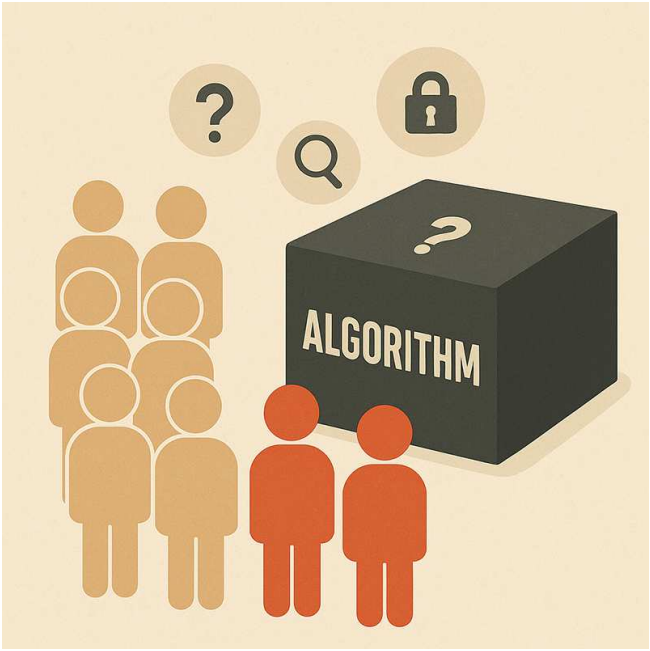


Figura 2. Representación gráfica sobre la importancia de la transparencia y explicabilidad de los algoritmos de aprendizaje.

Transparencia y explicabilidad

La falta de transparencia en los algoritmos puede llevar a problemas éticos, ya que usuarios y afectados no comprenden el proceso de toma de decisiones debido al comportamiento como el de una "caja negra". Según [4], esta "caja" puede generar desconfianza, especialmente cuando el sistema toma decisiones sobre cuestiones sensibles. Garantizar la explicabilidad es esencial para que los desarrolladores y usuarios identifiquen y comprendan el impacto del modelo en distintos grupos.

Responsabilidad y rendición de cuentas

La responsabilidad en el desarrollo y uso de algoritmos es crucial para mitigar riesgos. Como ilustra el caso de Amazon [5], en el que un sistema de contratación sesgado discriminó a candidatas mujeres, los desarrolladores y empresas deben ser responsables de los resultados de sus algoritmos. Políticas de rendición de cuentas pueden ayudar a prevenir la adopción de sistemas sin una adecuada evaluación ética.

Propuesta de regulación adicional

Considerando estos desafíos, una regulación integral para entornos sensibles debería incluir la evaluación de

riesgos antes del despliegue, auditorías de equidad periódicas y la obligación de contar con comités éticos independientes que supervisen el desarrollo y uso de estos sistemas. También es relevante adoptar el marco de "ética por diseño" que recomienda la *European Union Agency for Fundamental Rights* (2020), donde se incorporan principios éticos y de justicia durante todas las etapas de desarrollo del sistema.

Conclusiones

El sesgo en los datos representa uno de los mayores desafíos en la implementación de algoritmos de aprendizaje automático, ya que puede perpetuar y, en algunos casos, amplificar desigualdades sociales existentes. Las aplicaciones de estos sistemas en áreas críticas como la justicia penal, la contratación y el acceso a servicios financieros requieren una vigilancia y un compromiso con la justicia, dado que los errores en estas áreas pueden afectar profundamente a individuos y comunidades vulnerables. Abordar el sesgo no es solo una cuestión técnica; implica una responsabilidad ética de crear herramientas que contribuyan a una sociedad equitativa. Para avanzar hacia un futuro más justo, es esencial que los desarrolladores, investigadores y legisladores trabajen en conjunto para integrar la equidad desde el diseño hasta la implementación de algoritmos.

Entre las recomendaciones clave para minimizar el sesgo se incluyen:

- Compromiso con la diversidad en los datos: Asegurarse de que los datos utilizados en el entrenamiento de algoritmos representen de manera justa a todas las poblaciones afectadas, evitando sesgos de muestreo y exclusión.
- Auditorías de equidad y transparencia: Implementar auditorías constantes que identifiquen y corrijan posibles sesgos en el modelo antes de su implementación y a lo largo de su ciclo de vida, garantizando así que los resultados sean justos y explicables.
- Normas éticas y marcos regulatorios: Establecer regulaciones claras que guíen el desarrollo y uso de algoritmos en áreas sensibles. Estas normativas deben asegurar la transparencia, la rendición de cuentas y el respeto por la privacidad de las personas.
- Capacitación continua: Fomentar la formación ética y técnica en aprendizaje automático para que tanto desarrolladores como usuarios comprendan el impacto de sus decisiones y estén mejor equipados para gestionar los riesgos asociados al sesgo.

Abordar el sesgo en los datos y algoritmos es fundamental para que los sistemas de aprendizaje automático contribuyan al progreso social. Con un enfoque ético y regulaciones adecuadas, el aprendizaje automático puede ser una herramienta poderosa que promueva la igualdad y reduzca las disparidades en áreas clave, ayudando a construir un futuro más justo y equitativo para todos.✷

REFERENCIAS

1. Mehrabi, N., Morstatter, F., Saxena, N., Lerman, K., y Galstyan, A. (2021). A survey on bias and fairness in machine learning. *ACM Computing Surveys*, 54(6), 1-35.
2. Barocas, S., Hardt, M., y Narayanan, A. (2019). Fairness and machine learning: Limitations and opportunities. *MIT Press*. Recuperado el 2 de junio del 2025, de `https://mitpress.mit.edu/9780262048613/fairness-and-machine-learning/`.
3. Chouldechova, A. (2017). Fair prediction with disparate impact: A study on bias in recidivism prediction instruments. *Big Data*, 5(2), 153-163.
4. Angwin, J., Larson, J., Mattu, S., y Kirchner, L. (2016). Machine Bias. *ProPublica*. Recuperado el 2 de junio del 2025, de `https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing`.
5. Dastin, J. (2018). Amazon Scraps Secret AI Recruiting Tool That Showed Bias Against Women. *Reuters*. Recuperado el 2 de junio de 2025, de `https://www.reuters.com/article/world/insight-amazon-scraps-secret-ai-recruiting-tool-that-showed-bias-against-women-idUSKCN1MK0AG/`.
6. Maharana K., Mondal S., y Nemade B. (2022). A review: Data pre-processing and data augmentation techniques. *Global Transitions Proceedings*, 3(1), 91-99.
7. Rangel, R. A. P., Joaquín Pérez, O., Juan Javier González, B., Gelbukh, A., Sidorov, G., y Rodríguez, M. M. J. (2005). A Domain Independent Natural Language Interface to Databases Capable of Processing Complex Queries. *Lecture Notes in Artificial Intelligence*,3789. Springer-Verlag, 833–842.
8. Ferrar E. (2024). The Butterfly Effect in artificial intelligence systems: Implications for AI bias and fairness. *Machine Learning with Applications*, 15.

SOBRE LOS AUTORES



Daniel Sánchez Ruiz obtuvo el grado de doctor en Ingeniería del Lenguaje y el Conocimiento en la Benemérita Universidad Autónoma de Puebla. Actualmente es profesor en la Unidad Profesional Interdisciplinaria de Ingeniería Campus Tlaxcala - IPN y es miembro del Sistema Nacional de Investigadoras e Investigadores Nivel Candidato. Sus intereses son el reconocimiento de patrones, la ingeniería de características y el uso de las tecnologías de forma inclusiva y con perspectiva de género.

© 2025 - Sociedad Mexicana de Inteligencia Artificial ISSN 2007-0691



Ricardo Ramos Aguilar obtuvo el grado de Doctor en Filosofía (PhD) en Ingeniería del Lenguaje y del Conocimiento en la Facultad de Ciencias de la Computación de la Benemérita Universidad Autónoma de Puebla. Sus principales áreas de investigación incluyen el procesamiento del lenguaje natural sobre voz, análisis de bioseñales en la salud y la ingeniería de características sobre series de tiempo.



Eric Ramos Aguilar es actualmente candidato a doctor en Ingeniería de Lenguaje y Conocimiento en la Facultad de Ciencias de la Computación de la Benemérita Universidad Autónoma de Puebla y Docente de la UPIIT-IPN en el área de Ciencia de Datos e Inteligencia Artificial. Sus principales áreas de investigación incluyen el procesamiento de lenguaje natural, el reconocimiento de patrones, el procesamiento y análisis de audio digital, y el estudio de las lenguas indígenas de México.



Jesús García Ramírez obtuvo el grado de Doctor en Ciencias de la Computación en el Instituto Nacional de Astrofísica y Electrónica (INAOE). Sus principales áreas de investigación incluyen el aprendizaje profundo, aprendizaje por refuerzo y la explicabilidad de los modelos.

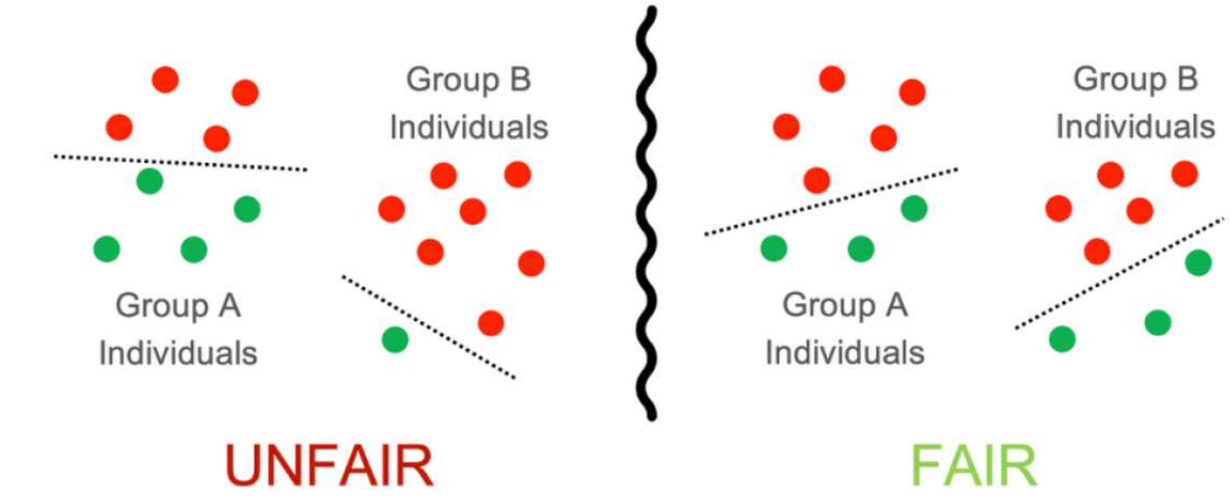


Imagen tomada de: https://www.linkedin.com/pulse/group-fairness-machine-learning-najeeb-jebreel-ck3jf .

© 2025 - Sociedad Mexicana de Inteligencia Artificial ISSN 2007-0691

IA & Educación

Julieta Noguez, Lucía Barrón y Yasmín Hernández

El uso responsable de la inteligencia artificial en educación

El avance acelerado de la inteligencia artificial (IA) ha generado nuevas oportunidades de su aplicación en una gran diversidad de áreas, entre las que destacan facilitar los diagnósticos médicos, facilitar las conexiones humanas a través de las redes sociales, aumentar la eficiencia laboral mediante la automatización de tareas, realizar labores especializadas para las empresas, tales como mercadotecnia o finanzas, etc. Sin embargo, los rápidos cambios también plantean riesgos y profundos dilemas éticos, que surgen del potencial que tienen los sistemas basados en IA para reproducir prejuicios, contribuir a la degradación del clima y amenazar los derechos humanos, entre otros [1].

Aunque en términos generales el uso responsable de la IA implica desarrollarla y utilizarla de forma ética, transparente y segura, asegurando que se beneficie a la sociedad sin causar daños, vale la pena reflexionar sobre su uso y las implicaciones que está teniendo en educación. Un de los principales retos que enfrentan los docentes es el uso que los estudiantes están dando a las herramientas de la IA generativa, particularmente los IA chatbots, tales como ChatGPT, Gemini, Copilot o Deepsek entre otros, no sólo para mantener conversaciones en lenguaje natural y crear contenido original, sino para resolver tareas con poco esfuerzo.

La cuestión clave es determinar cómo afecta el uso de las herramientas de la IA generativa en el aprendizaje y si realmente los estudiantes adquieren habilidades y aprendizaje de largo alcance conforme van utilizando esas herramientas. Recientemente, se publicó el artículo ¿La IA generativa puede perjudicar el aprendizaje? (Generative AI Can Harm Learning?) [2], el cual estudia el impacto específico del GPT-4 de OpenAI, en el aprendizaje humano en el contexto de clases de matemáticas de un instituto, con una participación de casi mil estudiantes. Se evaluaron dos tutores basados en GPT, uno que imita la interfaz estándar de ChatGPT (llamado GPT Base) y otro con indicaciones diseñadas para salvaguardar el aprendizaje (llamado GPT Tutor). Los resultados obtenidos mostraron que el acceso a GPT-4 mejoró significativamente el rendimiento (48 % de mejora para GPT Base y 127 % para GPT Tutor). Sin embargo, también se observó que, cuando se les retiró el acceso a estas herramientas, los alumnos obtuvieron peores resultados que los que nunca tuvieron acceso (una reducción del 17 % en el caso de GPT Base). Es decir, en ese caso, el acceso a GPT-4 perjudicó los resultados del aprendizaje.

Por otra parte, se publicó entre otros, un estudio donde se demuestra que actualmente no es posible confiar plenamente en las herramientas generativas de IA para el aprendizaje de la Física, por su bajo rendimiento académico en la resolución de problemas [3]. En este trabajo se presentó una comparación detallada de la precisión de dos chatbots de IA, ChatGPT-4 y Gemini 1.5 Pro, midiendo su capacidad para responder correctamente a preguntas de opción múltiple de cuestionarios estándar de física para estudiantes universitarios. Se analizó una muestra total de 790 preguntas, que abarcaban 13 temas de mecánica, desde unidades y medidas hasta gravitación. En concordancia parcial con estudios similares, se observó que ChatGPT logró una precisión global de sólo 58 puntos (en una escala de 0 a 100), mientras que Gemini alcanzó 79, lo que representa una mejora del 36 %. Un análisis de varianzas demostró que esta diferencia era significativa. En ese sentido se recomienda a los profesores y estudiantes considerar estos resultados como una advertencia sobre la actual falta de precisión y el potencial engaño de estos chatbots de IA cuando se utilizan para resolver preguntas en temas de física de pregrado que requieren procedimientos matemáticos.

Los estudios para verificar el impacto de la IA generativa en el aprendizaje se multiplican y los profesores deberán establecer estrategias para garantizar que su uso realmente beneficie a los estudiantes. Es muy importante concientizar al estudiante del uso responsable de estas herramientas, promoviendo que verifique los resultados y analizando con espíritu crítico las respuestas obtenidas.

La Inteligencia Artificial en educación, aunque puede generar controversias, puede ser utilizada para producir cambios significativos en la calidad educativa y el mejoramiento del desempeño de los docentes en sus labores de enseñanza. Es muy importante que tanto profesores como estudiantes exploren cómo modificar la forma de enseñar y aprender a través de las diferentes aplicaciones de la IA en educación. Adicional a la IA generativa, existen otras tecnologías de IA que pueden emplearse con mucho éxito en la educación, por ejemplo, los tutores virtuales inteligentes, herramientas de aprendizaje personalizadas , ambientes de evaluación adaptativos, laboratorios virtuales y remotos. La propuesta es que se aprovechen las

© 2025 - Sociedad Mexicana de Inteligencia Artificial ISSN 2007-0691

herramientas de IA para transformar la visión de la educación y lograr que los estudiantes incrementen sus capacidades y asuman con una mayor responsabilidad su uso.✷

REFERENCIAS

1. Ramos G. (2024). Ética de la inteligencia artificial. UNESCO. https://www.unesco.org/es/artificial-intelligence/recommendation-ethics Consultada el 30 de mayo de 2025.
2. Bastani H., Bastani O., Sungu A., Ge H., Kabakcı Ö., and Mariman R. (2024) Generative AI Can Harm Learning. The Wharton School Research Paper. http://dx.doi.org/10.2139/ssrn.4895486
3. Robledo-Rella, V., Gonzalez-Nucamendi, A., Neri, L., García-Castelán, R. M., Noguez, J., & Valverde-Rebaza, J. (2025). Can We Trust AI Chatbots to Teach University Physics? A Performance Comparison of AI Chatbots. In 2025 IEEE Global Engineering Education Conference (EDUCON) (pp. 1-7). DOI:10.1109/EDUCON62633.2025.11016485



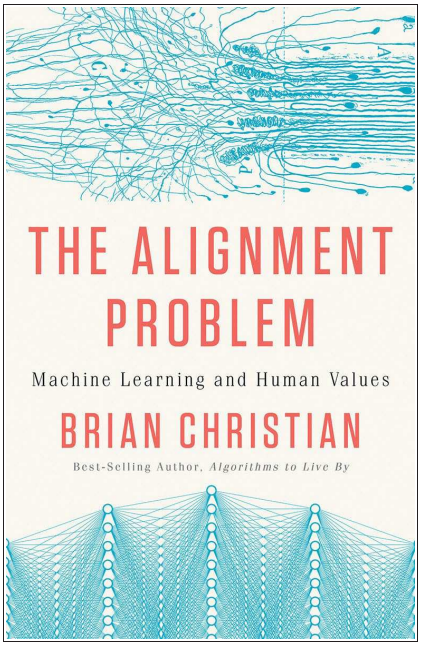
© 2025 - Sociedad Mexicana de Inteligencia Artificial ISSN 2007-0691

Deskubriendo Konocimiento

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

Revisión de "The Alignment Problem" *

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



Portada del libro*

El libro que hoy nos ocupa es acerca del aprendizaje automático y los valores humanos, acerca de sistemas que aprenden de datos sin ser programados explícitamente, y de cómo y qué estamos tratando de enseñarles. El interés en estos temas ha aumentado en los últimos años al seno de dos comunidades: los estudiosos de los riesgos éticos de la tecnología actual y los interesados en el riesgo futuro que conllevan sistemas cada vez más capaces, flexibles y autónomos. La seguridad en torno a la Inteligencia Artificial (IA) ha pasado de ser marginal, a constituirse en uno de los temas centrales más activos del área.

Curiosamente, esta preocupación estaba ahí desde un principio. A finales de los años cincuenta del sigo pasado, Norbert Wiener vio en acción al jugador de damas chinas de Arthur Samuel, sistema que mejoraba su desempeño usando las partidas previamente jugadas en un proceso que el mismo Samuel llamó aprendizaje automático. El padre de la cibernética quedó tan impresionado que escribió un artículo** donde observa: "Si usamos, para lograr nuestros propósitos, una agencia mecánica con cuya operación no podemos interferir eficientemente una vez puesta en marcha... entonces sería mejor estar seguros de que el propósito depositado en la máquina es el propósito que deseamos y no una mera imitación colorida de éste". Quizás la formulación más antigua y concisa de lo que hoy llamamos el problema de la alineación.

¿Qué podría ir mal si el problema no se atiende adecuadamente? La introducción de nuestro libro lo ejemplifica con cuatro casos: sistemas etiquetando fotografías de humanos como gorilas; concluyendo por analogía que doctor - hombre + mujer es enfermera; evaluando el riesgo de evadir la libertad condicional con un marcado sesgo contra la población afroamericana; dando vueltas en círculos frente a la meta hasta incendiarse, tras hacer una esplendida carrera. Y entonces, ¿Porqué dejamos de interesarnos en ello? El mismo Wiener parece explicar y predecir *** el olvido y resurgimiento del interés por el problema de la alineación: "En el pasado, una visión parcial e inadecuada del propósito humano ha sido relativamente inocua únicamente porque ha estado acompañada de limitaciones técnicas." Desafortunadamente, éste ya no es el caso de los actuales sistemas basados en aprendizaje automático. Pasamos rápidamente de clasificar flores y gatos en el laboratorio, a producir sistemas que son cada vez más omnipresentes y poderosos. El problema de la alineación devino urgente.

El libro presenta una extensa y detallada narrativa de cómo estamos tratando de asegurarnos de que nuestros modelos construidos con aprendizaje automático capturen nuestras normas y valores, entiendan lo que queremos decir o lo que intentamos, y sobre todo, hagan lo que queremos. Para ello, el texto se organiza en tres partes. La primera de ellas explora la manera en que nuestros sistemas actuales divergen de nuestras mejores in-

*B. Christian. *T*he Alignment Problem: Machine Learning and Human Values. W.W. Norton & Company, New York, NY, USA, 2020.
**N. Wiener. *Moral and Technique Consequences of Automation.* Science, New Series, Vol. 131, No. 34, May 6, 1960, pp. 1355-1358
***N. Wiener. *God & Golem, Inc.* MIT Press, Boston, MA, USA, 1966.

© 2025 - Sociedad Mexicana de Inteligencia Artificial ISSN 2007-0691

tenciones y discute la problemática de tratar de hacer que esas intenciones sean explícitas en estos sistemas que sentimos somos capaces de supervisar. Los problemas de sesgo, equidad, transparencia son revisados en detalle, ejemplificando sus implicaciones y explorando las soluciones que la IA está proponiendo.

La segunda parte del libro aborda los sistemas basados en el aprendizaje por refuerzo, la versión moderna de los esfuerzos de Samuel con las damas chinas. Los llamados agentes inteligentes no sólo clasifican y predicen, sino que además actúan intentando maximizar la utilidad que el medio ambiente les proporciona, usando el mundo como un crítico. Los problemas aquí son de diseño de recompensas, de la exploración versus la explotación de lo aprendido y de convergencia ¿El sistema hace lo que realmente queremos? Estos temas que tienen que ver con la evolución, la motivación en los humanos y la delicadeza de los incentivos.

La tercera parte aborda las técnicas de vanguardia en la seguridad de la IA, lo que estamos haciendo para alinear nuestros complejos sistemas autónomos con normas y valores que son sutiles o demasiado elaborados para ser especificados directamente. Los conceptos fundamentales para ello, provienen del estudio del aprendizaje social: imitación y aprendizaje colaborativo. Los problemas tienen que ver con la forma computacional de establecer esta cooperación entre humanos y sistemas basados en IA, por ejemplo, mediante aprendizaje por refuerzo inverso y la observación de patrones de activación en nuestras redes neuronales.

Nuestro autor (`https://brianchristian.org`) es un investigador, programador y poeta, interesado en las ciencias cognitivas computacionales, los modelos formales de la cognición humana y la alineación de la IA. Actualmente, Christian está afiliado al Centro para la IA Compatible con lo Humano, en la Universidad de Berkeley; y al Instituto para el Estudio Avanzado, famoso por haber tenido a Albert Einstein en su claustro original. Sus libros, *The Most Human Human* (Anchor, 2011) y *Algorithms to Live by* (Henry Holt and Co., 2016), han sido éxitos de venta ampliamente recomendados.

Me parece que estamos ante una lectura obligada para todos los interesados en la IA y, en particular, para las primeras generaciones de estudiantes focalizados explícitamente en la ética y seguridad de esta disciplina. La amplitud y profundidad de la investigación llevada a cabo, incluyendo casi un centenar de entrevistas, dan cuenta de dos cosas muy importantes: La diversidad y ubicuidad de los problemas que emergen de una mala alineación de valores; y el hecho de que la alineación es un problema que trasciende a la IA misma, requiriendo interacciones con las ciencias sociales y cognitivas. La buena nueva es que tal interacción comienza a darse en ambas direcciones, abriendo nuevas áreas de oportunidad para la aplicación de la IA y contribuyendo la seguridad de la misma. ✷

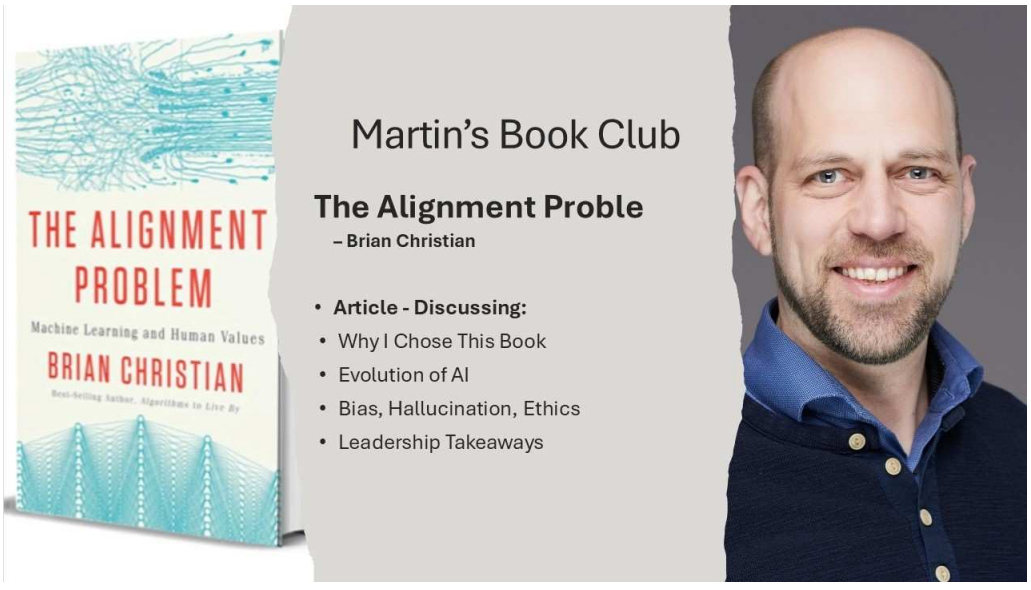
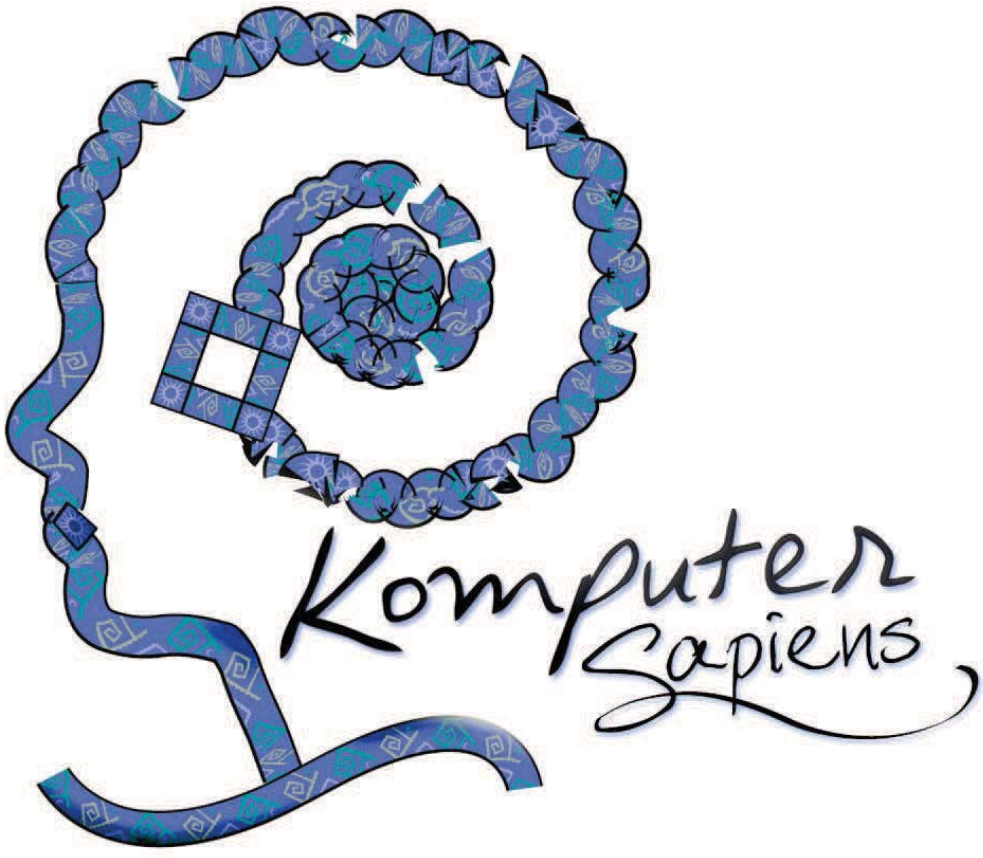


Imagen tomada del sitio:
https://www.linkedin.com/pulse/alignment-problem-its-complicated-martin-vogel-k4b9c.

© 2025 - Sociedad Mexicana de Inteligencia Artificial ISSN 2007-0691



¡Publique en Komputer Sapiens!

Ciencia y Tecnología
Secretaría de Ciencia, Humanidades, Tecnología e Innovación

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