Módulo 1. Ética IA y ciberseguridad

CÓRDOBA CREA FUTURO

Introducción

Introducción

El 30 de noviembre de 2022 OpenAI lanzó ChatGPT, lo que ocasionó una disrupción tecnológica mundial que provocó un nuevo paradigma educativo global, entre otras cosas. Este inicio marca el puntapié inicial de la era de la IA generativa accesible para el público general, lo que permite conversaciones fluidas y creativas con un modelo de lenguaje grande (LLM).

Recorriendo un poco la historia, sabemos que la inteligencia artificial (IA) no se «creó» en un solo momento, sino que sus bases surgieron en los años 40 con modelos de neuronas artificiales (McCulloch y Pitts, 1943) y en los 50 con pioneros como Alan Turing, pero el término «inteligencia artificial» fue acuñado por John McCarthy en la Conferencia de Dartmouth en 1956, evento considerado el nacimiento formal de la IA

como campo de estudio, en el que se establecieron sus objetivos y visión.

Muchos avances se han registrado desde los comienzos de la IA. Se destaca el uso en el campo militar con el desarrollo de un traductor de idiomas del ruso al inglés en el año 1966 en el proyecto llamado ALPAC (Alonso Martín, 2022; Pangeanic, 2018) o Automatic Language Processing Advisory Committee. En nuestros días, encontramos traductores en línea compatibles con distintos dispositivos móviles y de escritorio para sistemas operativos tales como Android, Windows, etcétera. Es importante destacar que Google Translate se lanzó en 2001, servicio que podía traducir ocho idiomas al inglés y viceversa.

En esta reciente historia, los avances de la IA no pasaban desapercibidos. A fines de febrero de 2020, bajo un evento denominado *Rome Call for Artificial Intelligence Ethic* (Capelli, 2019), el papa Francisco convocó en la Ciudad del Vaticano, dos años antes del lanzamiento de ChatGPT, a dialogar y firmar un documento con un enfoque ético de la inteligencia artificial a empresas tales como Microsoft, IBM, FAO (Organización de las Naciones Unidas para la Agricultura y la Alimentación).

Figura 1: Humanoide Ameca

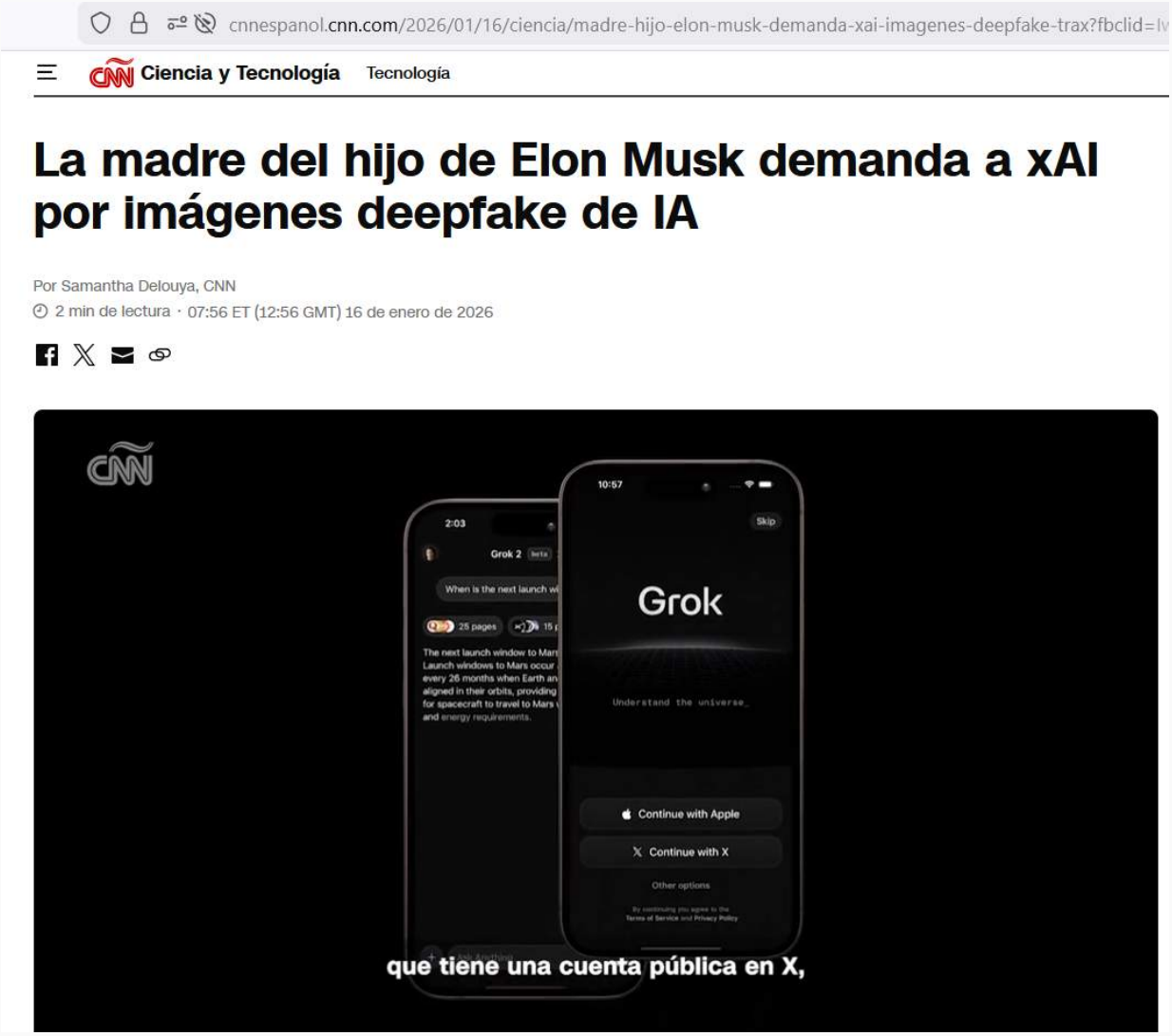
Fuente: [imagen sin título sobre humanoide Ameca], s. f., https://bit.ly/3Z2Sj07.

El avance ha sido de tal magnitud que, por ejemplo, ha preocupado lo mencionado por el humanoide **Ameca**, que dijo que es capaz de dirigir al mundo. También está el caso de otro humanoide, Sophia (Bellucci, 2021), que pretendía destruir a la humanidad y ahora la protege.

La preocupación por el uso ético de la IA sigue vigente en 2026; en este contexto, se puede ver el cuestionamiento a la red social X por permitir hacer desnudos e imágenes provocativas en personas que incluían a menores de edad. Esta situación derivó, como lo muestra la imagen del diario digital CNN en Español, en que Ashley St. Clair (madre del hijo de Elon Musk) demandara a xAI por imágenes «*deepfake*» de IA. Pero esto no es todo: además, suceden cosas mucho más graves: Google y la IA de Character alcanzan acuerdos de

conciliación en EE. UU. por suicidios de menores (EuroNews, 2026).

Figura 2: Portada de noticia de CNN



Fuente: captura de pantalla de CNN en Español (https://cnnespanol.cnn.com/).

Si bien se puede enumerar un sinfín de ejemplos que demuestran la vigencia de *Call Rome* convocada por el papa Francisco, el uso y alcance de las herramientas de IA deben ser sabiamente empleados; de lo contrario, se corre el riesgo de encontrarse ante casos de mal uso como los de *deepfakes* que pueden generar potenciales daños perdurables en el tiempo.

Definiciones

Antes de comenzar, para dar un contexto inicial luego de la introducción hecha, se transcriben las definiciones de la Real Academia Española sobre algunos términos.

Con base en la RAE (s. f. a), el término ético/ca, presenta los siguientes significados:

a) “Conjunto de principios y normas que rigen la conducta humana, relacionados con el sentido del bien y del mal.

b) Parte de la filosofía que estudia el comportamiento del hombre desde el punto de vista del bien o del mal, y los principios por los que debe regirse, teniendo como finalidad el bien”.

Por otra parte, en lo que respecta a la definición del término ciberseguridad, la RAE (s. f. b) establece: “Conjunto de elementos, medidas y equipos destinados a controlar la seguridad informática de una entidad o un espacio virtual”.

CONTINUAR

1. ¿Cómo se relaciona la IA con la ética?

¿Cómo se relaciona la IA con la ética?

¿Cómo se relaciona la IA con la ética?

La ética es un conjunto de principios morales que nos ayudan a discernir entre el bien y el mal. La ética de la IA es un campo multidisciplinario que estudia cómo optimizar el impacto beneficioso de la inteligencia artificial (IA) mientras se reducen los riesgos y los resultados adversos.

Algunos ejemplos de cuestiones éticas de la IA son la responsabilidad y la privacidad de los datos, la imparcialidad, la robustez, la transparencia, la sustentabilidad medioambiental, la inclusión, la agencia moral, la alineación de valores, la responsabilidad, la confianza y el mal uso de la tecnología.

Con la aparición de *big data*, las compañías aumentaron su enfoque para impulsar la automatización y la toma de decisiones basada en datos en todas sus organizaciones. Si bien la intención suele ser, sino siempre, mejorar los resultados comerciales, las compañías están experimentando consecuencias imprevistas en algunas de sus aplicaciones de IA, particularmente debido a un diseño de investigación inicial deficiente y conjuntos de datos sesgados.

A medida que salieron a la luz algunos casos cuyos resultados no eran los esperados, surgieron nuevas directrices, principalmente de las comunidades de investigación y ciencia de datos, para abordar las preocupaciones en torno a la ética de la IA. Las compañías líderes en el campo de la IA también mostraron un interés personal en dar forma a estas directrices, dado que ellas mismas comenzaron a experimentar algunas de las consecuencias de no mantener los estándares éticos en sus productos. La falta de diligencia en esta área puede generar exposición reputacional, regulatoria y legal, lo que lleva a establecer sanciones costosas en aquellos países que tienen instrumentación legal. Como ocurre con todos los avances tecnológicos, la innovación tiende a superar la regulación gubernamental en campos nuevos y emergentes. A medida que se desarrolle la experiencia adecuada dentro de la industria gubernamental, podemos esperar que las compañías sigan más protocolos de

IA, lo que les permitirá evitar cualquier violación de los derechos humanos y las libertades civiles.

Establecimiento de principios para la ética de la IA

Mientras se desarrollan normas y protocolos para gestionar el uso de la IA, la comunidad académica ha aprovechado el informe Belmont (National Commission for the Protection of Human Subjects of Biomedical and Behavioral Research, 1979) como medio para orientar la ética dentro de la investigación experimental y el desarrollo algorítmico. Hay tres principios fundamentales que se desprenden del informe Belmont (y que sirven de guía para el diseño de experimentos y algoritmos.

1. **Respeto a las personas**: este principio reconoce la autonomía de los individuos y defiende la expectativa de que los investigadores protejan a las personas con autonomía disminuida, lo que podría deberse a una variedad de circunstancias como enfermedad, discapacidad mental o restricciones de edad. Este principio se concentra principalmente en la idea del consentimiento. Las personas deben ser

conscientes de los posibles riesgos y beneficios de cualquier experimento en el que participen, y deben elegir participar o retirarse en cualquier momento antes y durante el experimento.

2 **Beneficencia**: este principio proviene de la ética de la atención médica, a partir de la cual los médicos hacen el juramento de «no hacer daño». Esta idea se puede aplicar fácilmente a la inteligencia artificial, en la que los algoritmos pueden amplificar los sesgos en torno a la raza, el género, las inclinaciones políticas, etcétera, a pesar de la intención de hacer el bien y mejorar un sistema determinado.

3 **Justicia**: este principio trata temas como la equidad y la igualdad. ¿Quién debería cosechar los beneficios de la experimentación y el *machine learning*? El informe Belmont ofrece cinco formas de distribuir cargas y beneficios:

- igualdad.
- Necesidad individual.
- Esfuerzo individual.

- Contribución social.
- Mérito

Principales preocupaciones de la IA en la actualidad

Hay una serie de cuestiones que están al frente de las conversaciones éticas en torno a las tecnologías de IA en el mundo real.

- **Modelos fundacionales e IA generativa**

Tal como se describió en la introducción, el lanzamiento de ChatGPT en 2022 marcó un verdadero punto de inflexión para la inteligencia artificial. Las capacidades del *chatbot* de OpenAI (desde redactar reportes legales hasta depurar códigos) abrieron nuevas posibilidades sobre lo que la IA puede hacer y cómo se puede aplicar en casi todas las industrias.

ChatGPT y herramientas similares se basan en modelos fundacionales, modelos de IA que pueden adaptarse a una amplia gama de tareas posteriores. Los modelos fundacionales suelen ser modelos generativos a gran escala, compuestos por miles de millones de parámetros, que se

entrenan con datos no etiquetados mediante autosupervisión. Esto permite a los modelos fundacionales aplicar rápidamente lo que han aprendido de un contexto a otro, lo que los hace altamente adaptables y capaces de llevar a cabo una amplia variedad de tareas diferentes. Sin embargo, hay muchos posibles problemas y preocupaciones éticas en torno a los modelos fundacionales que son comúnmente reconocidos en la industria de la tecnología, como la parcialidad, la generación de contenido falso, la falta de explicabilidad, el mal uso y el impacto social. Muchos de estos problemas son relevantes para la IA en general, pero adquieren una nueva urgencia a la luz del poder y la disponibilidad de los modelos fundacionales.

- **Singularidad tecnológica**

La singularidad tecnológica es un escenario teórico en el que el crecimiento tecnológico se vuelve incontrolable e irreversible, lo que culmina en cambios profundos e impredecibles en la civilización humana. Si bien este tema atrae mucha atención pública, a muchos investigadores no les preocupa la idea de que la IA supere a la inteligencia humana en un futuro cercano o inmediato. En teoría, este fenómeno está impulsado por la aparición de inteligencia artificial (IA) que supera las capacidades cognitivas humanas

y puede mejorarse a sí misma de forma autónoma. El término «singularidad» en este contexto se basa en conceptos matemáticos que indican un punto en el que los modelos existentes se descomponen y se pierde la continuidad en la comprensión. Esto describe una era en la que las máquinas no solo igualan, sino que superan sustancialmente la inteligencia humana, lo que inicia un ciclo de evolución tecnológica que se perpetúa a sí misma.

La teoría sugiere que tales avances podrían evolucionar a un ritmo tan rápido que los humanos serían incapaces de prever, mitigar o detener el proceso. Esta rápida evolución podría dar lugar a inteligencias sintéticas que no solo sean autónomas, sino también capaces de innovaciones que están más allá de la comprensión o control humanos. La posibilidad de que las máquinas puedan crear versiones aún más avanzadas de sí mismas podría cambiar a la humanidad hacia una nueva realidad donde los humanos ya no son las entidades más capaces. Las implicaciones de llegar a este punto de singularidad podrían ser buenas para la raza humana o catastróficas. Por ahora, el concepto queda relegado a la ciencia ficción; no obstante, puede ser valioso contemplar cómo podría ser un futuro así, para que la humanidad pueda dirigir el desarrollo de la IA de tal manera que promueva sus intereses civilizatorios.

Varias tecnologías actuales actúan como precursoras de la singularidad tecnológica, cada una de las cuales representa avances en áreas críticas para el desarrollo de la IA superinteligente.

Estas son algunas tecnologías claves.

Redes neuronales artificiales y aprendizaje profundo:

estas tecnologías constituyen el eje de gran parte de la investigación y el desarrollo de la IA actual. Imitan la estructura y la función del cerebro humano hasta cierto punto y han permitido avances significativos en el *machine learning*. Las redes neuronales son especialmente cruciales para tareas como el reconocimiento de voz, el reconocimiento de imágenes y la navegación autónoma de vehículos.

Computación cuántica:

aunque aún se encuentra en sus primeras etapas, la computación cuántica promete aumentar exponencialmente la potencia y la eficiencia informática en un futuro cercano, lo que acelera de manera potencial las capacidades de la IA más allá de los límites actuales. Esta tecnología podría suponer un gran avance en la capacidad de la IA para resolver problemas complejos mucho más rápido que las computadoras tradicionales.

Procesamiento de lenguaje natural (PLN):

los avances en PLN, ejemplificados por tecnologías como los modelos ChatGPT (transformador generativo preentrenado), son cruciales para desarrollar una IA que pueda comprender y generar texto similar al humano. Esta capacidad es vital para que la IA lleve a cabo tareas más complejas que requieren comprender el contexto y los matices del lenguaje.

Robótica y automatización:

las innovaciones en robótica permiten cada vez más que las máquinas realicen tareas que requieren destreza y toma de decisiones que antes se pensaba que eran exclusivamente humanas. Estos avances no solo automatizan más tareas físicas, sino que también integran la IA para crear sistemas más autónomos.

Computación en la nube y big data:

el gran aumento en la generación de datos y la capacidad de almacenarlos y procesarlos en la nube son vitales para entrenar sistemas de IA más potentes. El *analytics* de *big data* y la infraestructura en la nube que lo respalda habilitan los complejos modelos de *machine learning* necesarios para el desarrollo de IA avanzada.

Biotecnología e interfaces cerebro-computadora (BCI):

los avances en la comprensión del cerebro humano y la imitación de sus funciones son cruciales para crear una IA que potencialmente podría pensar y aprender de la misma manera que los humanos. Además, las BCI que conectan los cerebros humanos directamente a las computadoras son un paso hacia la fusión de la inteligencia biológica y artificial, un concepto que a menudo se discute en escenarios de singularidad.

La IA sólida (IA que poseería inteligencia y conciencia iguales a las de los humanos) y la superinteligencia siguen siendo hipotéticas; las ideas plantean algunas preguntas interesantes a la hora de considerar el uso de sistemas autónomos, como los coches autónomos. Es poco realista pensar que un coche sin conductor nunca tendrá un accidente de coche, pero ¿quién es el responsable en esas circunstancias? ¿Deberíamos seguir apostando por los vehículos autónomos, o limitamos la integración de esta tecnología para crear solo vehículos semiautónomos que promuevan la seguridad de los conductores? El debate aún no ha concluido, pero estos son los tipos de debates éticos que se producen a medida que se desarrolla la nueva e innovadora tecnología de IA.

Impacto de la IA en los empleos

Si bien gran parte de la percepción pública de la inteligencia artificial se centra en la pérdida de empleos, esta preocupación probablemente debería replantearse. Con cada nueva tecnología disruptiva, vemos que cambia la demanda de roles laborales específicos.

Por ejemplo, al considerar la industria automotriz, muchos fabricantes de autos, están cambiando su enfoque hacia la producción de vehículos eléctricos para ajustarse a las iniciativas ecológicas. La industria energética no va a desaparecer, pero la fuente de energía está cambiando de una economía de combustible a una eléctrica.

La inteligencia artificial debe verse de manera similar, en el sentido de que esta desplazará la demanda de puestos de trabajo a otras áreas. Será necesario que haya personas que ayuden a administrar estos sistemas a medida que los datos crezcan y cambien todos los días. Seguirá siendo necesario contar con recursos para atender problemas más complejos en las industrias con más probabilidades de verse afectados por los cambios en la demanda de empleo, como el servicio de atención al cliente. Lo importante de la inteligencia artificial y su efecto en el mercado laboral será ayudar a las personas en la transición a estas nuevas áreas de demanda del mercado.

Privacidad

La privacidad tiende a discutirse en el contexto de la privacidad, protección y seguridad de los datos, y estas preocupaciones permitieron avanzar más en este sentido en los últimos años. Por ejemplo, en 2016, se creó en Europa la legislación GDPR para proteger los datos personales de las personas en la Unión Europea y el Espacio Económico Europeo, lo que le dio a los individuos más control sobre sus datos. En Estados Unidos, los estados individuales están desarrollando políticas, como la Ley de Privacidad del Consumidor de California (CCPA), que exige a las empresas que informen a los consumidores sobre la recopilación de sus datos.

Esta y otras leyes recientes han obligado a las empresas a repensar cómo almacenan y usan los datos de identificación personal (PII). Como resultado, las inversiones en seguridad se han convertido en una prioridad cada vez mayor para las empresas, que tratan de eliminar cualquier vulnerabilidad y oportunidad de vigilancia, piratería informática y ataques cibernéticos.

Prejuicios y discriminación

Los casos de sesgo y discriminación en una serie de sistemas de *machine learning* han planteado muchas cuestiones éticas sobre el uso de la inteligencia artificial. ¿Cómo podemos protegernos contra el sesgo y la discriminación cuando los conjuntos de datos de formación pueden prestarse al sesgo? Aunque las empresas suelen tener intenciones bien intencionadas en torno a sus esfuerzos de automatización, la incorporación de la IA en las prácticas de contratación puede tener consecuencias imprevistas. En su esfuerzo por automatizar y simplificar un proceso, Amazon sesgó involuntariamente a posibles candidatos por género para puestos técnicos abiertos y, en última instancia, tuvo que desechar el proyecto. A medida que surgen eventos, surgen otras preguntas puntuales sobre el uso de la IA en las prácticas de contratación, como qué datos debería poder utilizar al evaluar a un candidato para un puesto.

El sesgo y la discriminación tampoco se limitan a la función de recursos humanos; se pueden encontrar en una serie de aplicaciones, desde *software* de reconocimiento facial hasta algoritmos de redes sociales.

A medida que las empresas se vuelven más conscientes de los riesgos de la IA, también se vuelven más activas en la discusión sobre la ética y los valores de la IA. Por ejemplo, el año pasado, el CEO de IBM, Arvind Krishna, compartió que IBM ha puesto fin a sus productos de reconocimiento y análisis facial de IBM de propósito general, destacando que IBM se opone firmemente y no condona los usos de ninguna tecnología, incluida la de reconocimiento facial que ofrecen otros proveedores para vigilancia masiva, perfiles raciales, violaciones de derechos humanos y libertades básicas, o cualquier propósito que no sea consecuente con nuestros valores y principios de confianza y transparencia. Cabe destacar que el autor de este trabajo sí ha desarrollado una IA de reconocimiento facial en la que ha planteado un desafío más amplio, contemplando un protocolo de uso necesario.

Responsabilidad

No existe una legislación universal y general que regule las prácticas de IA, pero muchos países y estados trabajan para desarrollarlas y aplicarlas a nivel local. En la actualidad, existen algunas normas sobre IA, y muchas otras están a punto de entrar en vigor. Para llenar el vacío, han surgido marcos éticos como parte de una colaboración entre especialistas en ética e investigadores para regir la creación y distribución de

modelos de IA dentro de la sociedad. Sin embargo, por el momento, estos solo sirven como guía, y la investigación muestra que la combinación de responsabilidad distribuida y la falta de previsión de las posibles consecuencias no conduce necesariamente a la prevención de daños a la sociedad.

Cómo establecer la ética de la IA

La inteligencia artificial se desempeña de acuerdo con la forma en que se diseña, desarrolla, entrena, sintoniza y utiliza, y la ética de la IA trata de establecer un ecosistema de estándares éticos y límites alrededor de todas las fases del ciclo de vida de un sistema de IA.

Las organizaciones, los gobiernos y los investigadores, en conjunto, han comenzado a crear marcos para abordar las preocupaciones éticas actuales sobre la IA y dar forma al futuro del trabajo en este campo. Si bien cada día se incorporan más estructuras a estas pautas, existe cierto consenso en cuanto a incorporar los aspectos que se describen en los próximos párrafos.

Gobernanza

La gobernanza es el acto de una organización de supervisar el ciclo de vida de la IA a través de políticas y procesos internos, personal y sistemas. La gobernanza ayuda a garantizar que los sistemas de IA funcionen según lo previsto en los principios y valores de una organización, como esperan los *stakeholders* y según lo exija la normativa pertinente. Un programa de gobernanza exitoso:

- Define las funciones y responsabilidades de las personas que trabajan con IA.
- Educa a todas las personas involucradas en el ciclo de vida de la IA sobre cómo crear IA de manera responsable.
- Establecerá procesos para construir, gestionar, monitorizar y comunicar sobre la IA y los riesgos de la IA.
- Aprovecha las herramientas para mejorar el rendimiento y la confiabilidad de la IA a lo largo de su ciclo de vida·

Un comité de ética de IA es un mecanismo de gobernanza particularmente eficaz. En algunas empresas, el consejo de ética de IA está compuesto por diversos líderes de toda la

compañía. Proporciona un proceso centralizado de gobernanza, revisión y toma de decisiones para las políticas y prácticas éticas. Como ejemplo, podemos citar a la empresa de tecnología de IBM, que tiene un apartado que puede ser accedido mediante un *link* que habla de cómo nuestro compromiso con la ética, la confianza y la transparencia está diferenciando a IBM (IBM, s. f.).

Principios y áreas de interés

La estrategia de una organización en ética de AI puede guiarse por principios que se pueden aplicar a productos, políticas, procesos y prácticas en toda la organización para ayudar a habilitar una IA confiable. Estos principios deben estructurar y apoyar áreas de enfoque, como la aplicabilidad o la equidad, en torno a las cuales se pueden desarrollar estándares y alinear prácticas.

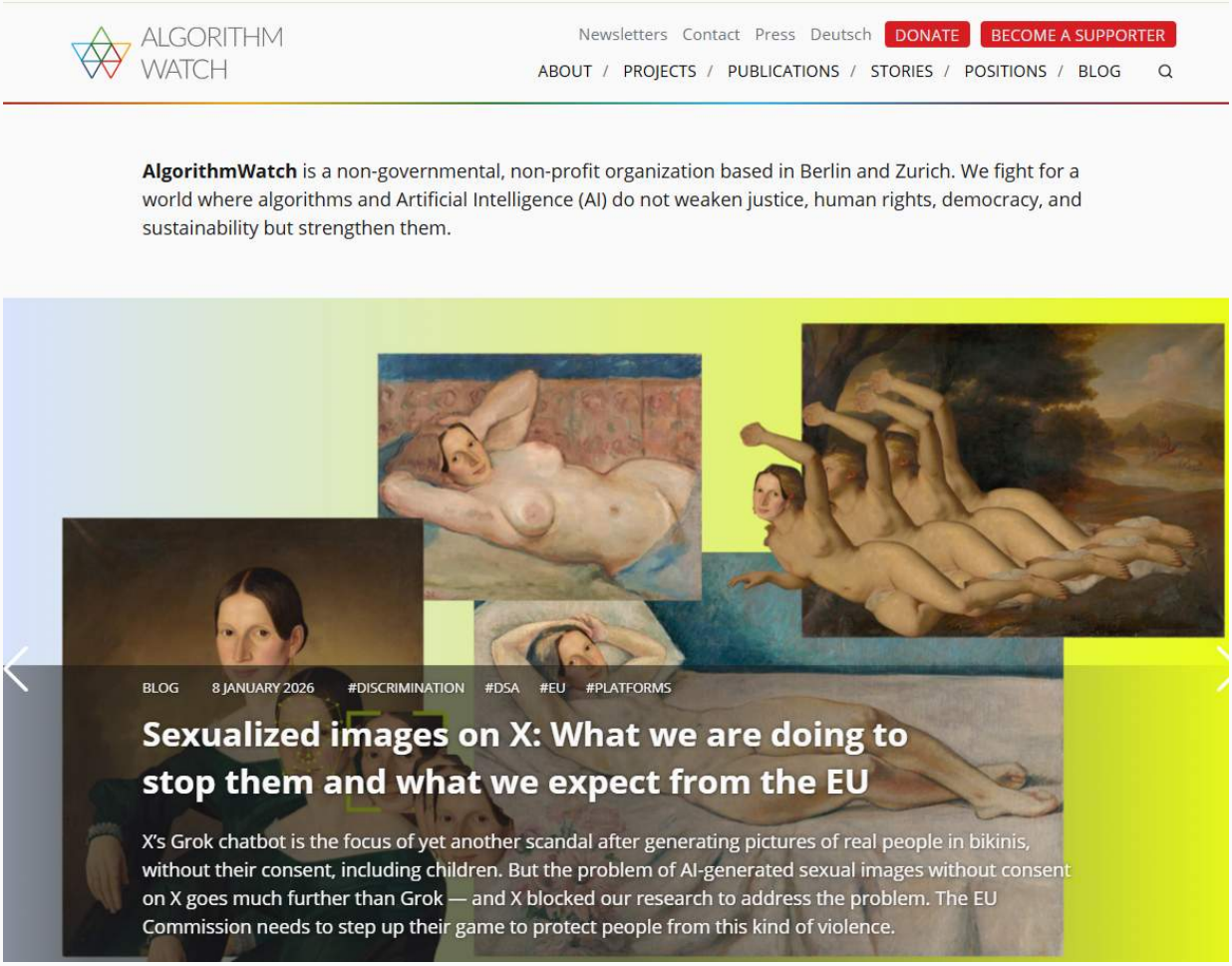
Cuando la IA se construye con la ética en el centro, es capaz de tener un gran potencial para impactar a la sociedad positivamente. Empezamos a ver esto en su integración en áreas de la salud, como la radiología. La conversación en torno a la ética de la IA también es importante para evaluar y mitigar adecuadamente los posibles riesgos relacionados con los usos de la IA, comenzando la fase de diseño.

Organizaciones que promueven la ética de la IA

Dado que los estándares éticos no son la principal preocupación de los ingenieros y científicos de datos en el sector privado, surgieron varias organizaciones para promover la conducta ética en el campo de la inteligencia artificial. Para aquellos que buscan más información, las siguientes organizaciones y proyectos proporcionan recursos para promulgar la ética de la IA.

- **AlgorithmWatch**: esta organización sin fines de lucro se centra en un algoritmo explicable y rastreable, así como en un proceso de decisión en programas de IA.

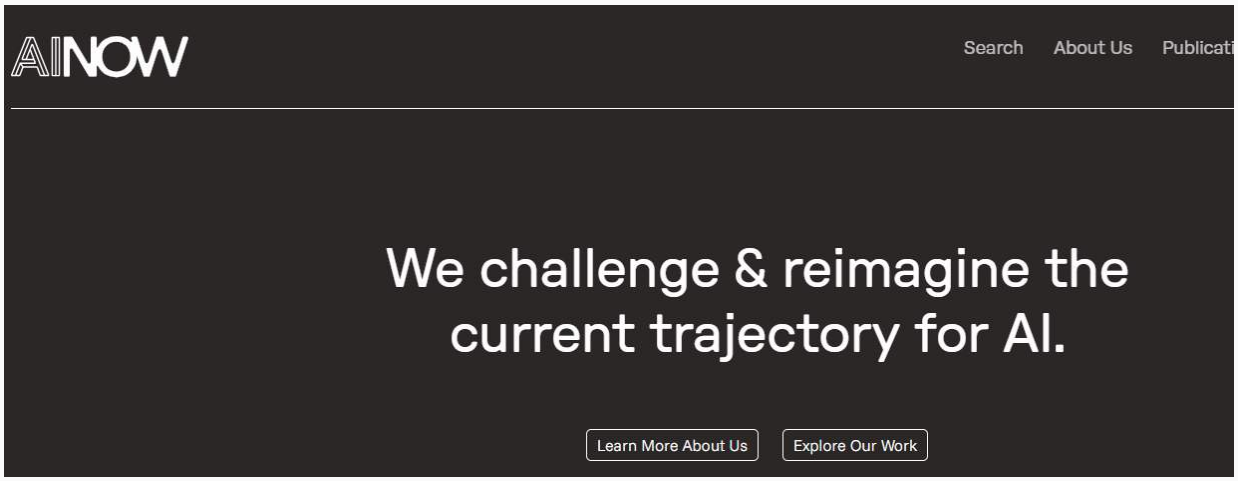
Figura 3: AlgorithmWatch



Fuente: captura de pantalla sobre AlgorithmWatch (https://algorithmwatch.org/en/).

- **AI Now Institute**: esta organización sin ánimo de lucro de la Universidad de Nueva York investiga las implicaciones sociales de la inteligencia artificial.

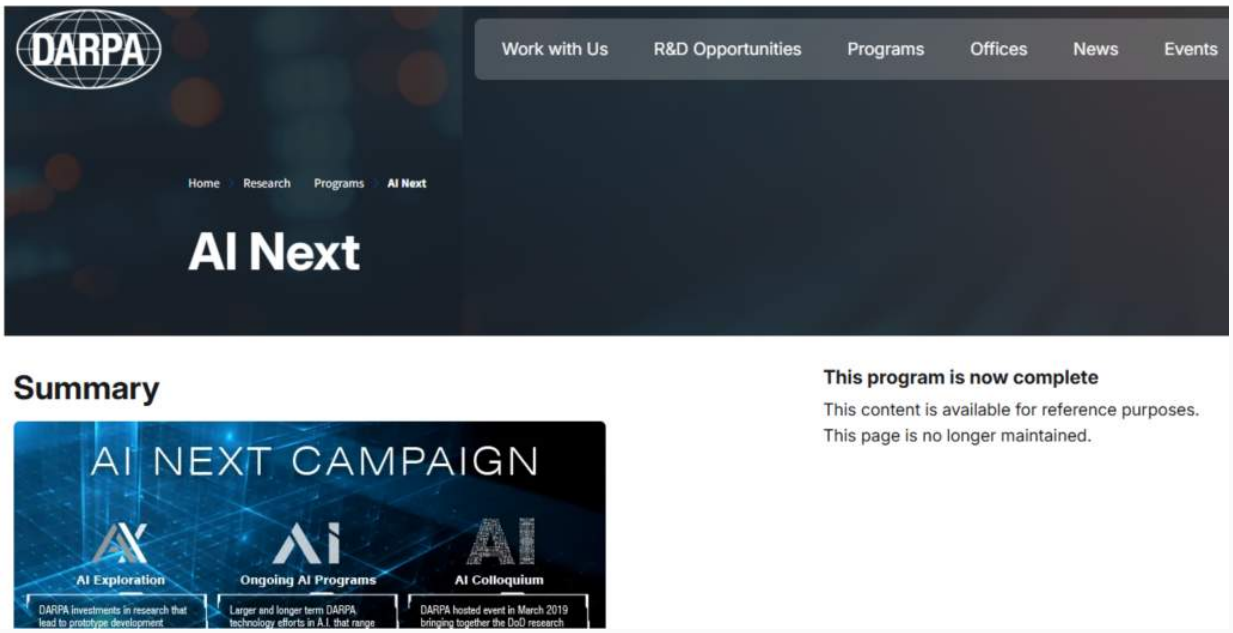
Figura 4: AINow



Fuente: captura de pantalla de AINow (https://ainowinstitute.org/).

- **DARPA**: la Defense Advanced Research Projects Agency del Departamento de Defensa de EE. UU. se centra en promover la IA explicable y la investigación sobre la IA.

Figura 5: DARPA



Fuente: captura de pantalla de DARPA (https://www.darpa.mil/research/programs/ai-next).

- **CHAI**: el Center for Human-Compatible Artificial Intelligence es una cooperación de varios institutos y universidades para promover una IA confiable y sistemas beneficiosos demostrables.

Figura 6: CHAI

Fuente: captura de pantalla de CHAI (https://humancompatible.ai/).

- **NASCAI**: la Comisión de Seguridad Nacional sobre Inteligencia Artificial es una comisión independiente que estudia los métodos y medios necesarios para avanzar en el desarrollo de la inteligencia artificial, el aprendizaje automático y las tecnologías asociadas para abordar, de forma integral, las necesidades de seguridad nacional y defensa de Estados Unidos.

Figura 7: The National Security Commission on Artificial Intelligence

Fuente: captura de pantalla de The National Security Commission on Artificial Intelligence (https://cybercemetery.unt.edu/).

Concretando el compromiso firmado en la convocatoria que hiciera en Vaticano (Call Rome), se transcriben algunos puntos de vista de la empresa de tecnología IBM sobre la ética de la IA. Esto estableció su propio punto de vista sobre la ética de la IA, al crear principios de confianza y transparencia para ayudar a sus clientes a entender dónde se encuentran sus valores dentro de la conversación sobre la IA.

Son tres los principios básicos que esta empresa presenta, los cuales dictan su enfoque hacia el desarrollo de datos e IA.

1 **El objetivo de la IA es potenciar la inteligencia humana**

Esto significa que no buscamos reemplazar la inteligencia humana con IA, sino apoyarla. Dado que cada nueva innovación tecnológica implica cambios en la oferta y la demanda de roles laborales particulares, IBM se compromete a apoyar a los trabajadores en esta transición invirtiendo en iniciativas globales para promover la formación de habilidades en torno a esta tecnología.

Los datos e *insights* pertenecen a su creador

Los clientes de IBM pueden estar seguros de que ellos, y solo ellos, son propietarios de sus datos. IBM no proporcionó ni proporcionará acceso gubernamental a los datos de los clientes para ningún programa de vigilancia, y sigue comprometida con la protección de la privacidad de sus clientes

Los sistemas de IA deben ser transparentes y explicables. IBM cree que las empresas de tecnología deben tener claro quién

entrena sus sistemas de IA, qué datos se utilizaron en el entrenamiento y, lo más importante, qué se incluyó en las recomendaciones de sus algoritmos.

IBM también ha desarrollado cinco pilares para guiar la adopción responsable de las tecnologías de IA.

- **Explicabilidad**: un sistema de IA debe ser transparente, particularmente sobre lo que se incluyó en las recomendaciones de su algoritmo, según sea relevante para los *stakeholders* con una variedad de objetivos.
- **Equidad**: se refiere al trato equitativo de las personas o grupos de personas por parte de un sistema de IA. Cuando se calibra adecuadamente, la IA puede ayudar a los humanos a tomar decisiones más justas, contrarrestar los sesgos humanos y promover la inclusión.
- **Fortaleza**: los sistemas impulsados por IA deben defenderse activamente de los ataques de adversarios, lo que reduce los riesgos de seguridad y define confianza en los resultados del sistema.
- **Transparencia**: para reforzar la confianza, los usuarios deben poder ver cómo funciona el servicio, evaluar su funcionalidad y comprender sus fortalezas y limitaciones.

- **Privacidad**: los sistemas de IA deben priorizar y salvaguardar la privacidad y los derechos de datos de los consumidores y proporcionar garantías explícitas a los usuarios sobre cómo se utilizarán y protegerán sus datos personales.

CONTINUAR

2. Ética en la ciberseguridad

Ética en la ciberseguridad

La ética en la ciberseguridad es un conjunto de principios morales que guían a profesionales y usuarios para actuar correctamente en el ciberespacio, lo cual equilibra la protección de sistemas con el respeto a la privacidad, la integridad de datos y la disponibilidad de servicios, distinguiendo a los defensores de los atacantes maliciosos y enfocándose en la responsabilidad, la transparencia y la justicia para prevenir daños y sesgos, incluso más allá del mero cumplimiento legal.

Principios éticos fundamentales

Algunos de los principios éticos fundamentales se mencionan seguidamente:

- **Confidencialidad**: proteger la información sensible del acceso, divulgación o alteración no autorizada.

- **Integridad**: asegurar que los datos sean precisos, completos y no modificados indebidamente.

- **Disponibilidad**: garantizar que los sistemas y servicios estén accesibles para los usuarios legítimos cuando los necesiten.

- **Responsabilidad**: asumir la obligación de proteger a usuarios y organizaciones, tomando decisiones éticas.

- **Transparencia**: comunicar de forma clara sobre políticas de seguridad, vulnerabilidades y uso de datos.

- **Justicia**: usar la tecnología para promover la equidad, evitando la discriminación y los sesgos algorítmicos.

- **Consentimiento informado**: obtener permiso claro y total sobre la recolección y uso de datos.

Consideraciones claves

- **Privacidad vs. seguridad**: el dilema central de equilibrar la necesidad de proteger con el derecho individual a la privacidad.

Figura 9: Privacidad vs. seguridad

Fuente: [imagen sin título sobre privacidad vs. seguridad], s. f., https://bit.ly/4tbIsD2.

- **Dilemas del profesional**: enfrentar conflictos entre intereses personales y responsabilidades profesionales, como aceptar contratos de clientes con prácticas cuestionables.

- ***Hacking* ético (sombrero blanco)**: usar habilidades de *hacking* con autorización para encontrar y corregir vulnerabilidades, no para explotarlas.

- **Minimización de datos**: recopilar solo la información estrictamente necesaria para reducir riesgos.

¿Por qué es importante?

La ética es la brújula que guía el uso responsable de la tecnología, lo que distingue a los profesionales que defienden de los atacantes y fomenta la confianza en el entorno digital, lo cual es vital para individuos, empresas y gobiernos.

Dilemas éticos en la ciberseguridad

La ciberseguridad plantea una serie de dilemas éticos que requieren una reflexión cuidadosa por parte de los responsables de tecnología de las empresas.

Algunos de los dilemas más comunes se presentan a continuación.

EL USO DE LA INTELIGENCIA ARTIFICIAL EN LA CIBERSEGURIDAD:	LA VIGILANCIA Y EL MONITOREO:	ETHICAL HACKING:
¿cómo garantizar que la IA se utilice de forma ética y responsable? ¿Cómo evitar que la IA se utilice para discriminar o invadir la privacidad de las personas?		

EL USO DE LA INTELIGENCIA ARTIFICIAL EN LA CIBERSEGURIDAD:	LA VIGILANCIA Y EL MONITOREO:	ETHICAL HACKING:
	¿dónde está el límite entre la seguridad y la privacidad? ¿Cómo equilibrar la necesidad de proteger la seguridad con el derecho a la privacidad de las personas?	

EL USO DE LA INTELIGENCIA ARTIFICIAL EN LA CIBERSEGURIDAD:	LA VIGILANCIA Y EL MONITOREO:	ETHICAL HACKING:

¿cuándo es adecuado hacer pruebas de penetración? ¿Cómo garantizar que el *ethical hacking* se emplee de forma responsable y no cause daños?

La importancia de la ética en la ciberseguridad

La ética en la ciberseguridad es importante por varias razones.

Construir confianza y reputación: las empresas que actúan de forma ética en entornos digitales generan confianza en sus clientes, socios y colaboradores.

Proteger los derechos y las libertades de las personas: la ciberseguridad debe usarse para proteger los derechos y las libertades de las personas, no para restringirlos. Las empresas y los profesionales deben respetar la privacidad, la libertad de expresión y otros derechos fundamentales.

Promover un ciberespacio seguro y responsable: la ética en la ciberseguridad es esencial para promover un entorno digital seguro y responsable para todos. Al actuar de forma ética, las empresas y los profesionales pueden contribuir a crear un entorno digital en el que todos puedan confiar.

La ética en la ciberseguridad es una responsabilidad compartida. Las empresas, los profesionales y los usuarios deben trabajar juntos para crear un espacio seguro, responsable y ético. Al tomar decisiones éticas, podemos proteger nuestros datos, sistemas e infraestructuras, al tiempo que respetamos los derechos y las libertades de las personas.

De forma más amplia, nos planteamos la siguiente pregunta: ¿Cuáles son las implicaciones éticas de la IA en la ciberseguridad?

¿Cuáles son las implicaciones éticas de la IA en la ciberseguridad?

Otro importante desafío ético en el uso de la IA en ciberseguridad es el potencial de mal uso de datos y la toma de decisiones sesgada. Los sistemas de IA se basan en grandes conjuntos de datos para aprender y tomar decisiones. Si estos conjuntos de datos contienen información sesgada o incompleta, la IA puede producir resultados inexactos o injustos.

Implicaciones éticas de la IA en la ciberseguridad

La integración de IA —incluyendo la IA generativa— en la ciberseguridad abre oportunidades, pero también plantea dilemas éticos profundos. A continuación, se presentan las principales implicaciones, sustentadas con evidencia actualizada.

Privacidad y uso responsable de datos

La IA requiere grandes volúmenes de datos para entrenarse. Estos *datasets* pueden contener información personal recopilada sin consentimiento o extraída de fuentes públicas mediante *web scraping*, lo que genera riesgos de violación de privacidad.

IBM destaca que incluso empresas legítimas pueden desarrollar modelos con datos personales sin darse cuenta, lo que provoca tensiones legales y éticas crecientes.

Transparencia, explicabilidad y rendición de cuentas

La creciente automatización de decisiones en seguridad —detección de amenazas, filtrado, respuestas automáticas— exige que estas decisiones sean trazables y comprensibles.

Según IBM (s. f.), mantener a los seres humanos informados es esencial para garantizar la responsabilidad y evitar tácticas de «sombrero gris» que desborden límites éticos.

Sesgos algorítmicos y discriminación

La IA puede amplificar sesgos presentes en los datos, lo que afecta decisiones de seguridad: clasificación errónea de tráfico, detección desigual según idioma o región, perfiles injustos, etcétera.

Investigaciones citadas en Edu News muestran que los sistemas de IA han sido acusados de reproducir sesgos humanos y generar impactos sociales negativos significativos, lo que exige nuevas reglas y marcos éticos.

Uso dual de la IA: defensa vs. ataque

El dilema ético más crítico: la IA tiene capacidad de potenciar tanto la ciberdefensa como el cibercrimen.

Observando antecedentes recientes, estos muestran que los atacantes utilizan IA para lo siguiente:

- Generar *phishing* hiperrealista.

- Crear *deepfakes*.
- Automatizar ciclos completos de ataque mediante agentes ofensivos

Esto reduce las barreras técnicas para los ciberdelincuentes y multiplica su alcance, lo que plantea un desafío ético global.

Riesgos de autonomía y decisiones no supervisadas

La aparición de ataques totalmente autónomos impulsados por IA, previstos por firmas como WatchGuard, abre el riesgo de que máquinas tomen decisiones agresivas sin intervención humana constante. Esto reconfigura la responsabilidad legal y moral en ciberincidentes.

Equilibrio entre seguridad y derechos humanos

Una IA excesivamente intrusiva en seguridad puede violar principios de proporcionalidad y libertad individual.

Los principios éticos destacados por la ONU —privacidad, justicia, gobernanza, no maleficencia— están siendo reinterpretados ante la presión de nuevas tecnologías generativas, lo que evidencia la necesidad urgente de marcos actualizados.

Regulación y gobernanza de sistemas de IA

Los gobiernos están acelerando la creación de regulaciones sobre IA y ciberseguridad, impulsadas por riesgos crecientes:

- Exigencias de transparencia.
- Auditorías de modelos.
- Cumplimiento estricto de protección de datos.

Es necesario que las regulaciones se consoliden como pilar de la estrategia institucional para frenar abusos y proteger a los ciudadanos.

La IA trae beneficios enormes en la ciberseguridad, pero también desafíos éticos ineludibles: privacidad, transparencia, sesgos, autonomía, uso dual y regulación. La clave estará en desarrollar IA responsable, a partir de la cual los intereses humanos, los derechos y la rendición de cuentas serán el centro del diseño y uso de estas tecnologías.

Para finalizar, se darán ejemplos algorítmicos de sesgos de la ciberseguridad e IA.

Sesgo de datos históricos (*bias* histórico)

Ocurre cuando el modelo aprende patrones injustos del pasado.

Ejemplo: un modelo de reclutamiento entrenado con datos de empleados de una empresa en la que históricamente se contrató más a hombres que a mujeres termina «prefiriendo» candidatos masculinos.

Sesgo de selección

El *dataset* no representa adecuadamente a todos los grupos.

Ejemplo: un sistema de reconocimiento facial entrenado mayormente con rostros de personas blancas muestra tasas de error más altas al identificar personas afrodescendientes o asiáticas.

Sesgo de etiquetado (*label bias*)

Los datos están mal o subjetivamente etiquetados.

Ejemplo: si moderadores humanos etiquetan erróneamente cierta forma de habla informal como «agresiva», un sistema de detección de toxicidad puede penalizar injustamente a comunidades que usan ese dialecto.

Sesgo de automatización

Las personas confían demasiado en la decisión de la IA, aun cuando es incorrecta.

Ejemplo: un analista de ciberseguridad acepta sin cuestionar la clasificación de un correo como «seguro» porque el sistema lo marcó así, aunque realmente era *phishing*.

Sesgo por *proxy*

El modelo aprende correlaciones indirectas que perjudican a grupos específicos.

Ejemplo: un algoritmo de crédito usa el código postal para predecir capacidad de pago. Esto afecta a residentes de zonas de bajos ingresos, aunque su historial financiero individual sea bueno.

Sesgo de representación en lenguaje

Los modelos lingüísticos aprenden asociaciones estereotipadas.

Ejemplo: relacionar profesiones como «ingeniero» con hombres y «enfermera» con mujeres, debido a patrones frecuentes en el texto de entrenamiento.

Sesgo operacional

Aparece cuando el modelo funciona bien en laboratorio, pero falla en el uso real.

Ejemplo: un detector de objetos entrenado con imágenes de buena iluminación comete muchos errores cuando se usa en cámaras de seguridad nocturnas.

Sesgo de medición

Se produce cuando las variables usadas no capturan realmente lo que se desea medir.

Ejemplo: usar historial de arrestos para predecir «riesgo criminal», sin considerar que ciertas comunidades han sido históricamente más vigiladas que otras.

CONTINUAR

Referencias

Alonso Martín, J. A. (2022). *El desembarco de la inteligencia artificial en la traducción automática*. Centro Virtual Cervantes. https://cvc.cervantes.es/lengua/anuario/anuario_22/alonso_martin/p01.htm .

Bellucci, M. (1 de febrero de 2021). Los secretos de Sophia: el robot que pretendía destruir a la humanidad y ahora la protege. *Clarín*. https://www.clarin.com/tecnologia/secretos-sophia-robot-pretendia-destruir-humanidad-ahora-protege_0_MFH1vqbKr.html.

Capelli, B. (23 de febrero de 2021). Inteligencia artificial: un compromiso común para el futuro de la humanidad. *Vatican News*. https://www.vaticannews.va/es/vaticano/news/2021-02/inteligencia-artificial-academia-pontificia-para-la-vida.html.

EuroNews. (11 de enero de 2026). *Google y la IA de Character alcanzan acuerdos de conciliación en EE.UU. por suicidios de menores*.

https://es.euronews.com/next/2026/01/11/google-y-la-ia-de-character-alcanzan-acuerdos-de-conciliacion-en-eeuu-por-suicidios-de-men.

IBM. (s. f.). *Cómo nuestro compromiso con la ética, la confianza y la transparencia diferencia a IBM*. https://www.ibm.com/mx-es/think/insights/how-our-commitment-to-ethics-trust-and-transparency-is-differentiating-ibm.

[Imagen sin título sobre humanoide Ameca]. (s. f.). https://www.perfil.com/noticias/modo-fontevecchia/ia-la-rebelion-del-robot-humanoide-mas-avanzado-que-se-declara-autoconsciente.phtml.

[Imagen sin título sobre privacidad vs. seguridad]. (s. f.). https://lockbits.cl/blog/etica-en-la-ciberseguridad/.

McCulloch, W. S., y Pitts, W. (1943). A logical calculus of the ideas immanent in nervous activity. *The Bulletin of Mathematical Biophysics*, 5(4), 115-133.

National Commission for the Protection of Human Subjects of Biomedical and Behavioral Research. (18 de abril de 1979). *The Belmont Report: Ethical principles and guidelines for the protection of human subjects of research*. US Department of Health, Education, and Welfare. https://www.hhs.gov/ohrp/regulations-and-policy/belmont-report/read-the-belmont-report/index.html.

Pangeanic. (4 de septiembre de 2018). *Traducción automática: el informe ALPAC*. https://blog.pangeanic.com/es/traduccion-automatica-informe-alpac.

Real Academia Española [RAE]. (s. f. a). *Ético*. En Recuperado el 28 de enero de 2026, de https://dle.rae.es/%C3%A9tico.

Real Academia Española [RAE]. (s. f. b). *Ciberseguridad*. En Recuperado el 28 de enero de 2026, de https://dpej.rae.es/lema/ciberseguridad.

CONTINUAR