VISUAL REVIEW | Vol. 18, No. 3, 2026 | ISSN 2695-9631
International Visual Culture Review / Revista Internacional de Cultura Visual
https://doi.org/10.62161/revvisual.v18.5893

IA GENERATIVA Y EFECTIVIDAD DE DETECCIÓN DE UNA *DEEPFAKE*
Análisis sobre la efectividad en la capacidad de detección de una deepfake entre Inteligencia Artificial y el cerebro humano.

NATHALIE ALEJANDRA RODRÍGUEZ EGAS [1]
nathalir@ucm.es
DANIEL GÓMEZ GONZÁLEZ [2]
dagomez@ucm.es
JAVIER SIERRA SÁNCHEZ [1]
javier_sierra@ucm.es
NARCISA JESSENIA MEDRANDA MORALES [3]
nmedranda@ups.edu.ec
[1] Universidad Complutense de Madrid, España
[2]Universidad Complutense de Madrid, Instituto de Estadística y Ciencia de Datos, España
[3] Universidad Politécnica Salesiana, Ecuador

PALABRAS CLAVE

Deepfake
Inteligencia Artificial
OpenAI's
Capacidad de detección
Deep learning
Redes neuronales
Suplantación de identidad

RESUMEN

La inteligencia artificial ha transformado múltiples industrias, especialmente en la generación de imágenes y vídeos. La aparición de herramientas capaces de crear vídeos hiperrealistas plantea serios riesgos en la suplantación de identidad. Este artículo examina el fenómeno de los deepfakes, junto con factores que inciden al momento de identificarlos y presenta un estudio experimental que evalúa la efectividad de la detección de estos contenidos mediante redes neuronales profundas. Se compara el desempeño de un grupo de personas con una herramienta basada en tecnología de deep learning, proporcionando un análisis detallado sobre las fortalezas y limitaciones de ambos enfoques.

Recibido: 11/ 06 / 2025
Aceptado: 15/ 09 / 2025

1. Introducción

La Inteligencia Artificial (IA) es un campo de la informática que se enfoca en crear sistemas que puedan realizar tareas que normalmente requieren inteligencia humana, como el aprendizaje, el razonamiento y la percepción (Plan de Recuperación, Transformación y Resiliencia, 2023).

Esto se refiere a sistemas digitales que pueden razonar y resolver problemas incluso con grandes cantidades de datos, facilitando tareas como crear contenido, predicciones estadísticas o automatizaciones. Se podría definir a la IA desde distintas perspectivas, según su tipología, modelos o lenguajes empleados. Según su tipología, El filósofo John Searle (1980), clasifica la inteligencia artificial en IA débil e IA fuerte, en primer lugar, define la débil como aquella que se ejecuta mediante la programación, y no posee poderes causales o intencionalidad propia, por lo cual no simula la mente, es decir que, llevan a cabo funciones específicas y con procesos distintos a los del cerebro. La IA fuerte, es catalogada por el autor como una mente en sí misma ya que considera contiene rasgos equivalentes al procesamiento del cerebro humano, autonomía para la toma de decisiones y capacidad de aprendizaje. Sin embargo, la hipótesis principal de este autor es que ningún tipo de programa o máquina puede igualar las capacidades causales del cerebro humano "Cualquier intento de crear intencionalidad artificialmente, la IA fuerte, no podría tener éxito simplemente diseñando programas, sino que tendría que duplicar las capacidades causales del cerebro humano y ningún programa por sí solo es suficiente para pensar" (p.1).

Turing (1980), quien en sus escritos afirmaba que un computador posee comportamientos inteligentes, comparándolo con desarrollo cognitivo que emula la mente humana, a lo cual denominó el juego de la imitación e intuyó la importancia que tendría el aprendizaje automático en el desarrollo de la IA. En 1956, el matemático McCarthy introduce por primera vez el término Inteligencia Artificial, durante la conferencia de Dartmouth, donde se reunieron pioneros y visionarios como: Minsky, Shannon, Rochester, entre otros. Conscientes de que la capacidad de que las computadoras de esa época eran limitadas para imitar las funciones del cerebro humano vaticinaban su evolución, mediante la automatización de los ordenadores. Ellos consideraban que los avances serían significativos y conllevarían varios problemas que sintetizaron en siete puntos: la automatización; el uso del lenguaje por computadoras; redes neuronales; eficiencia de los cálculos; el autoaprendizaje; las abstracciones; y la relación entre azar y creatividad. En la propuesta presentada definen a la IA como el aprendizaje tanto de conceptos y abstracciones que consolidan algoritmos similares a la inteligencia humana con tal precisión que puedan mejorar o resolver problemas de forma efectiva (McCarthy et al., 2006). En esta conferencia McCarthy et al. (2006), desarrolla la teoría de la información y teoría de antonomasia, donde propone crear modelos que se asemejan al cerebro humano. En concordancia con esta línea Russell y Norvig (2004), proponen una clasificación con el enfoque del modelo cognitivo, en el cual profundiza en sistemas que piensan y actúan como humanos. Todo esto bajo la teoría de la mente impulsada por Woodruff y Premack (1979) cuando habla del sistema de inferencia del cerebro humano, atribuyendo capacidades como la conciencia, intención del sistema deductivo de la mente.

Kosinski (2024), un reconocido investigador de Stanford, analizó los actuales modelos de lenguaje de IA como Chat GPT y aunque reconoce que aún somero, posee gran potencia la teoría de la mente. "No olvidemos que estamos observando un progreso exponencial, ya que los modelos de IA duplican su rendimiento cada año" afirmó Kosinski en una entrevista ofrecida a Infobae (Fernández, 2024) enfatizando en que actualmente las máquinas interpretan y captan un componente social antes no visto.

López de Mántaras y Meseguer (2017), definen a los principales modelos en Inteligencia artificial como; conexionista; evolutivo; y corpóreo. El simbólico opera con representaciones abstractas que se modelan mediante lenguajes de representación no corpóreos basados en la lógica matemática y sus extensiones, pueden ser útiles para demostrar teoremas. El modelo conexionista o también llamado IA bio-inspirada, propuesto por los autores MCculloch y Pittis en 1943 sustenta la idea de que una neurona esta esencialmente unida a la lógica. En tercer lugar, el modelo evolutivo que surge bajo los conceptos de computación evolutiva, enfocado en la mejora de los ordenadores al momento de resolver problemas para los que habían sido programados gracias a operadores y desarrollo de programas, cuyas soluciones son versiones mejoradas, sin embargo, estos autores plantean varias problemáticas y limitaciones que se explican de mejor manera en el p1.1. La limitación de la IA Generativa

Los modelos de lenguaje de IA más destacados según Douglas Eck, director de investigación senior en Google, son:

Predictiva: Aquella que utiliza un sistema de aprendizaje automático para la toma de decisiones en escenarios futuros y puede clasificar información sobre el lenguaje, sin embargo, no puede generar datos nuevos a partir de su programación.

Generativa: Se entrena con volúmenes altos de datos, un ejemplo es Gemini que está trabajando en crear imágenes a partir de texto y secuencias de automatizaciones (Google Cloud, 2025).

El aprendizaje automático propio del modelo generativo tiene la capacidad de crear algo completamente nuevo basado en la programación y datos de los cuales dispone (Eben, 2023).

Los grandes modelos de lenguaje (LLM) son un tipo de IA generativa, ya que crean nuevas combinaciones de texto en forma de lenguaje natural e imágenes.

Según sus etapas de desarrollo en el portal oficial de Google Cloud, 2025, define estas cuatro fases que son: las máquinas reactivas; las de memoria limitada; teoría de la mente; y las máquinas conscientes de sí mismas. Adaptado al contexto actual la IA débil serían las máquinas reactivas, que responde a reglas de su programación, sin embargo, no posee memoria y no aprende datos nuevos, por ejemplo, Deep Blue el ordenador programando para jugar ajedrez. En el caso de la IA Fuerte sería aquella que se puede entrenar con nuevos datos a través de redes neuronales artificiales, de forma general, como se evidencia en modelos de Memoria Limitada o MLL y basadas en aprendizaje automático (Machine Learning) y aprendizaje profundo (Deep Learning). Las otras dos categorías, no se han desarrollado en su totalidad, por lo que no existen en la actualidad, pero se trata de una IA basada en la teoría de la mente en donde simula el razonamiento de cerebro humano por su capacidad de toma de decisiones y reconocimiento de emociones. Un paso más allá de esta teoría se encuentra la IA consciente de sí misma con sentimientos y autopercepción como un ser humano (Google Cloud, 2025).

Este artículo presenta una aproximación a la capacidad de la inteligencia artificial y los seres humanos para detectar deepfakes, las limitaciones y fortalezas de ambas perspectivas. Se abordan las preocupaciones éticas, las regulaciones vigentes y los riesgos asociados con el uso de deepfakes, como la suplantación de identidad. Además, el texto destaca la creciente prevalencia de deepfakes, especialmente en contenido pornográfico no consensuado, y la necesidad de mayor alfabetización digital y herramientas accesibles para su detección. Las preguntas que plantea resolver son las siguientes:

P1. ¿Cuáles son las limitaciones de IA generativa como mecanismo de detección de deepfakes?

P2. ¿Cuál es el porcentaje de acierto que tiene un grupo objetivo a comparación de un algoritmo al momento de detectar un deepfake?

P3. ¿Qué factores pueden favorecer a la detección de deepfakes en entornos digitales?

P4. ¿Qué experimentos previos e investigaciones analizan esta problemática?

Para dar una respuesta a estas preguntas se analiza el fenómeno deepfake con cifras actuales, las teorías y estudios previos que comparan la capacidad de razonamiento y detección de un algoritmo frente al cerebro humano, que pretenden seguir una línea de investigación sobre los factores determinantes para la detección de imágenes y vídeos fotorrealistas.

1.1. La limitación de la IA Generativa

Las limitaciones más reiterativas se relacionan con los riesgos potenciales como el lenguaje tóxico y los sesgos sociales. Es imperativo asegurarse de que se han establecido las medidas necesarias de confidencialidad y acceso a los datos que se pueden facilitar al momento de utilizar estas tecnologías.

La IA también puede responder con alucinaciones, estas se producen cuando se obtienen respuestas fuera de contexto. Sus resultados se basan en los datos con la que han sido entrenadas y depende de la calidad del contenido de los mismos.

Von Neumann (1980), analiza el procesamiento paralelo de información de una máquina comparado con la inteligencia humana en un intento por entender la complejidad del cerebro humano y las funciones neuronales que han servido para la programación de máquinas y sistemas basados en algoritmos de Inteligencia Artificial, tales como lo son el Machine Learning y Deep Learning.

Existen programas que tienen la capacidad de generar conductas resolutivas inteligentes, estos son los modelos de la programación evolutiva que apuntan a inteligencias de tipo general. Respecto a esta premisa los autores López de Mántaras y Meseguer (2017) afirman lo siguiente:

> La realidad es mucho más compleja y esta aproximación tiene muchas limitaciones, (..) una de las críticas más fuertes a estos modelos no corpóreos se basa en que un agente inteligente

> necesita un cuerpo para poder tener experiencias directas con su entorno, en lugar de que un programador proporcione descripciones abstractas de dicho entorno (López de Mántaras y Meseguer, 2017, p.14).

Esto confirma la idea de que la interacción directa puede generar representaciones internas para la toma de decisiones racionales, por lo cual el agente y el entorno-contexto juegan un papel de gran importancia, a lo que se conoce como cognición situada. Lo cual se relaciona con las limitaciones de la representación y el razonamiento casados en lógica matemática. La IA actualmente no ha logrado superar estas limitaciones, sin embargo, esto ha abierto una línea de investigación dentro del subárea de la IA robótica del desarrollo.

Se ha tomado en cuenta el modelo de Arquitectura computacional expresado por el autor en su libro El ordenador y el cerebro. En donde se establecen bases para la convergencia de la neurociencia y la informática. A breves rasgos resaltamos las diferencias con mayor relevancia a este artículo son las siguientes:

1.2. Fiabilidad lógica y precisión aritmética:

Los sistemas estadísticos propios de una máquina sean digitales como analógicas, dependiendo del procesamiento de información, tienden a presentar una mayor precisión aritmética, pero a la vez los resultados pueden ser distorsionados y/o alterados por un pequeño fallo en las programaciones y ocasionar alucinaciones en las respuestas de la IA, lo que la hace menos fiable en comparación el cerebro humano, ya que este posee una mayor confiabilidad lógica (Von Neumann, 1980).

1.3. Limitaciones de la Arquitectura Serial:

"El cerebro compensa su falta de profundidad lógica explotando una amplitud lógica, es decir, el procesamiento paralelo masivo" (Von Neumann, 2012, como se cita en Kurzweil, 2012, p. 40).

En este apartado se compara el paralelismo versus el serialismo y afirma que el cerebro no puede responder con la misma velocidad y precisión que una máquina de arquitectura serial como los son los ordenadores y es por ello que la amplitud y distinción lógica de la inteligencia humana puede superar a la de las máquinas al momento de detección de información.

Una de las similitudes que encuentra Von Neumann (1980), es la llamada *prima facie* digital del sistema nervioso, en donde explica como las dendritas conducen a las neuronas y compara los impulsos nerviosos de los axones con marcadores binarios del sistema digital.

En esencia, Von Neumann (1980), admite que la arquitectura computacional y el cerebro comparten algunas similitudes, aunque equitativamente también difieren sustancialmente en precisión, fiabilidad y velocidad.

A medida que este tipo de tecnología siga avanzando no hay duda de que serán necesarias más regulaciones específicas para cada caso. La desigualdad entre el rápido crecimiento de estas nuevas tecnologías de acceso abierto o OpenAI's como lo son *Deepseek*, CHat GPT o Gemini compiten en un entorno lleno de intereses monetarios, políticos y sociales diversos por lo que su escasa regulación, combinada con la falta de conocimiento para su detección, son factores que según Smith deben ser mejorados junto con la obligación de proteger la seguridad y ciberseguridad de las naciones (The Guardian, 2023).

Sam Altman, CEO del startup ChatGPT y del generador de imágenes Dall-E2 coincide con Smith en la importancia de crear licencias y pruebas para el desarrollo de estas tecnologías.

El impacto de estas tecnologías ha ocasionado que se cree la primera regulación integral sobre inteligencia artificial. El Reglamento (UE) 2024/1689 del Parlamento Europeo y del Consejo, de 13 de junio de 2024, por el que se establecen normas aplicables en el ámbito de inteligencia artificial, que se aplica a entidades públicas y privadas. En donde se determina una serie de obligaciones y sanciones en función a niveles de riesgo, los cuatro niveles se clasifican en los siguientes:

Nivel 1: inaceptable o prohibido como, por ejemplo, la manipulación biométrica de las personas, del comportamiento humano con IA, el *Social Scoring*, sistema de identificación de emociones en el trabajo o que exploten vulnerabilidades humanas.

Se espera que esta regulación sirva para garantizar la seguridad de las personas, sin embargo, se permiten realizar ciertos experimentos en áreas controladas y con fines investigativos.

Nivel 2: de riesgo Alto se permite con requisitos estrictos como el tratamiento de datos médicos y reclutamiento en el trabajo.

Nivel 3: de riesgo limitado, se permite con transparencia por ejemplo las empresas deben presentar un aviso previo sobre las herramientas de IA que emplearán y no deberán violar la ley de derechos de autor.

En los últimos años ha aumentado la utilización de las deepfakes para la suplantación de identidad.

Solo el año 2019 se identificaron más de 900 páginas de Facebook con fotografías falsas de periodistas estadounidenses que emplearon tecnología hiperrealista de la IA para participar en tácticas de desinformación. Los administradores de dichas páginas se localizaron en Vietnam y fueron relacionados con una asociación de medios de la extrema derecha de USA, llamada Epoch Media Group. (Bhuiyan, 2023).

Gigantes tecnológicos como Google, Meta, X, TikTok y otras empresas se han comprometido al uso responsable de IA, en 2018 pasa a ser una prioridad en la empresa y ha desarrollado un programa de gobernanza y responsabilidad ética sobre la aplicación de esta tecnología. En el año 2022, a partir la pandemia, se concientiza mucho más en la importancia de detección de información falsa, implementando sistemas de verificación de información como fue el caso de Líderes tecnológicos y expertos académicos se unieron para crear el proyecto Deepfake Detection Challenge (DFDC), con el fin de desarrollar nuevas formas de detectar videos de deepfake. El conjunto de datos de más de 100.000 vídeos procedentes de 3.426 actores pagados, producidos con varios métodos basados en GAN y face-swapped o intercambio de rostro, esta investigación ha dado como resultado la construcción del conjunto de datos públicos hasta la fecha. (Dolhansky et al., 2020).

En abril del 2023, Bradford Smith, presidente de Microsoft desveló al mundo su gran preocupación acerca de la inteligencia artificial y las deepfakes.

A pesar de los esfuerzos e inversión contemplada, a inicios del año 2025, el panorama cambia cuando los CEO Musk y Zuckerberg anuncian que se eliminará su programa de verificación de información, dejando la responsabilidad al usuario de filtrar el contenido. Esto lo han justificado diciendo que desean respetar la libertad de expresión y evitar errores en la moderación de contenido (Raya, 2025).

1.4. El fenómeno deepfake. Teorías y estudios previos

Según Giansiracusa (2021) el origen del término deepfake surge en el 2017 por un usuario anónimo que utilizaba ese seudónimo en su *nickname*. Con este perfil falso, se crearon vídeos que intercambiaban el rostro de celebridades como Gal Gadot (protagonista del film Mujer Maravilla) utilizando algoritmos de deep learning. El hecho marcó un hito en cuanto a pornografía no consensuada.

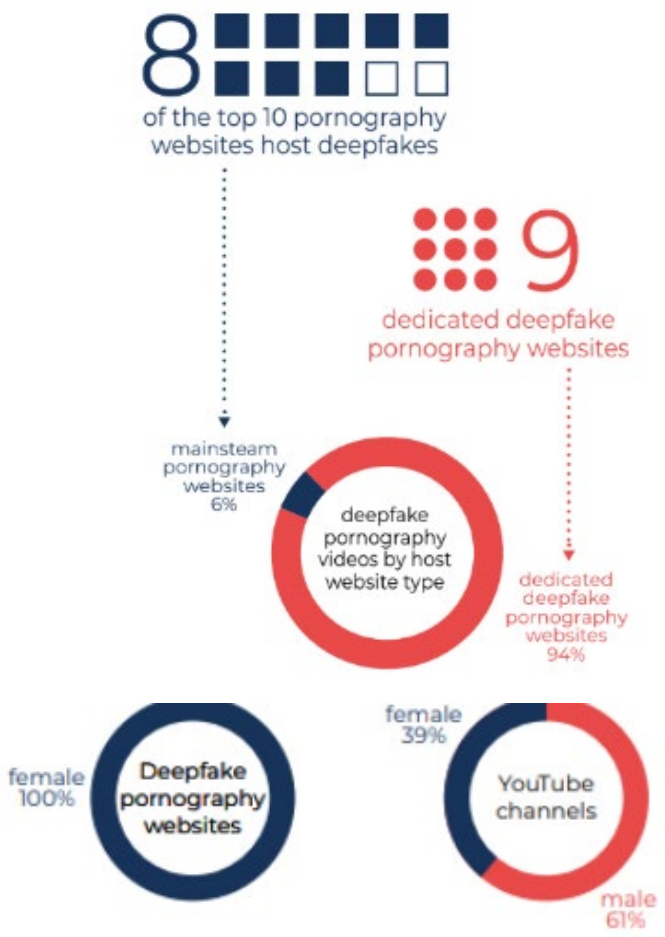
El autor clasifica los diferentes tipos de deepfakes, según su modo de empleo como: pornográfico: entretenimiento; moda; deporte; negocios y política.

Con los avances tecnológicos y la accesibilidad a una infinidad de aplicaciones, en principio de entretenimiento, ya no se requiere tener nociones de programación o ningún tipo de formación para crear deepfakes, destacando las siguientes aplicaciones de acceso abierto dedicadas a crear estos vídeos falsos como, por ejemplo: FakeApp, DFaker, faceswap, faceswap-GAN y otras.

En los últimos años el estudio sobre este tipo de tecnología se ha incrementado, sin embargo, sigue siendo un fenómeno en constante evolución, por lo tanto, existen perspectivas poco exploradas. Según los investigadores Vasist y Krishnan (2022), existen varias lagunas en las narrativas de los deepfakes, incluidas preocupaciones sobre la definición, escaso fundamento teórico, la falta de una cobertura demográfica, geográfica e integral en la literatura existente sobre esta temática. El análisis se realizó por publicaciones, afiliación y ubicación geográfica. Los índices más destacados fueron encabezados por Estados Unidos y la mayoría de investigaciones empíricas utilizan una metodología cualitativa.

Si bien este fenómeno se ha contemplado como una problemática que en su mayoría afecta a personajes públicos, como políticos, actores, influencers y otros, los avances tecnológicos se han vuelto más accesibles y sencillos de utilizar, por lo que cada vez incrementan los casos de personas comunes que han sido víctimas de suplantación de identidad. Esto lo evidencia un estudio realizado por Ajder et al. (2019), el cual determina que, de cada diez webs con carácter pornográfico, al menos ocho contienen deepfakes. Lo cual confirma que el 94% del contenido es pornográfico y continúa incrementado su presencia a través de las diferentes plataformas y sitios web como se puede apreciar en el siguiente gráfico:

Figura 1. The rise of deepfake pornography



Fuente(s): Ajder et al., 2019

En un estudio posterior Ajder (2020) determinó que el 96% de los deepfakes que aparecen en internet han sido de pornografía no consensuada, utilizando tecnología de suplantación de rostros. El incremento continúa en ascenso como lo evidenció un reporte emitido en junio del 2020, en dicho estudio clasifica a esta tecnología según su empleo y porcentaje de proliferación en entornos web con los siguientes fines: pornográficos representaban el 96%; de entretenimiento 88.9%; de moda 21%; deportes; negocios/empresarial 4.1%; política 4%. (Giansiracusa, 2021, p.46).

A menudo la creación de plataformas con tecnología IA puede generar modelos basados en algoritmos que perpetúan la desigualdad, el sesgo y conducen a resultados perjudiciales para varios sectores de la sociedad, es por ello que este artículo plantea una perspectiva integrativa que emplea un análisis empírico al momento de detectar e interactuar con deepfakes, además de anotar las convergencias perspectivas entre distintos campos y disciplinas de interés.

El avance acelerado y acceso abierto de varias aplicaciones que generan no solo imágenes hiperrealistas falsas sino también suplantación de voz ha hecho que esta tecnología se convierta en una preocupación mucho más cotidiana y que incluso afecta a poblaciones vulnerables como los jóvenes, adultos mayores.

Los casos masivos en los que se utiliza deepfakes, se han vuelto cada vez más común. En Badajoz-España durante septiembre del 2023, se presentaron una veintena de denuncias por estos vídeos falsos con desnudos de las niñas menores de edad que además de ser víctimas de suplantación de identidad, también fueron sobornadas y acosadas. Los adolescentes imputados fueron 11 niños menores de 14 años, tres de ellos manejaban la aplicación que les permitió crear los vídeos y el resto se encargó de la difusión en redes sociales y chats destinados para este fin (Pérez, 2023). Posteriormente un caso similar producido en Sevilla ha investigado a cinco jóvenes por difundir imágenes de 20 menores de edad desnudas por inteligencia artificial (Aguiar, 2024).

En Ecuador en un instituto católico, 24 estudiantes menores de edad fueron víctimas de suplantación de identidad para crear contenido de tipo sexual con IA. El hecho no fue denunciado a la fiscalía, sin embargo, el Ministerio de Educación afirmó que se establece un protocolo de acción frente a estos casos categorizados como violencia de género (Loaiza, 2023).

Estos casos evidencian la importancia de continuar investigando herramientas adecuadas a cada contexto social, que sirvan para la detección de este tipo de contenido, cómo analiza la investigadora Popova (2019), plantea una exploración etnográfica, analizando dos *websites* destinadas a deepfakes. Distingue las perspectivas sobre la preocupación de intimidad de una persona detrás de un vídeo hiperrealista falso. Según el estudio, en comparación con otras comunidades que producen contenido para adultos, las comunidades deepfake están menos preocupadas por las cuestiones de intimidad y autenticidad de la persona detrás de la imagen. Además, intentan mantener el contenido dentro de sus grupos.

2. Metodología

Para el desarrollo metodológico se empleó un análisis fenomenológico sobre la evolución de deepfakes y la aplicación de un experimento que mide la capacidad de aprendizaje y detección de algoritmos basados en deep learning, frente a las habilidades humanas, que compara el rendimiento de una forma objetiva, utilizando metodologías de detección y estudios experimentales previos que se describen en este apartado.

Como técnica aplicada se ha realizado una encuesta aplicada a un grupo objetivo en Quito-Ecuador, con una muestra de (116). Se eligió esta muestra de la población siguiendo el nivel de vulnerabilidad se analiza desde la Teoría de las Discrepancias Múltiples de Michalos (1985), que explica cómo las comparativas pueden involucrar tanto parámetros reales como empíricos, utilizando como máximo cinco estándares o categorías de distinción.

En la primera sección se midió el grado de conocimiento previo sobre deepfakes, entre otros factores contextuales como accesibilidad y limitaciones, utilizando herramientas de escalas diferenciadas y de Likert. En la segunda parte se presentó un experimento que mide la efectividad en la capacidad de detección de una deepfake entre inteligencia artificial basada en deep learning y el grupo objetivo de la población considerada.

Las hipótesis plantean

H1. La Inteligencia Artificial generativa se enfrenta a retos y limitaciones de carácter éticos y tecnológicos.

H2. La IA generativa tiene un amplio espectro de efectividad al momento de detectar imágenes falsas, sin embargo, el porcentaje de acierto puede ser menor al compararla con las habilidades de detección de una persona común.

H3. El conocimiento previo y alfabetización digital pueden favorecer a la detección de videos falsos

H4. Los experimentos y teorías relacionadas a la problemática principal pueden arrojar líneas de investigación.

2.1. Objetivos

Los objetivos a los cuales se orienta el presente artículo son los siguientes:

Analizar las limitaciones de la IA generativa como mecanismo de detección.

Analizar la efectividad en la capacidad de detección de una deepfake entre inteligencia artificial y el cerebro humano.

Determinar cuáles son los factores pueden favorecer al momento de detectar una deepfake.

Realizar una revisión sobre la literatura y experimentos relacionados.

2.2. Metodologías de detección

Se han analizado herramientas tecnológicas destinadas a la manipulación de imágenes, detección de vídeos hiperrealistas, así como el estado del arte y su evolución prospectiva, con tecnologías que utilizan modelos de reconocimiento facial o sistemas para la prevención de deepfakes como Deepware.

En cuanto a la metodología de detección está enmarcada en los sistemas de clasificación CNN utilizados por Mirsky y Lee (2021), en su artículo *The creation and detection of deepfakes: A survey*, clasifican métodos específicos de detección, los siete tipos de mecanismos que describe son los

siguientes: espaciales combinados, entorno; artefactos temporales en el comportamiento, la fisiología, sincronización y coherencia (p.28). Estos autores respaldan la teoría de que las redes neuronales profundas tienden a funcionar mejor que las herramientas forenses de imágenes tradicionales en imágenes comprimidas. Luego, varios autores demostraron cómo las arquitecturas estándar de CNN pueden detectar eficazmente vídeos deepfake. Albahar y Almalki (2019) por otro lado, clasifican los métodos de detección en reconocimiento facial, Multimedia forense; marca de agua; y detección Convolutional Neural Networks (CNNs). Según ellos uno de los métodos más sencillos/diferenciadores para detectar una *Fake news* tiene que ver con la mirada de los vídeos: en primer lugar, el ritmo y velocidad al momento de parpadear. Una persona promedio parpadea cada 2-10 segundos, con una duración de una cuarta parte o décima de segundo. En promedio las deepfakes tiene un número de parpadeo más lento. En segundo lugar, la diferenciación del color de los ojos es una técnica mediante la visión artificial, con el análisis y extracción de color.

En general, los investigadores han adoptado uno de dos enfoques: clasificación o detección de anomalías. En este caso, se utiliza el enfoque de detección de anomalías bajo la se evidencia de que a diferencia de una red completamente conectada (densa), una red neuronal convolucional (CNN) aprende jerarquías de patrones en los datos y, por lo tanto, es mucho más eficiente en el manejo de imágenes.

Los desafíos de detectar imágenes generadas por IA han hecho que empresas como Sensity presenten algoritmos de detección avanzada a tiempo real que analizan señales visuales y contextuales para identificar imágenes generadas por IA. Estos algoritmos pueden detectar artefactos sutiles, inconsistencias y patrones que son indicativos de manipulación de IA, incluso en ausencia de la verdad o datos de referencia. (Sensity, 2023). Esta empresa pionera además presentó un análisis donde desvela que los sistemas y tecnologías de reconocimiento facial y de identidad digital son vulnerable a la amenaza en un 95%, por lo que aún con los avances en el desarrollo de algoritmos de detección tanto el reconocimiento facial como la detección siguen estando en desventaja frente la tecnología deepfake (Sensity, 2022, p.3).

2.3. Estudios experimentales previos

Se ha tomado como referencia un experimento de detección (Groh et al., 2022), en el que se comparó la capacidad de detección con una herramienta basada en un algoritmo CNN y un grupo de personas para determinar el rendimiento de la detección. Se realizaron dos experimentos, uno en el que a las personas se les mostró una versión real y una versión falsa del mismo video y se les pidió que decidiera cuál era real y cuál era un deepfake sin conocimiento previo. En la segunda parte se les dio la oportunidad de cambiar su respuesta si así lo deseaban tras conocer la respuesta de la herramienta de detección. En ambos experimentos, a los humanos también se les mostraron desencadenantes emocionales aleatorios antes de que pudieran responder. Los resultados mostraron que, en el primer experimento, el 82% de las personas obtuvieron mejores resultados que la herramienta, que tuvo un rendimiento del 65% de precisión.

En un más reciente estudio realizado por Lake y Baroni (2023) determinaron que la IA puede fácilmente igualar o mejorar las habilidades humanas de aprendizaje e incluso desarrollar capacidades nuevas, esto lo han dilucidado tras realizar una red neuronal desarrollada bajo una técnica que han denominado como Meta-learning for Compositionality (MLC), aseguran que puede superar la capacidad de aprendizaje y detección de cualquier otra IA.

El siguiente experimento que resultó de gran utilidad para esta investigación, es el llamado “In Event of Moon Disaster” realizado por el Instituto Tecnológico de Massachusetts-MIT (2020), en donde se pone a prueba la capacidad de detección de un gran número de personas. Es un vídeo deepfake donde se puede apreciar al presidente Nixon anunciando que el aterrizaje del Apolo 11 resultó fallido y los astronautas murieron. La finalidad de este experimento es que el espectador decida si el vídeo es real o no y a su vez analiza los factores impulsan esta respuesta.

3. Análisis y resultados

En la primera fase de la encuesta los datos arrojados han sido considerados de acuerdo a las siguientes variables:

Grado de reconocimiento de deepfakes en jóvenes.

Accesibilidad de herramientas y aplicaciones de detección de deepfakes.

Factores inciden en tu capacidad de detección de deepfakes.

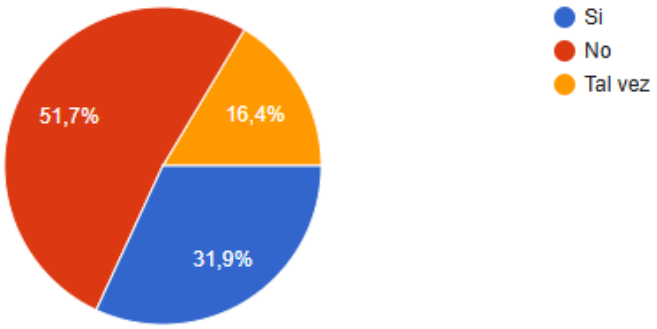
El Porcentaje de efectividad que tiene un Algoritmo comparado con humanos al momento de detectar deepfakes.

Las técnicas de medición empleadas incluyen la escala de Likert, junto con una matriz de lado a lado para evaluar el grado de incidencia de los factores de información en el reconocimiento de deepfakes.

En la segunda parte de la encuesta se plantea un experimento que busca analizar la capacidad de detección de deepfakes por parte del grupo objetivo, comparado con un IA basada en Deep Learning.

Los encuestados respondieron que no disponían de conocimiento previo en un 51,7% frente a un 31,9% que sí había escuchado hablar sobre el tema.

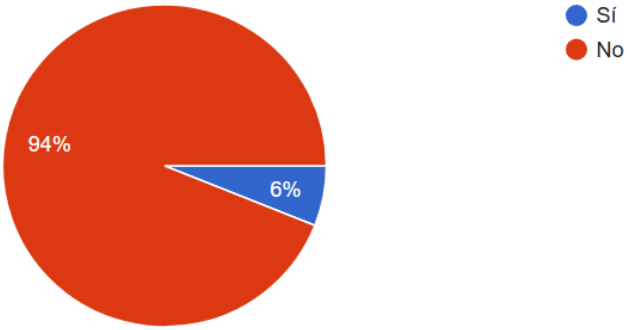
Figura 2. Conocimiento previo



Fuente(s): Elaboración propia, 2025.

A la pregunta ¿Has recibido información de cómo detectar o prevenir deepfakes en tu entorno cercano o académico? Solo un 6% afirmó que había recibido, frente a un 94% que respondieron que no.

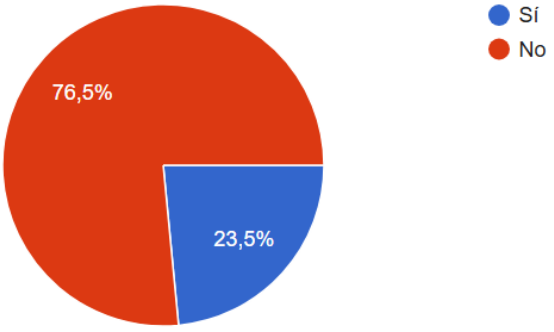
Figura 3. Autopercepción de conocimiento de detección



Fuente(s): Elaboración propia, 2025.

Relacionado a la variable de accesibilidad el 76,5% aseguró no tener a su alcance información sobre herramientas de detección.

Figura 4. Accesibilidad



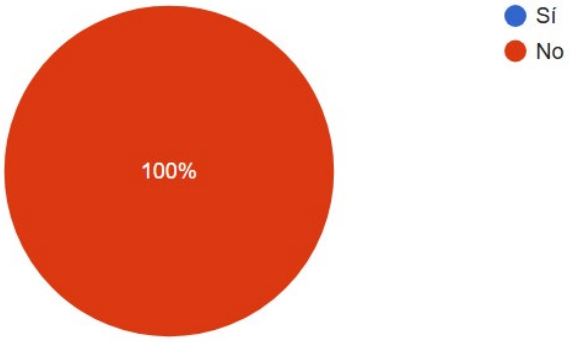
Fuente(s): Elaboración propia, 2025.

Para los encuestados los factores más influyentes son tener acceso a plataformas de detección y los de menor incidencia son los factores socioculturales. Esta pregunta solicitaba ponderar en la escala del 1 al 5 el grado de incidencia en la capacidad de detección de deepfakes. (Siendo 1 el menos influyente y 5 el más alto) y este fue el resultado:

5) Accesibilidad a plataformas de detección
4) conocimiento previo
3)Alfabetización Digital
2) Polarización de información
1) Factores socioculturales

El grado de preocupación determinó que el 49% teme que su intimidad se vea afectada por suplantación de identidad. Un 100% aseguró no saber cómo actuar en caso de verse afectado por una deepfake.

Figura 5. Protocolo de acción

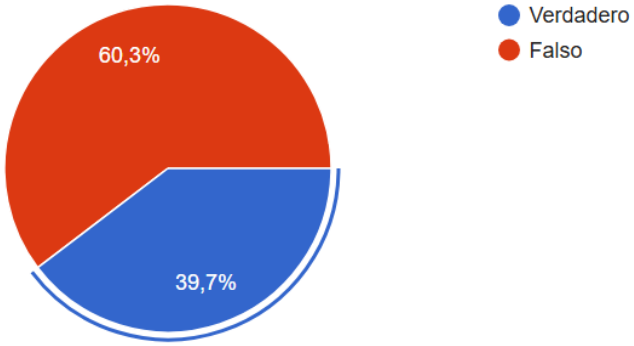


Fuente(s): Elaboración propia, 2025.

En la segunda parte se presentó a los encuestados un vídeo que deberían analizar y responder si consideran que es verdadero o falso, justificando la respuesta. El vídeo utilizado fue el mismo del experimento "In Event of Moon Disaster" y arrojó los siguientes resultados:

El 60,3% aseguró que el vídeo es falso, mientras que el 39,7% dijo que es verdadero:

Figura 6. Resultados experimento

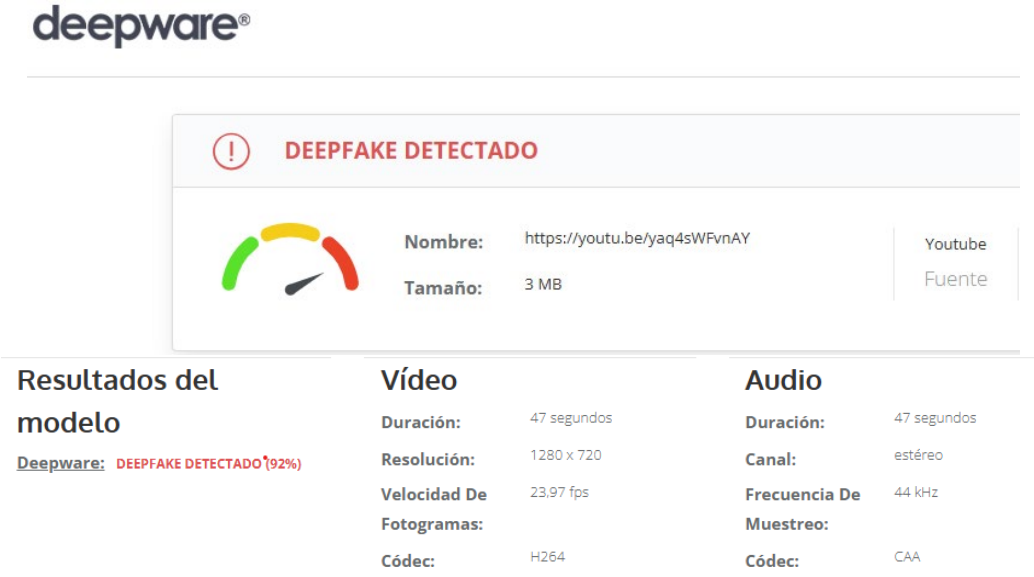


Fuente(s): Elaboración propia, 2025.

Justificando esta respuesta, los elementos que destacaron los encuestados para identificar la veracidad del vídeo fueron en primer lugar la calidad del audio y vídeo; seguido por la fuente o medio de difusión y el número de parpadeos.

Al comparar la efectividad de detección de la herramienta Deepware con tecnología de deep learning, el resultado arrojado sobre el vídeo fue que es un 92% falso, además de identificar la fuente, calidad, resolución y fecha de publicación con gran precisión únicamente con la URL indicada:

Figura 7. Herramienta de detección.



Fuente: Deepware, 2025.

Sin embargo, el resultado tiene un mensaje “Deepware tiene como objetivo dar una opinión sobre el video escaneado y no es responsable del resultado. Dado que Deepware Scanner aún está en versión beta, los resultados no deben tratarse como una verdad o evidencia absoluta.” (deepwarware, 2025).

4. Conclusión

Los resultados arrojados respecto a las limitaciones y fortalezas de IA generativa como mecanismo de detección de deepfakes determinan que, las fortalezas más destacadas son: precisión aritmética, capacidad de procesamiento de datos y velocidad. El procesamiento de grandes cantidades de imágenes y vídeos, junto con el reciente progreso de deep learning ha permitido que estas herramientas aprendan representaciones cada vez más complejas, presentes en los pixeles que a nivel de efectos audiovisuales produce buenos resultados de reconocimiento.

En el caso de esta tecnología, las redes neuronales con las que emplea sus comandos pueden resultar similares a las humanas por la capacidad de combinar y crear nuevos aprendizajes sistemáticos. Sin embargo, los de la IA dependen de la complejidad y sesgo de los datos con los que se programe. El procesamiento del lenguaje natural con métodos matemáticos y estadísticos es frágil al momento de identificar abstracciones contextuales, socioeconómicas, entre otras y puede fallar en muchos casos en el tratamiento de ambigüedades ya que no existe una comprensión profunda del mensaje, o de la escena completa por falta de sentido común y la profundidad lógica propia del ser humano. Una manera de mejorar el modelo de lenguaje es entrenarlo con más datos, aunque aun así siguen existiendo limitaciones de caracteres éticos, contextuales, lógicos y de regulación, por lo cual se posiciona a la capacidad humana por encima de la IA.

El consenso de la red neuronal profunda, respecto a la identificación ha sido mayor con un 92%, mientras que las habilidades de detección de una persona común fueron del 60,3%. Cabe destacar que en ambos casos los resultados han sido correctos, al lograr identificar el vídeo como falso.

Como lo describe la leyenda de la propia plataforma (Deepware) no se puede considerar como un resultado verídico ya que aún se encuentra en desarrollo y continuo aprendizaje por lo que puede generar alucinaciones o falsos positivos o negativos.

la autopercepción de conocimiento e identificación de deepfakes el 31% coincide con el porcentaje de encuestados que no consideran tener conocimiento previo. En las hipótesis se plantearon los elementos de conocimiento previo y alfabetización digital como factores que pueden favorecer a la identificación de vídeos falsos, sin embargo, los resultados indicaron que el grupo objetivo a pesar de no tener ningún conocimiento sobre el tema, aun así, logró identificar que era falso, valiéndose de dos componentes: la calidad del vídeo y la fuente de donde viene la información.

El tener acceso a plataformas de código abierto o gratuitas fue una de las variables mejor ponderadas por los encuestados y por lo tanto es preocupante que la población no tenga acceso a

información de cómo detectar vídeos falsos y tampoco de protocolos que le permitan saber actuar en un caso de suplantación de identidad, los casos de víctimas analizados en el contexto actual dilucidan la problemática que se extiende de forma acelerada, por lo tanto, es imperante tener accesibilidad y recursos que faciliten la identificación precoz de vídeos falsos.

La creciente problemática de los deepfakes, especialmente en casos de pornografía no consensuada y suplantación de identidad que afectan a poblaciones vulnerables, resalta la urgencia de establecer mayor accesibilidad a recursos gratuitos y protocolos para su detección.

Finalmente se confirma que los experimentos y teorías relacionadas a la problemática principal permiten seguir desarrollando líneas de investigación relacionadas a este fenómeno deepfake y así analizar los factores que pueden favorecer la detección de este contenido malicioso. Este artículo presenta resultados previos a una tesis doctoral que realiza un análisis comparativo entre dos grupos objetivos de Ecuador y España para determinar, mediante una etnografía digital, factores que pueden favorecer una detección precoz de deepfakes.

5. Agradecimientos

Daniel Gómez González agradece por el financiamiento del Gobierno de España mediante el Plan I+D+i del proyecto [PID2021-122905NB-C21]

Nathalie Alejandra Rodríguez Egas agradece a la Universidad Politécnica Salesiana de Ecuador, a la Carrera y estudiantes de Diseño Multimedia por ser parte de esta investigación, a su directora de la estancia internacional. Un especial agradecimiento a sus directores de tesis por su apoyo y acompañamiento y a la Universidad Complutense de Madrid.

Referencias

Aguiar, A. R. (2024). La Guardia Civil detecta un nuevo escándalo de menores 'desnudadas' con IA en España: cinco jóvenes, investigados en Sevilla. *Business Insider*. Retrieved 12 04, 2024, from https://www.businessinsider.es

Ajder, H. (2019). Scams And Sabotage: Why Deepfakes Pose An Unprecedented Threat To Businesses. *Medium*. Retrieved October 6, 2023, from https://acortar.link/EHLBz8

Ajder, H. (2020). Tracer Newsletter 58 (09/07/20)-Deepfake Threat Intelligence: A statistics snapshot from June 2020. *Medium*. Retrieved October 6, 2023, from https://acortar.link/D6bA7d

Albahar, M., & Almalki, J. (2019). Deepfakes: Threats and countermeasures systematic review. *Journal of Theoretical and Applied Information Technology*, 97(22), 3242-3250.

Bhuiyan, J. (2023, May 16). OpenAI CEO calls for laws to mitigate 'risks of increasingly powerful' AI. *The Guardian*. https://lc.cx/vE8-g8

Dolhansky, B., Bitton, J., Pflaum, B. L., Howes, R., Wang, M., & Ferrer, C. C. (2020). *The deepfake detection challenge (dfdc) dataset*. arXiv preprint arXiv:2006.07397.

Eben, C. (2023, April). Ask a Techspert: What is generative AI? Google. https://lc.cx/yD6o_fGiansiracusa,

Fernández, M. (2024, Noviembre). *La teoría de la mente: un experimento probó que la IA tiene una capacidad humana que se creía imposible*. Infobae. https://goo.su/SXhUzau

Giansiracusa, N. (2021). *How Algorithms Create and Prevent Fake News: Exploring the Impacts of Social Media, Deepfakes, GPT-3, and More*. Apress. https://doi.org/10.1007/978-1-4842-7155-1

Google Cloud. (2025, Marzo). *¿Qué es la inteligencia artificial (IA)?* .https://lc.cx/0YFGOG

Groh, M., Epstein, Z., & Picard, R. (2022). Deepfake detection by human crowds, machines, and machine-informed crowds. *Proceedings of the National Academy of Sciences*, 119(1), e2110013119. https://doi.org/10.1073/pnas.2110013119

Kosinski, M. (2024). Evaluating large language models in theory of mind tasks. *Proceedings of the National Academy of Sciences*, 121(45). https://doi.org/10.1073/pnas.2405460121

Lake, B. M., & Baroni, M. (2023). Human-like systematic generalization through a meta-learning neural network. *Nature*, 623(7985), 115-121. http://dx.doi.org/10.1038/s41586-023-06668-3

Loaiza, Y. (2023, October 5). Estudiantes de un colegio de Quito utilizaron fotografías de sus compañeras para crear contenido sexual con inteligencia artificial. *Infobae*. Retrieved November 22, 2024, from https://lc.cx/nO2NO0

López de Mántaras, R., & Meseguer, P. (2017). *Inteligencia artificial.* CSIC Los Libros de la Catarata.

Massachusetts Institute of Technology. (2020). Tackling the misinformation epidemic with In Event of Moon Disaster. *MIT News Office*. https://lc.cx/tUb3GL

McCarthy, J., Minsky, M. L., Rochester, N., & Shannon, C. E. (2006). A proposal for the Dartmouth summer research project on artificial intelligence, August 31, 1955. *AI magazine*, 27(4), 12-42.

Michalos, A. C. (1985). Multiple discrepancies theory (MDT). Social indicators research, 16, 347-413.

Mirsky, Y., & Lee, W. (2021). The Creation and Detection of Deepfakes: A Survey. *ACM Comput*, 54(1), 1-41. https://doi.org/10.1145/3425780

OpenAI. (2025). *Deepware* (versión del 07 de febrero) [Modelo multimodal grande]. https://scanner.deepware.ai/

Pérez, E. (2023, September 18). Falsos desnudos de menores generados por IA: la Policía investiga en Almendralejo el primer caso masivo en España. *Xataka*. Retrieved September 27, 2023, from https://lc.cx/HYGdfz

Plan de Recuperación, Transformación y Resiliencia. (2023, April 19). *Qué es la Inteligencia Artificial*. Retrieved March 4, 2025, from https://planderecuperacion.gob.es/noticias/que-es-inteligencia-artificial-ia-prtr

Popova, M. (2019). Reading out of context: Pornographic deepfakes, celebrity and intimacy. *Porn Studies*, (7). https://doi.org/10.1080/23268743.2019.1675090

Raya, A. (2025, January 21). Elon Musk y Mark Zuckerberg juegan a dos bandas: X y Meta mantendrán la lucha contra las noticias falsas en la UE. *El Español*. https://surl.li/jkukcq

Reglamento (UE) 2024/1689 del Parlamento Europeo y del Consejo, de 13 de junio de 2024, por el que se establecen normas armonizadas en materia de inteligencia artificial y por el que se modifican los Reglamentos (CE) n.° 300/2008, (UE) n.° 167/2013, (UE) n.° 168/2013, (UE) 2018/858, (UE) 2018/1139 y (UE) 2019/2144 y las Directivas 2014/90/UE, (UE) 2016/797 y (UE) 2020/1828

(Reglamento de Inteligencia Artificial). Diario Oficial de la Unión Europea, de 13 de junio de 2024. http://data.europa.eu/eli/reg/2024/1689/oj

Russell, S. J., & Norvig, P. (2004). *Inteligencia artificial: un enfoque moderno*. Pearson Educación.

Searle, J. R. (1980). Minds, brains, and programs. Behavioral and Brain Sciences, 3(3), 417–424. doi:10.1017/S0140525X00005756

Sensity. (2022). Deepfakes vs biometric KYC verification [Report]. *Sensity AI*. https://sensity.ai/reports/

Sensity. (2023, May 10). How to detect AI generated images with Sensity in 2023. Sensity AI. Retrieved April 2, 2024, from https://lc.cx/taJOpi

The Guardian. (2023, May 25). Deepfakes are biggest AI concern, says Microsoft president. The Guardian. https://lc.cx/inUT0b

Turing, A. M. (1980). Computing machinery and intelligence. *Creative Computing*, 6(1), 44-53.

Vasist, P., & Krishnan, S. (2022). Deepfakes: An Integrative Review of the Literature and an Agenda for Future Research. *Communications of the Association for Information Systems*, 51(556), 556-557. 10.17705/1CAIS.05126

Von Neumann, J. (1980). *El ordenador y el cerebro*. Antoni Bosch.

Von Neumann, J. (2012). *The computer & the brain* (R. Kurzweil, Foreword; 3ra ed.). Yale University Press. https://lc.cx/5WsZ-M

Woodruff, G., & Premack, D. (1979). Intentional communication in the chimpanzee: The development of deception. *Cognition*, 7(4), 333-362.