Universidad Ean

Facultad de Ingeniería

Influencia de los Deepfakes en la Confianza Pública y la Credibilidad de la Información Digital en Colombia

Estefany Del Carmen Saavedra Camacho

Trabajo de grado presentado como requisito para optar al título de:

Especialista en Gerencia de Tecnología

Profesora:

Luisa Fernanda Carvajal Díaz

Unidad de estudio:

Seminario de Investigación

Bogotá D.C, Colombia

22/04/2026

La confianza pública y la credibilidad de la información digital.

Planteamiento del Problema

En los últimos años, el desarrollo de la inteligencia artificial ha propiciado numerosos avances tecnológicos, entre ellos la generación de contenido audiovisual altamente realista. Dentro de estas tecnologías se destaca el *deepfake*, fenómeno que lleva a cuestionar la autenticidad de imágenes y videos disponibles en redes sociales. Según EFE (2023), los *deepfakes* progresan más rápido que las herramientas diseñadas para detectarlos, siendo China el primer país en regularlos.

Serrano et al., (2025) indican que los *deepfakes* representan un vacío legal en la legislación de diversos países, pues pocos cuentan con marcos normativos preparados para proteger a los ciudadanos frente a la falsificación de identidad.

Esta situación evidencia una problemática que desborda la capacidad de respuesta de las entidades regulatorias a nivel mundial. Al respecto, Cotino (2022) explica que imponer sanciones por la difusión de información falsa en internet, especialmente en redes sociales, y por profesionales de la comunicación supone reconocer el incumplimiento de las obligaciones propias de su labor.

El término *deepfake* hace referencia al uso de imágenes manipuladas mediante inteligencia artificial para generar videos e imágenes de apariencia extremadamente realista. Este fenómeno afecta principalmente a personas de relevancia pública, políticos, actores, *influencers* y empresarios, así como a ciudadanos comunes que son engañados a través de las redes sociales. La frecuencia de este tipo de contenido aumenta progresivamente, dificultando cada vez más la distinción entre lo real y lo artificialmente generado. Un caso ilustrativo fue el video generado con inteligencia artificial que simuló un golpe de Estado contra el presidente francés Macron, demostrando el potencial de los *deepfakes* para viralizarse y erosionar la confianza ciudadana en la información audiovisual disponible en redes sociales (Euronews, 2025).

El impacto de los *deepfakes* en la confianza pública es profundo. Parra y Pinela (2025) destacan que “la pérdida de credibilidad de los medios digitales se ve acelerada por la capacidad de los *deepfakes* de minar el pensamiento crítico, lo cual se vuelve un mayor reto para cualquier medio incluso los medios tradicionales” (p. 2100). Además, este deterioro de la credibilidad de la información que ven o escuchan en internet, puede llevar a que las personas se vuelvan más escépticas a incluso frente a información veraz, generando un ambiente de desconfianza que afecta tanto a los medios digitales como los tradicionales.

La confianza pública y la credibilidad de la información digital.

A su vez, Suastegui et al., (2026) advierten que la inteligencia artificial generativa ha incrementado de forma acelerada la producción de *deepfakes* y sus repercusiones en la confianza social, configurándose como una amenaza creciente para la credibilidad de la información digital.

La desinformación constituye un fenómeno en expansión, favorecido por la sobrecarga informativa que impide la toma de decisiones racionales y por la tendencia de contenido superficial orientado a captar la atención de la audiencia. En este contexto, los *deepfakes* se convierten en un vector especialmente efectivo de desinformación (Serrano et al., 2025).

Esta problemática motiva el presente estudio, cuyo propósito es analizar cómo los *deepfakes* afectan la percepción de confiabilidad de la información digital y sus efectos en la confianza pública, reconociendo al mismo tiempo las aplicaciones legítimas de esta tecnología en el entretenimiento, la educación y la producción audiovisual (Ballesteros & Ruiz, 2024).

La confianza pública y la credibilidad de la información digital.

Pregunta de Investigación

¿Cuál es el efecto de los *deepfakes* en la confianza pública y en la credibilidad de la información digital en Colombia en la actualidad?

Objetivos

Objetivo General

Evaluar el impacto de la tecnología *deepfake* en la confianza pública y en la credibilidad de la información digital en Colombia.

Objetivos Específicos

- Determinar el nivel de conocimiento y percepción que tienen los usuarios de medios digitales sobre los *deepfakes* y su capacidad para identificar contenido manipulado.
- Analizar la relación entre la exposición a *deepfakes* y el nivel de desconfianza hacia los medios digitales.
- Identificar las estrategias de verificación, que emplean los usuarios colombianos para hacer frente a la desinformación generada por *deepfakes* en el entorno digital.

La confianza pública y la credibilidad de la información digital.

Justificación

La presente investigación resulta relevante desde los ámbitos tecnológico, social y ético, y surge del cuestionamiento ante la falta de control sobre el contenido *deepfake* presente en medios de comunicación, redes sociales y páginas web.

La investigación analiza el impacto de los *deepfakes* en Colombia, visibilizando la importancia de reconocer contenido manipulado e identificar los principales riesgos sociales a los que se expone la ciudadanía en el entorno digital. El estudio se desarrolla en Colombia, considerando el crecimiento en el uso de medios digitales y la vulnerabilidad frente a la desinformación generada por los *deepfakes* (Parra & Pinela, 2025). En este contexto, se hace necesario fortalecer las estrategias de educación digital para promover el desarrollo de habilidades que permitan adoptar una postura reflexiva frente a los contenidos manipulados por la inteligencia artificial, con el fin de reducir el impacto de la desinformación digital y así fomentar el uso responsable de las tecnologías emergentes.

La vulnerabilidad del contexto colombiano resulta especialmente significativa. Un informe de Microsoft citado por Infobae (2025) revela que el 78% de los usuarios de internet en Colombia ha estado expuesto a algún tipo de riesgo en línea. Dentro de los riesgos identificados, la desinformación lidera con un 59% de incidencia, seguida por la exposición a contenidos violentos (47%) y el discurso de odio (39%). Particularmente alarmante es el aumento del 20% en *deepfakes* pornográficos en el país, que lo posiciona como el segundo con mayor incidencia entre los 15 países analizados. Adicionalmente, el 82% de los colombianos expresó preocupación por el uso indebido de la inteligencia artificial generativa, lo que evidencia una conciencia creciente sobre la amenaza, aunque sin los mecanismos suficientes para hacerle frente.

Ballesteros y Ruiz (2024) señalan que, aunque la inteligencia artificial ofrece oportunidades en campos como el artístico y el audiovisual, también representa un gran desafío para la sociedad, al dificultar la distinción entre lo real y lo creado por la inteligencia artificial, lo que la convierte en una herramienta con gran capacidad de crear contenido engañoso. Esta dualidad hace necesaria una reflexión crítica sobre los límites éticos y sociales de su uso. En este sentido, resulta fundamental desarrollar en los usuarios la capacidad crítica y ética que les permita interpretar los contenidos que consumen en los medios digitales y enfrentar los desafíos derivados de estas tecnologías.

En tecnología, los *deepfakes* muestran la urgencia de desarrollar estrategias de regulación y alfabetización digital que fortalezcan las habilidades críticas ciudadanas, especialmente en población joven. Un abordaje integral desde la Alfabetización Mediática

permite gestionar la tecnología inmersiva, optimizando sus ventajas y reduciendo riesgos (Sánchez-Acedo et al., 2024). Además, fortalecer estos aspectos ayuda a promover un consumo responsable y facilitar la identificación de contenidos manipulados.

Desde la perspectiva del sector financiero, Asobancaria (2025) advierte que los *deepfakes* representan una amenaza importante para la seguridad digital y la integridad de datos en Colombia. El informe señala que los ciberdelincuentes emplean estas tecnologías para evadir mecanismos de autenticación biométrica y engañar tanto a usuarios como a entidades bancarias, configurándose como una de las amenazas más complejas en el ecosistema digital financiero del país. Esta realidad exige que las instituciones financieras y los usuarios adopten medidas preventivas para mitigar los riesgos de suplantación de identidad con inteligencia artificial, y evidencia la urgencia de desarrollar marcos regulatorios y capacidades técnicas de respuesta a nivel nacional.

La confianza pública y la credibilidad de la información digital.

Marco Teórico

Definición y Contexto de los *Deepfakes*

Los *deepfakes* son contenidos audiovisuales, como imágenes, videos y audios, manipulados mediante herramientas de inteligencia artificial que permiten el intercambio de rostros y la modificación de la voz, de modo que resulta difícil distinguir si son reales o falsos. (Garriga et al., 2024). Lo anterior se ha popularizado con el surgimiento cada vez más común de nuevas inteligencias artificiales cuyas capacidades son más avanzadas y dificultan la identificación de estos archivos modificados.

Dentro de los *deepfakes* se encuentran categorías tales como:

Deepfaces

Consisten en crear imágenes o videos desde cero, o manipular contenido existente de medios confiables.

Deepvoices

En este caso se busca suplantar la voz de una persona, o modificar un audio existente para agregar conversación inexistente.

Argumentación

Según como lo menciona Serrano et al., (2025):

> La proliferación de *deepfakes* no se limita a un ámbito geográfico específico; sin embargo, su impacto puede variar considerablemente según el contexto social, económico y cultural de cada región. A nivel global, la adopción creciente de tecnologías digitales y la presencia significativa en plataformas de redes sociales hacen que los *deepfakes* representen una amenaza emergente que requiere una atención particular (p. 20).

Complementando la visión de Serrano, es importante analizar cómo la brecha digital en regiones en desarrollo afecta la capacidad de respuesta ante este tipo de contenido, especialmente en territorios donde relativamente hace poco no tenían conexión a internet. En la actualidad, personas de diferentes grupos sociales tienen acceso a internet y consumen más de una red social de forma frecuente. En grupos sociales con menor acceso a educación de buena calidad (o educación inexistente), o grupos sociales de bajos ingresos se puede percibir que muchas personas desconocen lo que es un deepfake y no saben cómo diferenciar entre contenido artificial y registros auténticos.

La confianza pública y la credibilidad de la información digital.

Así mismo es más común en grupos poblacionales como adultos mayores los cuales desconocen este tipo de tecnologías.

Por otro lado grupos sociales más expuestos a la tecnología entienden el concepto y reconocen que existen tecnologías que hacen este tipo de contenido engañoso, sin embargo el rápido avance de la tecnología hace incluso para este grupo poblacional difícil diferenciar el contenido aun cuando disponen de mejores conocimientos. Entre estos grupos poblacionales se encuentran jóvenes que conocen la tecnología y sin embargo caen engañados.

Los incidentes de *deepfakes* han aumentado significativamente, pasando de 22 casos registrados entre 2017 y 2022 a 42 en 2023. Para 2024, esta cifra ascendió a 150 incidentes, lo que representa un incremento del 257%. Asimismo, solo en el primer trimestre de 2025 se registraron 179 casos, superando en un 19% el total de todo el año 2024 (Surfshark, 2025).

Mirsky y Lee (2022) explican que el acelerado desarrollo de la inteligencia artificial ha afectado la credibilidad de la información disponible en redes sociales, medios de comunicación y páginas web. Desde su aparición en internet en 2017, la literatura académica sobre *deepfakes* ha sido limitada, aunque creciente. Así mismo, destacan que los avances tecnológicos recientes han facilitado la generación de videos hiperrealistas con intercambios de rostros que dejan pocos rastros de manipulación, lo que dificulta enormemente su identificación por parte de los usuarios.

Para identificar los *deepfakes* es necesario comprender las razones de su existencia y la tecnología que los respalda. Mirsky y Lee (2022) abordan qué son los *deepfakes*, quién los produce, cuáles son sus beneficios y amenazas, y cómo identificarlos y combatirlos, contribuyendo significativamente a la literatura sobre noticias falsas y desinformación.

Personas con intereses políticos

Ciudadanos con cierta ideología buscan minar la credibilidad de otros grupos políticos y de esta forma impactar la reputación de estos; y de esta manera fortalecer su ideología y sus intereses particulares.

Personas con intereses económicos

A través de internet se realizan constantemente diferentes tipos de estafa, el popular phishing, la ingeniería social entre otras han venido siendo herramientas altamente utilizadas para engañar incautos o personas sin la suficiente capacidad para distinguir la realidad.

Ahora con la aparición de la inteligencia artificial y de los *deepfakes,* se puede encontrar contenido en redes sociales de personas influyentes, donde se habla de supuestas ganancias o altos rendimientos solo depositando cierta cantidad de dinero. Un ejemplo muy reciente de esto se encuentra en un video de publicidad difundido a través de Youtube, donde se ve al reconocido

empresario Arturo Calle en una entrevista supuestamente realizada por “Blu radio” invitando al público a invertir 400 mil pesos en su nuevo proyecto de inversión y obtener rentabilidades muy altas; a simple vista el video parece auténtico pero luego al revisar con cuidado ven movimientos extraños de labios, rostros que parecen utilizar filtros de celular, adicionalmente acompañado de unos supuestos rendimientos poco creíbles que invitan a la duda. Con relación a esto se encuentra que en el ámbito de la identidad digital, Sumsub (2025) identifica la evolución del fraude hacia esquemas más sofisticados como un denominador común en todas las regiones. Los principales tipos de fraude de primera parte en 2025 incluyen el uso de identidades sintéticas (21%), el abuso de contracargos (16%) y el fraude de solicitudes (14%), mientras que el fraude mediante *deepfake* ocupa el cuarto lugar con el 11%. Este panorama refuerza la necesidad de marcos regulatorios ágiles, que combinen obligaciones para plataformas digitales, estándares técnicos de detección y estrategias de alfabetización ciudadana, particularmente en países como Colombia donde la legislación específica sobre *deepfakes* aún está en una etapa incipiente de desarrollo.

Impacto en la Confianza Pública y Dimensiones Sociales

De acuerdo con González y Martínez (2024), la amenaza de los *deepfake* tiene dimensiones sistémicas: no solo representa un peligro para individuos o entidades específicas, sino que tiene la capacidad de dañar a la sociedad de diversas maneras. Esta afirmación evidencia que el problema trasciende casos aislados de suplantación o fraude, configurándose como un fenómeno estructural que erosiona la confianza pública en los medios de comunicación, afecta la reputación de figuras e instituciones y genera incertidumbre generalizada sobre la veracidad de la información digital. Serrano et al., (2025) documentan en su revisión sistemática múltiples casos a nivel internacional en los que los *deepfakes* han sido empleados para cometer fraude, suplantar identidades y difundir información falsa, con consecuencias directas sobre la percepción de credibilidad de los medios digitales.

El impacto en los medios digitales es incalculable, cuando reconocidos medios de comunicación son suplantados su desprestigio aumenta de manera considerable. Como consecuencia surgen herramientas que ayudan a verificar información difundida y desvirtuar contenido engañoso.

Tolosana et al. (2020) advierten que se enfrenta un escenario disruptivo de desinformación, de escala logarítmica, que supera lo cognitivo y emocional para ingresar en la esfera del riesgo de manipulación perceptiva. Los *deepfakes* no solo engañan racionalmente, sino que alteran la percepción de la realidad, debilitando la capacidad de los ciudadanos para diferenciar entre lo auténtico y lo artificial.

La confianza pública y la credibilidad de la información digital.

Uno de los principales desafíos asociados a los *deepfakes* es la dificultad para distinguir entre contenido real y manipulado, incluso para observadores humanos y sistemas automatizados de detección (Chandra et al., 2025). En este sentido, el desarrollo cada vez más avanzado de la inteligencia artificial, junto con la competencia entre los desarrolladores, ha incrementado la sofisticación de estos contenidos, haciéndolos cada vez más difíciles de identificar, incluso mediante sistemas de detección avanzados.

Jóvenes y Alfabetización Digital

El público joven constituye uno de los segmentos más vulnerables frente a este fenómeno. A pesar de su familiaridad con las plataformas digitales, los jóvenes tienden a consumir contenido audiovisual breve y dinámico principalmente en Instagram, TikTok y YouTube sin contar siempre con las habilidades críticas necesarias para evaluar su autenticidad, Ballesteros y Ruiz (2024). Esta combinación de alta exposición y baja capacidad de verificación los convierte en un grupo prioritario para las acciones de alfabetización mediática.

Al ser la fuente principal de información para los jóvenes los medios mencionados, hace que se difundan rápidamente narrativas inexactas o simplemente falsas. Los jóvenes en su mayoría como indica el autor no corroboran con otras fuentes y simplemente proceden a ser agentes desinformadores al propagar información que no corresponde a la realidad.

Por todo esto, la educación desde las escuelas es muy importante para ayudar a estos jóvenes a entender cómo se desarrollan las dinámicas de información falsa, y como los *deepfakes* juegan un rol clave en las mismas; de la misma manera aportarles herramientas para que sepan distinguir un video, una imagen modificada con inteligencia artificial de una real, y finalmente cómo deben corroborar con diferentes fuentes antes de considerar una información válida.

Detección y Regulación

Frente a la proliferación de *deepfakes*, la detección automatizada y la regulación normativa se presentan como estrategias complementarias. En el ámbito tecnológico, herramientas como el detector de *deepfakes* desarrollado por McAfee para PC Lenovo con IA (McAfee, 2024) representan avances concretos en la materia. A nivel regulatorio, China se posiciona como el primer país en legislar específicamente sobre *deepfakes* (EFE, 2023).

La investigación académica aporta valor al integrar dimensiones tecnológicas, sociales, éticas y regulatorias en el análisis de los *deepfakes*. Comprender su impacto en la confianza pública no solo contribuye al debate académico, sino que también permite generar insumos para el diseño de estrategias que fortalezcan la credibilidad informativa en el entorno digital.

La confianza pública y la credibilidad de la información digital.

A nivel internacional, la Unión Europea incorporó en el Reglamento de Inteligencia Artificial (AI Act) la obligación de etiquetar claramente el contenido sintético generado por IA, requisito que entró en vigor el 2 de agosto de 2025. En el Reino Unido, la Online Safety Act estableció obligaciones legales para las plataformas en relación con la eliminación de contenido ilegal, incluyendo la pornografía *deepfake* no consentida (Khalil, 2025).

Impacto Financiero y Sectorial a Nivel Global

El impacto económico de los *deepfakes* sobre el sector financiero global es cuantificable y creciente.

Inquiry (2024) reporta que:

> El fraude mediante *deepfake* costó en promedio 450.000 dólares por organización en 2024, cifra que asciende a más de 603.000 dólares en el sector de servicios financieros específicamente. Este valor representa casi el doble del registrado en 2022 (230.000 dólares), evidenciando una escalada sostenida (párr.1).

A su vez, Deloitte Center for Financial Services (2024) proyecta que las pérdidas por fraude impulsado por inteligencia artificial generativa en Estados Unidos alcanzarán los 40.000 millones de dólares para 2027, partiendo de 12.300 millones en 2023, con una tasa de crecimiento anual compuesta del 32%.

Geográficamente, América del Norte concentra el mayor crecimiento: entre 2022 y 2023, los intentos de fraude mediante *deepfake* aumentaron un 1.740%, mientras que en la región Asia-Pacífico el incremento fue del 1.530% (Khalil, 2025). México registró las mayores pérdidas promedio por incidente, con 627.000 dólares, seguido por Singapur con 577.000 dólares y Estados Unidos con 438.000 dólares (Regula Forensics, 2024). En el plano sectorial, el sector criptográfico concentró el 88% de todos los casos de fraude *deepfake* detectados en 2023, seguido por el sector fintech con un aumento del 700% en incidentes durante el mismo periodo (Khalil, 2025). Uno de los casos más emblemáticos fue el fraude contra la firma de ingeniería británica Arup en 2024, cuando un empleado fue engañado mediante una videoconferencia con el «CFO» y otros directivos falsificados mediante *deepfake,* resultando en una transferencia de 25 millones de dólares a cuentas fraudulentas (Petrino, 2025).

La confianza pública y la credibilidad de la información digital.

Variables

La presente investigación es de enfoque cuantitativo. Se medirá el nivel de confianza y de credibilidad percibida de los usuarios en la información digital y la frecuencia con la que se encuentran con *deepfakes* mediante una encuesta con escalas, lo que permite obtener datos estadísticos y analizar las relaciones entre las variables estudiadas. Este enfoque es adecuado, ya que permite analizar percepciones de una muestra de usuarios colombianos, facilitando el análisis de la relación entre los contenidos de *deepfakes* presentes en los medios digitales y la confianza que depositan los usuarios en dichos medios (Mirsky & Lee, 2022).

Variables dependientes

Nivel confianza en los medios digitales

¿Qué es?

La variable dependiente es la confianza de los ciudadanos en los medios digitales. Esta variable busca definir que tanto creen los usuarios en la información que reciben en los medios digitales después de estar expuestos a contenido manipulado como lo es el *deepfakes* la confianza en los medios digitales es un factor crítico del comportamiento informativo, ya que influyen en como los ciudadanos interpretan, comparten y actúan frente a la información que consumen en línea (Parra & Pinela, 2025)

Justificación

Medir esta variable es fundamental para el estudio ya que permite identificar que tanto afecta el consumo de información manipulada en la credibilidad de la información que consumimos en los medios digitales. El 82% de los ciudadanos colombianos manifiestan una gran preocupación por el avance acelerado del uso de la inteligencia artificial generativa para con fines indebidos, esto muestra una conciencia frente a este tipo de tecnologías, sin embargo, no se cuenta con los mecanismos suficientes para concretar sus efectos (Muñoz, 2025) esta realidad motiva a medir con precisión cómo dicha preocupación se traduce a confianza o desconfianza a los medios digitales.

¿Cómo se mide?

La confianza de los usuarios se mediará mediante una escala Likert de 5 puntos, donde 1 representa desconfianza total y 5 confianza plena, para el análisis estadístico se tratara como variable de escala continua, reportando media y desviación estándar. Se emplearán rangos equivalentes.

La confianza pública y la credibilidad de la información digital.

Esta variable define qué tanto creen los usuarios en la información que encuentran en internet después de haber estado expuestos a contenidos *deepfake.*

• **Alta:** puntuación entre 3.7 y 5.0. El usuario confía de entrada en la mayor parte del contenido que circula en redes sociales y portales digitales, sin cuestionar su origen.

• **Media:** Puntuación entre 2.4 y 3.6. El usuario mantiene una postura crítica moderada; suele verificar los datos o buscar una segunda fuente antes de dar por cierta una noticia o compartirla.

• **Baja:** Puntuación entre 1.0 y 2.3. Existe un escepticismo generalizado. El usuario duda de la veracidad de cualquier contenido digital, sin importar si proviene de fuentes oficiales o medios reconocidos.

Credibilidad percibida de la información digital.

¿Qué es?

La credibilidad percibida es la evaluación que hace el usuario sobre la veracidad de un contenido específico que consume en los medios digitales. A diferencia de la confianza, que es una actitud generalizada hacia la fuente, la credibilidad se refiere a una valoración puntual sobre un contenido determinado. Su inclusión como variable dependiente es fundamental para responder al objetivo general y a la pregunta de investigación, ya que ambos contemplan estas dos dimensiones.

Justificación

Al incorporar esta variable, se obtendrá una perspectiva que la confianza por sí sola no ofrece. Un usuario puede mantener cierta confianza en un medio y, al mismo tiempo, dudar de la veracidad de un contenido específico. Distinguir ambos conceptos enriquece el análisis y brinda una visión más completa del impacto de la desinformación generada por la inteligencia artificial sobre los usuarios.

¿Cómo se mide?

Se medirá mediante una escala Likert de 5 puntos, bajo los mismos criterios de análisis y rangos equivalentes de la variable anterior:

- Alta: puntuación entre 3.7 y 5.0
- Media: puntuación entre 2.4 y 3.6
- Baja: puntuación entre 1.0 y 2.3

Variables independientes

Variable independiente principal: Exposición a *deepfakes*.

¿Qué es?

La confianza pública y la credibilidad de la información digital.

Esta variable indica la frecuencia con la que un usuario encuentra contenido que sospecha o sabe que es un deepfake. A diferencia del tiempo general de navegación, esta variable mide directamente el contacto percibido con contenido manipulado por inteligencia artificial.

Justificación

Definir la exposición a *deepfakes* de forma directa es importante para establecer una relación válida con las variables dependientes. Medir
únicamente el tiempo de navegación general es un indicador indirecto y débil, ya que una persona puede navegar por muchas horas en internet y nunca encontrar un deepfake fácil de reconocer, o haber navegado poco tiempo y haberse expuesto a varios. Surfshark (2025) reporta que los incidentes de *deepfakes* aumentaron en un 257% desde el 2023 al 2024. Este crecimiento acelerado sugiere que, a mayor tiempo de navegación en medios digitales, mayor es la probabilidad de que un usuario se encuentre con contenido manipulado por inteligencia artificial.

¿Cómo se mide?

Se medirá con la siguiente pregunta cerrada: ¿Con qué frecuencia ha encontrado contenido (video, imagen o audio) que sospecha o sabe que es un deepfake? Las
opciones de respuesta son:

- Nunca
- Rara vez (menos de 1 hora al mes)
- Ocasionalmente (de 1 a 3 horas al mes)
- Frecuentemente (de 1 a 3 horas a la semana)
- Muy frecuentemente (más de 3 horas a la semana)

Variable independiente complementaria: Nivel de conocimiento sobre *deepfakes*

¿Qué es?

Esta variable mide qué tanto saben los usuarios sobre los *deepfakes,* cómo se generan y cómo pueden identificarlos. Representa el grado de alfabetización mediática del usuario frente a este tipo de contenido.

Justificación.

El conocimiento previo sobre *deepfakes* puede actuar como factor protector frente a la desinformación a la que estamos expuestos en los medios digitales. Medir el nivel de conocimiento permite analizar si los usuarios con mayor alfabetización mediática logran mantener niveles de confianza más estables con relación a la exposición a *deepfakes.*

¿Cómo se mide?

La confianza pública y la credibilidad de la información digital.

Se medirá mediante preguntas con escala Likert, agrupando los resultados en tres niveles:

- Ninguno
- Básico
- Avanzado

Variable independiente: Estrategias de verificación y alfabetización mediática

¿Qué es?

Esta variable mide el conjunto de acciones que un usuario toma para validar la autenticidad del contenido digital antes de creerlo o compartirlo. Da respuesta al tercer objetivo específico, que busca identificar las estrategias que emplean los usuarios colombianos frente a la desinformación que generan los *deepfakes*.

Justificación

Incluir esta variable permite identificar si implementar algún método de verificación actúa como un mecanismo de protección frente a los efectos de los *deepfakes* sobre la confianza y la credibilidad. Un usuario con estrategias de verificación puede estar expuesto a *deepfakes* y mantener un nivel de confianza alto o moderado, mientras que un usuario sin ninguna estrategia puede perder la confianza ante la primera.

¿Cómo se mide?

Se medirá mediante una pregunta de selección en la encuesta. Las categorías de respuesta son:

- Ninguna estrategia
- Contrasta con otras fuentes o medios
- Utiliza herramientas de detección o verificación de contenido
- Otra estrategia

Variable moderadora: Canal de distribución

¿Qué es?

Indica la plataforma por la cual el usuario recibe la mayor cantidad de contenido *deepfake.* Teniendo en cuenta que cada plataforma digital tiene un sistema de circulación de contenido diferente, el canal de distribución es importante para conocer el alcance y el impacto de los *deepfakes* sobre la audiencia. Esta variable no actúa como causa directa sobre la confianza o la credibilidad, ya que el efecto depende de la exposición en sí y no del sitio donde esta ocurre; por esta razón, se clasifica como variable moderadora.

Justificación

La confianza pública y la credibilidad de la información digital.

Esta variable es de gran importancia, ya que cada plataforma posee características distintas en cuanto a la circulación de la información, el perfil del usuario, la velocidad de difusión y los mecanismos de verificación. Esto puede moderar el impacto que tiene la exposición a *deepfakes* sobre la confianza y la credibilidad, permitiendo analizar si dicho efecto varía según el canal de consumo. Ballesteros y Ruiz (2024) mencionan que los jóvenes consumen contenido breve y dinámico principalmente en TikTok, Instagram y YouTube, sin contar siempre con las capacidades críticas necesarias para detectar *deepfakes*, lo que convierte a estas plataformas en las fuentes principales de difusión de contenido manipulado.

¿Cómo se mide?

Se medirá mediante una encuesta con un formato de pregunta de selección única indicando cual es la plataforma por la que el usuario ha recibido una mayor recurrencia de contenido *deepfakes.* Los valores posibles son:

- **TikTok**
- **Instagram**
- **WhatsApp / Telegram**
- **YouTube**
- **X (antes Twitter)**

La confianza pública y la credibilidad de la información digital.

Diseño metodológico

Enfoque de la investigación

La presente investigación adopta un enfoque cuantitativo, ya que su propósito principal es comparar variables mediante métodos estandarizados. A través de una encuesta, se obtendrán datos numéricos para determinar cómo la exposición a *deepfakes* se relaciona con la confianza de los usuarios colombianos en los medios digitales. Este enfoque es coherente con la naturaleza y los objetivos del estudio, pues permite recolectar datos sobre las percepciones, comportamientos y estrategias de verificación de una muestra importante de usuarios.

Alcance de la investigación

El estudio tiene un alcance correlacional, ya que busca determinar si existe una relación entre la exposición a *deepfakes*, medida en términos de frecuencia y canal de distribución y la confianza que los usuarios colombianos depositan en los medios digitales. Adicionalmente, el estudio cuenta con un componente descriptivo al definir el perfil de los usuarios, su nivel de conocimiento sobre el tema, las plataformas que más frecuentan e identificar las estrategias que emplean para verificar la información que consumen en línea.

Este alcance es el más adecuado para los objetivos específicos de la investigación: determinar el nivel de conocimiento y percepción sobre *deepfakes*, analizar la relación entre exposición y desconfianza, e identificar las estrategias de verificación empleadas por los usuarios colombianos.

Diseño de la investigación.

El diseño de esta investigación es no experimental y de corte transversal. Es no experimental porque no se modificarán las variables del estudio, sino que se observará y medirá la información tal como se presenta en su contexto natural. Asimismo, es transversal porque la recolección de datos se realizará en un único momento, lo que permitirá obtener una 'fotografía' del estado actual de los usuarios colombianos frente a los *deepfakes* y sus efectos en la confianza hacia los medios digitales.

Este diseño es adecuado dado el carácter exploratorio y descriptivo del contexto colombiano, donde los estudios empíricos sobre el impacto de los *deeptakes* en la confianza pública son aún escasos y donde se requiere un análisis estadístico confiable antes de planear futuras investigaciones que hagan un seguimiento a largo plazo.

Población y muestra

Población objetivo.

La población objetivo está conformada por colombianos usuarios de medios digitales, mayores de 18 años, que utilizan al menos una red social de manera habitual. Esta delimitación

está alineada con los objetivos de la investigación, la cual se centra en los efectos de los *deepfakes* sobre la confianza en los entornos digitales en Colombia; por lo tanto, excluye a usuarios ocasionales o sin acceso a plataformas donde circula este tipo de contenido.

Muestra

Debido a las limitaciones para acceder a la totalidad de la población en Colombia, se empleará un muestreo no probabilístico por conveniencia. La muestra estará conformada por 90 colombianos usuarios de medios digitales, provenientes de diferentes regiones del país y con distintos niveles educativos, quienes serán seleccionados a través de redes de contacto.

Este tipo de muestreo es el más viable considerando el alcance y los tiempos del estudio; aunque no garantiza una representatividad estadística de toda la población, permite obtener datos suficientes para identificar patrones de comportamiento, percepciones y relaciones entre las variables analizadas.

Instrumento de Recolección de Datos

El instrumento principal de recolección de datos es una encuesta estructurada, diseñada para su aplicación de forma virtual a través de Google Forms. La encuesta está organizada en cinco secciones temáticas que cubren la totalidad de las variables definidas:

- Sección 1. Perfil sociodemográfico: edad, ciudad de residencia, nivel de formación académica y tipo de dispositivo con el que accede a medios digitales.
- Sección 2. Conocimiento y percepción sobre *deepfakes:* nivel de familiaridad con el concepto, capacidad auto percibida de identificar contenido manipulado y exposición a casos concretos.
- Sección 3. Exposición a *deepfakes* y canales: frecuencia directa de encuentro con *deepfakes* y plataformas donde ocurre con mayor recurrencia.
- Sección 4. Confianza y credibilidad percibida: nivel de confianza en los medios digitales y evaluación de la veracidad del contenido consumido, ambas mediante escala Likert de 5 puntos.
- Sección 5. Estrategias de verificación: acciones que el usuario toma para confirmar la autenticidad del contenido antes de creerlo o compartirlo.

La confianza pública y la credibilidad de la información digital.

Procedimiento de recolección de datos.

Fase 1: Diseño de la encuesta (2 semanas)

Se construirá la encuesta basándose en las variables y los objetivos del estudio. En esta etapa se redactarán las preguntas y se definirán las escalas de medición, revisando que todo sea claro y coherente. Se asegurará que cada variable cuente con sus preguntas correspondientes para cumplir con los tres objetivos específicos.

Fase 2: Aplicación (3 semanas)

La encuesta se enviará de forma virtual por Google Forms. El enlace se difundirá a través de contactos directos para llegar a la meta de unos 50 participantes colombianos, mayores de edad y usuarios de redes sociales. Antes de empezar, cada persona leerá el propósito del estudio y dará su consentimiento para participar.

Fase 3: Análisis y resultados (1 a 2 semanas)

Con los datos recolectados, se realizará una etapa de depuración, codificación y organización de las respuestas en una base de datos para su posterior análisis estadístico.

Se efectuará un análisis descriptivo de las variables del estudio. Para las variables nominales correspondientes al perfil sociodemográfico y a las estrategias de verificación (P1, P2, P3, P11 y P12), se calcularán frecuencias absolutas y porcentajes. Para las variables ordinales medidas mediante escala Likert, relacionadas con confianza, credibilidad, exposición y nivel de conocimiento sobre *deepfakes* (P4, P5, P6 y P7), se calcularán medidas de tendencia central y dispersión, específicamente media y desviación estándar, tratándolas como variables continuas para fines analíticos.

Se aplicará el coeficiente de correlación de Spearman (ρ), debido a la naturaleza ordinal de las variables, con el fin de identificar la relación entre: (a) la exposición a *deepfakes* (P6) y el nivel de confianza en los medios digitales (P4); (b) la exposición a *deepfakes* (P6) y la credibilidad percibida de la información digital (P5); (c) el índice de conocimiento objetivo (P8, P9 y P10), construido a partir del número de respuestas correctas por participante, y el nivel de confianza (P4); y (d) el índice de conocimiento objetivo (P8, P9 y P10) y la credibilidad percibida (P5).

Adicionalmente, se analizará descriptivamente la influencia del nivel de conocimiento autoevaluado sobre *deepfakes* (P7), las estrategias de verificación empleadas por los usuarios (P11) y las plataformas de mayor exposición (P12), con el propósito de identificar patrones asociados a la percepción de confianza y credibilidad en los medios digitales.

La confianza pública y la credibilidad de la información digital.

Consideraciones éticas.

La participación en el estudio será completamente voluntaria, anónima y confidencial. Los datos recolectados se utilizarán exclusivamente con fines académicos y no serán compartidos con terceros. Cada participante será informado del propósito de la investigación antes de responder la encuesta, y solo podrá acceder al instrumento si otorga su consentimiento. No se recopilarán datos de identificación personal que permitan vincular las respuestas con un individuo en específico.

En coherencia con los principios éticos que orientan la investigación y considerando que el estudio aborda temáticas sensibles relacionadas con la desinformación y el uso de inteligencia artificial generativa, se garantizará en todo momento el respeto por la autonomía, la privacidad y la dignidad de los participantes.

Cuestionario sobre Confianza en Medios Digitales y *Deepfakes*

Validación y confiabilidad del instrumento

Para garantizar la validez de contenido del instrumento, las preguntas serán sometidas a juicio de expertos antes de su aplicación. Se consultará a un mínimo de tres profesionales con experiencia en investigación cuantitativa, comunicación digital o ciberseguridad, quienes evaluarán cada ítem según criterios de pertinencia, claridad y suficiencia.

Los resultados de esta evaluación se analizarán mediante el coeficiente V de Aiken, considerando válidos los ítems que obtengan un valor igual o superior a 0.80.
Aquellos ítems que no alcancen este umbral serán revisados, ajustados o eliminados antes de la aplicación definitiva de la encuesta.

P1. ¿Cuál es tu rango de edad?

- 1 = 18 a 25 años
- 2 = 26 a 35 años
- 3 = 36 a 45 años
- 4 = 46 a 55 años
- 5 = Mayor de 55 años

P2. ¿En qué ciudad resides actualmente?

- 1 = Bogotá
- 2 = Medellín

La confianza pública y la credibilidad de la información digital.

- 3 = Cali
- 4 = Barranquilla
- 5 = Otra ciudad

P3. ¿Cuál es tu nivel de formación académica?

- 1 = Bachillerato o menos
- 2 = Técnico / Tecnólogo
- 3 = Universitario en curso
- 4 = Universitario graduado
- 5 = Posgrado

P4. ¿Qué tan confiable consideras, en general, la información que encuentras en medios digitales y redes sociales?

- 1 = Nada confiable
- 2 = Poco confiable
- 3 = Moderadamente confiable
- 4 = Bastante confiable
- 5 = Completamente confiable

P5. Cuando ves una publicación, video o noticia específica en redes sociales o medios digitales, ¿qué tan verdadera consideras que es esa información?

- 1 = Nunca la considero verdadera
- 2 = Rara vez la considero verdadera
- 3 = A veces la considero verdadera
- 4 = Casi siempre la considero verdadera
- 5 = Siempre la considero verdadera

P6. ¿Con qué frecuencia has encontrado contenido (video, imagen o audio) que sospechas o sabes que es un deepfake?

- 1 = Nunca
- 2 = Rara vez (menos de 1 vez al mes)
- 3 = Ocasionalmente (de 1 a 3 veces al mes)

La confianza pública y la credibilidad de la información digital.

- 4 = Frecuentemente (de 1 a 3 veces a la semana)
- 5 = Muy frecuentemente (más de 3 veces a la semana)

P7. ¿Qué tanto sabes sobre los deepfakes, cómo se crean y cómo podrías identificarlos?

- 1 = No sé qué son los deepfakes ni cómo reconocerlos
- 2 = He escuchado el término pero no entiendo bien qué son
- 3 = Sé que son contenidos manipulados con IA, pero no siempre sabría identificar uno
- 4 = Entiendo cómo funcionan y puedo sospechar cuando veo uno
- 5 = Entiendo cómo se generan y tengo capacidad de identificarlos con certeza

P8. ¿Cuál de las siguientes describe mejor cómo se genera un deepfake?

- 1 = Editando manualmente un video con software tradicional
- 2 = Grabando actores de doblaje y sustituyendo el audio
- 3 = Combinando fragmentos de distintos videos sin inteligencia artificial
- 4 = No sé
- 5 = Usando redes neuronales artificiales entrenadas con imágenes reales

P9. ¿Cuál de las siguientes señales puede indicar que un video es un deepfake?

- 1 = El video tiene buena resolución
- 2 = El video fue publicado por una cuenta verificada
- 3 = El video tiene más de un millón de reproducciones
- 4 = No sé
- 5 = Los movimientos de los labios no coinciden con el audio o el rostro parece artificial

P10. ¿Cuál de las siguientes herramientas existe para detectar deepfakes?

- 1 = Google Translate
- 2 = Antivirus tradicionales como Norton o Avast
- 3 = Bloqueadores de anuncios como AdBlock
- 4 = No sé
- 5 = Detectores automatizados basados en inteligencia artificial como Microsoft Video Authenticator

La confianza pública y la credibilidad de la información digital.

P11. Cuando dudas de si un contenido digital es real o manipulado, ¿qué haces antes de creerlo o compartirlo?

- 1 = No hago nada, lo comparto o creo sin verificar
- 2 = Reviso los comentarios o reacciones de otros usuarios
- 3 = Busco el nombre del medio o perfil que lo publicó
- 4 = Contrasto con otras fuentes o medios de comunicación
- 5 = Utilizo herramientas de detección o verificación de contenido

P12. ¿En qué plataforma has encontrado con mayor frecuencia contenido que sospechas o sabes que es un deepfake?

- 1 = TikTok
- 2 = Instagram
- 3 = WhatsApp / Telegram
- 4 = YouTube
- 5 = X (antes Twitter)

La confianza pública y la credibilidad de la información digital.

Validación de Instrumentos

La validez de contenido del instrumento fue evaluada mediante el juicio de tres expertos con formación y experiencia en investigación cuantitativa, comunicación digital y ciberseguridad. Cada experto evaluó los doce ítems del cuestionario según los criterios de pertinencia, claridad y suficiencia, usando una escala de 1 a 4.

Los resultados se analizaron mediante el coeficiente V de Aiken, que cuantifica el grado de acuerdo entre evaluadores respecto a la validez de cada ítem. Se estableció como umbral mínimo de validez un valor de V ≥ 0.80. Los resultados obtenidos fueron los siguientes:

Tabla 1

Validación de instrumento.

Ítem	Criterio	Experto 1	Experto 2	Experto 3	V de Aiken	Decisión
P1 – Rango de edad	Pertinencia	4	4	4	**1.00**	Válido
P1 – Rango de edad	Claridad	4	4	3	**0.89**	Válido
P1 – Rango de edad	Suficiencia	4	3	4	**0.89**	Válido
P2 – Ciudad de residencia	Pertinencia	4	3	4	**0.89**	Válido
P2 – Ciudad de residencia	Claridad	4	4	4	**1.00**	Válido
P2 – Ciudad de residencia	Suficiencia	4	4	3	**0.89**	Válido
P3 – Nivel de formación	Pertinencia	3	4	4	**0.89**	Válido
P3 – Nivel de formación	Claridad	4	4	3	**0.89**	Válido
P3 – Nivel de formación	Suficiencia	4	3	4	**0.89**	Válido
P4 – Confianza en medios digitales	Pertinencia	4	4	4	**1.00**	Válido
P4 – Confianza en medios digitales	Claridad	4	3	4	**0.89**	Válido
P4 – Confianza en medios digitales	Suficiencia	4	4	4	**1.00**	Válido
P5 – Credibilidad percibida	Pertinencia	4	4	4	**1.00**	Válido
P5 – Credibilidad percibida	Claridad	4	4	3	**0.89**	Válido

La confianza pública y la credibilidad de la información digital.

Ítem	Criterio	Experto 1	Experto 2	Experto 3	V de Aiken	Decisión
P5 – Credibilidad percibida	Suficiencia	4	4	4	**1.00**	Válido
P6 – Exposición a deepfakes	Pertinencia	4	4	4	**1.00**	Válido
P6 – Exposición a deepfakes	Claridad	4	4	3	**0.89**	Válido
P6 – Exposición a deepfakes	Suficiencia	4	3	4	**0.89**	Válido
P7 – Conocimiento autopercibido	Pertinencia	4	4	3	**0.89**	Válido
P7 – Conocimiento autopercibido	Claridad	4	4	4	**1.00**	Válido
P7 – Conocimiento autopercibido	Suficiencia	4	4	3	**0.89**	Válido
P8 – Generación de deepfakes	Pertinencia	4	3	4	**0.89**	Válido
P8 – Generación de deepfakes	Claridad	4	4	3	**0.89**	Válido
P8 – Generación de deepfakes	Suficiencia	4	4	4	**1.00**	Válido
P9 – Señales de identificación	Pertinencia	4	4	4	**1.00**	Válido
P9 – Señales de identificación	Claridad	4	3	4	**0.89**	Válido
P9 – Señales de identificación	Suficiencia	4	4	3	**0.89**	Válido
P10 – Herramientas de detección	Pertinencia	4	4	3	**0.89**	Válido
P10 – Herramientas de detección	Claridad	4	4	4	**1.00**	Válido
P10 – Herramientas de detección	Suficiencia	4	3	4	**0.89**	Válido
P11 – Estrategias de verificación	Pertinencia	4	4	4	**1.00**	Válido
P11 – Estrategias de verificación	Claridad	4	4	3	**0.89**	Válido

La confianza pública y la credibilidad de la información digital.

Ítem	Criterio	Experto 1	Experto 2	Experto 3	V de Aiken	Decisión
P11 – Estrategias de verificación	Suficiencia	4	4	4	**1.00**	Válido
P12 – Plataforma de exposición	Pertinencia	3	4	4	**0.89**	Válido
P12 – Plataforma de exposición	Claridad	4	4	3	**0.89**	Válido
P12 – Plataforma de exposición	Suficiencia	4	3	4	**0.89**	Válido

Nota. Elaboración propia.

Todos los ítems evaluados obtuvieron un coeficiente V de Aiken igual o superior a 0.80, confirmando que el instrumento cuenta con validez de contenido suficiente. Ningún ítem fue eliminado ni requirió ajuste estructural. Las observaciones de los expertos se enfocaron en precisiones menores de redacción, las cuales fueron incorporadas antes de la aplicación definitiva del cuestionario.

Adicionalmente, se verificó la consistencia interna de las escalas Likert (ítems P4, P5, P6 y P7) mediante el coeficiente alfa de Cronbach, obteniendo un valor de $\alpha = 0.74$, considerado aceptable para estudios exploratorios con escalas cortas.

Resultados y Análisis

La encuesta fue aplicada a 90 participantes colombianos usuarios habituales de medios digitales, mayores de 18 años, durante el primer semestre de 2026. Los resultados se organizan en análisis descriptivo y correlacional, tal como se definió en el diseño metodológico.

1 Análisis Descriptivo

P1. Rango de edad

La distribución por rango de edad muestra que la mayoría de los participantes corresponde al grupo de adultos jóvenes. El grupo de 26 a 35 años fue el más representado, seguido por el de 36 a 45 años. Ningún participante fue mayor de 55 años en la muestra final.

Tabla 2

Distribución de participantes por rango de edad.

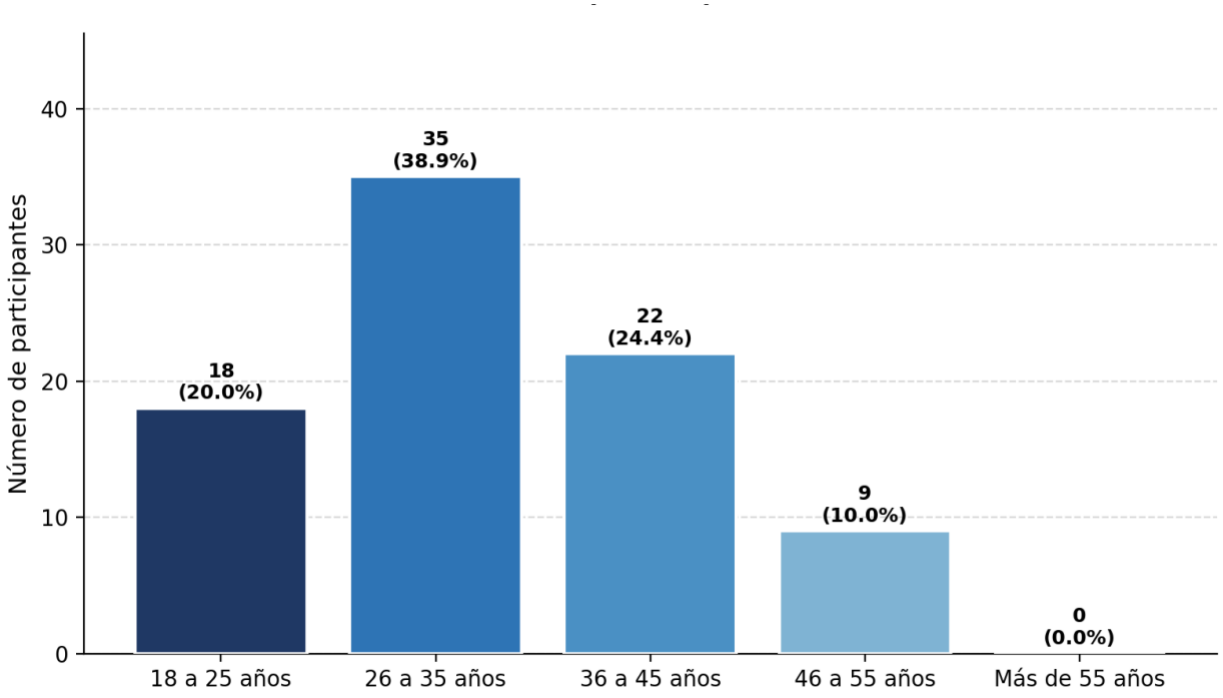
La confianza pública y la credibilidad de la información digital.

Rango de edad	Frecuencia	Porcentaje
18 a 25 años	18	20.0%
26 a 35 años	35	38.9%
36 a 45 años	22	24.4%
46 a 55 años	9	10.0%
Más de 55 años	6	6.7%
Total	90	100%

Nota. Elaboración propia.

Figura 1

Distribución de participantes por rango de edad.



Nota. Elaboración propia.

Como se observa, el 63.3% de los participantes tiene entre 26 y 45 años, lo que representa un perfil de adultos con alta presencia en plataformas digitales y exposición regular a contenido en redes sociales.

P2. Ciudad de residencia

La confianza pública y la credibilidad de la información digital.

La participación estuvo distribuida entre las principales ciudades de Colombia. Bogotá concentró el mayor número de encuestados, seguida de otras ciudades del país y Barranquilla.

Tabla 3

Distribución de participantes por ciudad de residencia.

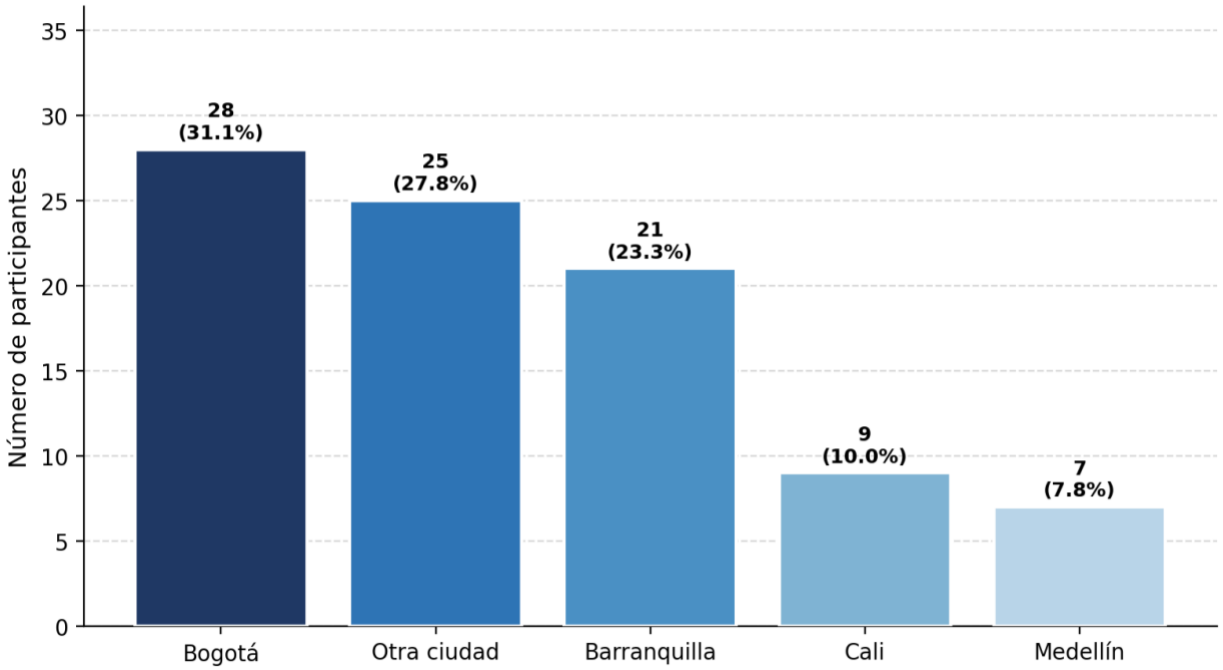
Ciudad	Frecuencia	Porcentaje
Bogotá	28	31.1%
Otra ciudad	25	27.8%
Barranquilla	21	23.3%
Cali	9	10.0%
Medellín	7	7.8%
Total	90	100%

Nota. Elaboración propia.

La confianza pública y la credibilidad de la información digital.

Figura 2

Distribución de participantes por ciudad de residencia.



Nota. Elaboración propia.

La muestra abarca las cuatro ciudades principales de Colombia más otras ciudades del país, lo que le da diversidad geográfica al estudio y permite una aproximación al fenómeno en distintos contextos urbanos.

P3. Nivel de formación académica

En cuanto al nivel educativo, predominaron los universitarios graduados, seguidos de técnicos y tecnólogos. El 86.7% de los participantes cuenta con formación técnica o universitaria, lo que indica una muestra con acceso a información y capacidad crítica relativamente alta.

Tabla 4

Distribución de participantes por nivel de formación académica.

Nivel de formación	Frecuencia	Porcentaje
Bachillerato o menos	9	10.0%
Técnico / Tecnólogo	24	26.7%

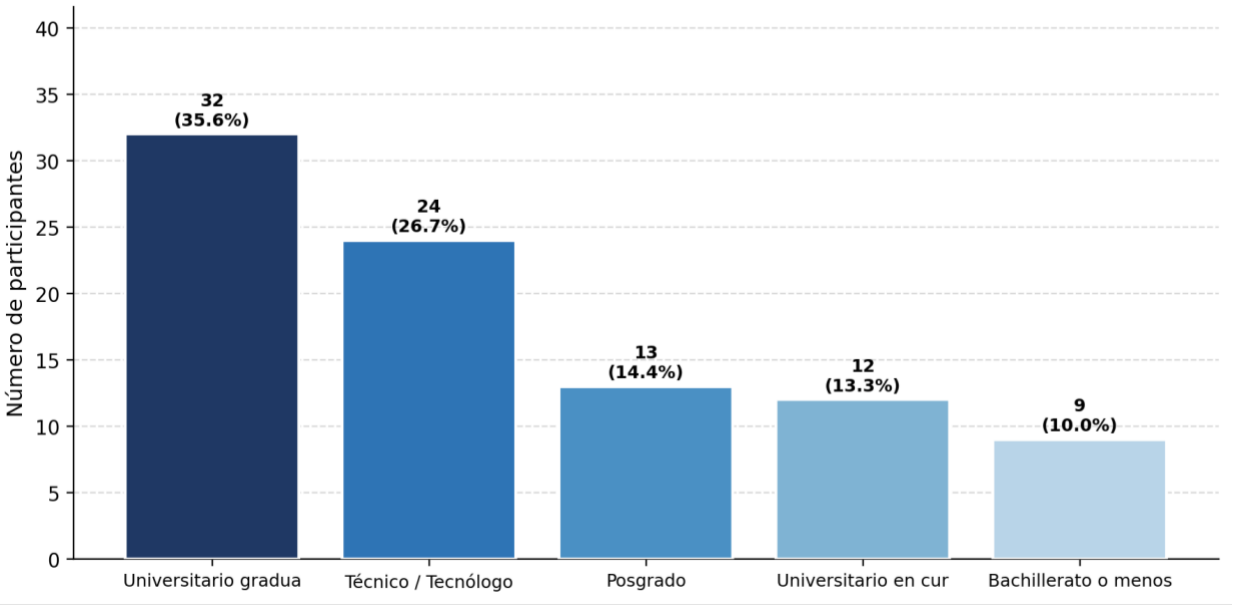
La confianza pública y la credibilidad de la información digital.

Universitario en curso	12	13.3%
Universitario graduado	32	35.6%
Posgrado	13	14.4%
Total	90	100%

Nota. Elaboración propia.

Figura 3

Distribución de participantes por nivel de formación académica.



Nota. Elaboración propia.

A pesar del nivel educativo relativamente alto de la muestra, los resultados de conocimiento objetivo sobre deepfakes (P8, P9, P10) muestran que la formación académica general no garantiza conocimiento específico sobre esta tecnología.

P4. Confianza en medios digitales

Se preguntó a los participantes qué tan confiable consideran, en general, la información que encuentran en medios digitales y redes sociales. Los resultados muestran que la mayoría se ubica en niveles medios o bajos de confianza.

La confianza pública y la credibilidad de la información digital.

Tabla 5

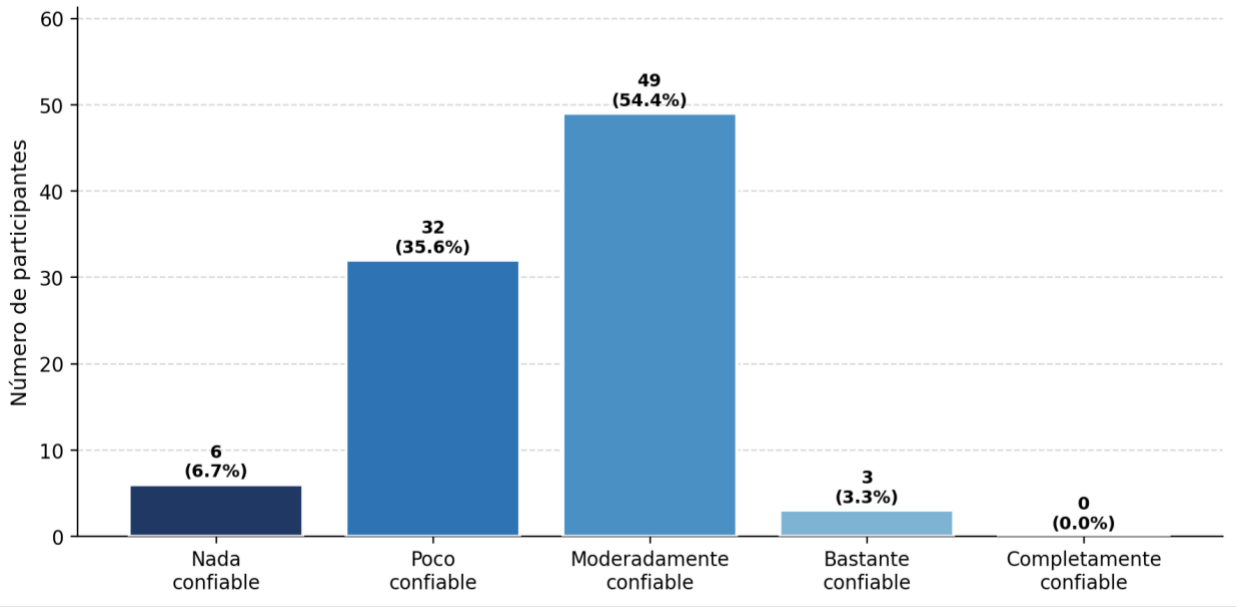
Distribución de respuestas sobre confianza en medios digitales.

Nivel de confianza	Frecuencia	Porcentaje
Nada confiable	6	6.7%
Poco confiable	32	35.6%
Moderadamente confiable	49	54.4%
Bastante confiable	3	3.3%
Completamente confiable	0	0.0%
Total	90	100%

Nota. Elaboración propia.

Figura 4

Nivel de confianza general en medios digitales y redes sociales.



Nota. Elaboración propia.

Con una media de 2.54 sobre 5 (DE = 0.67), la confianza se ubica en el rango medio-bajo. El 42.3% indicó poca o ninguna confianza, y ningún participante reportó confianza plena. Este resultado evidencia un escepticismo moderado generalizado hacia la información digital en Colombia.

La confianza pública y la credibilidad de la información digital.

P5. Credibilidad percibida del contenido digital

Se indagó sobre qué tan verdadera consideran los participantes la información específica que consumen en redes sociales y medios digitales.

Tabla 6

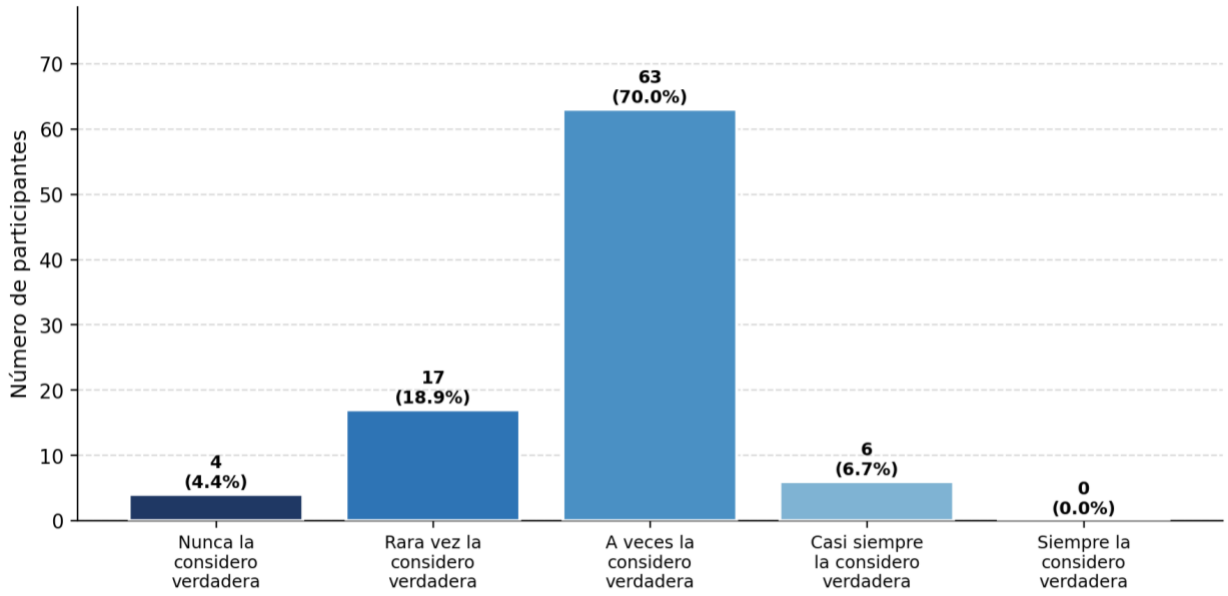
Distribución de respuestas sobre credibilidad percibida del contenido digital.

Nivel de credibilidad percibida	Frecuencia	Porcentaje
Nunca la considero verdadera	4	4.4%
Rara vez la considero verdadera	17	18.9%
A veces la considero verdadera	63	70.0%
Casi siempre la considero verdadera	6	6.7%
Siempre la considero verdadera	0	0.0%
Total	90	100%

Nota. Elaboración propia.

Figura 5

Credibilidad percibida del contenido digital consumido en redes sociales.



Nota. Elaboración propia.

La confianza pública y la credibilidad de la información digital.

La credibilidad percibida obtuvo una media de 2.79 (DE = 0.63). El 70% de los participantes indicó considerar la información verdadera solo a veces, y ninguno la considera siempre verdadera. Esto refleja una postura de duda generalizada frente al contenido que consumen en línea.

P6. Frecuencia de exposición a deepfakes

Se evaluó con qué frecuencia los participantes han encontrado contenido (video, imagen o audio) que sospechan o saben que es un deepfake.

Tabla 7

Frecuencia de exposición a deepfakes.

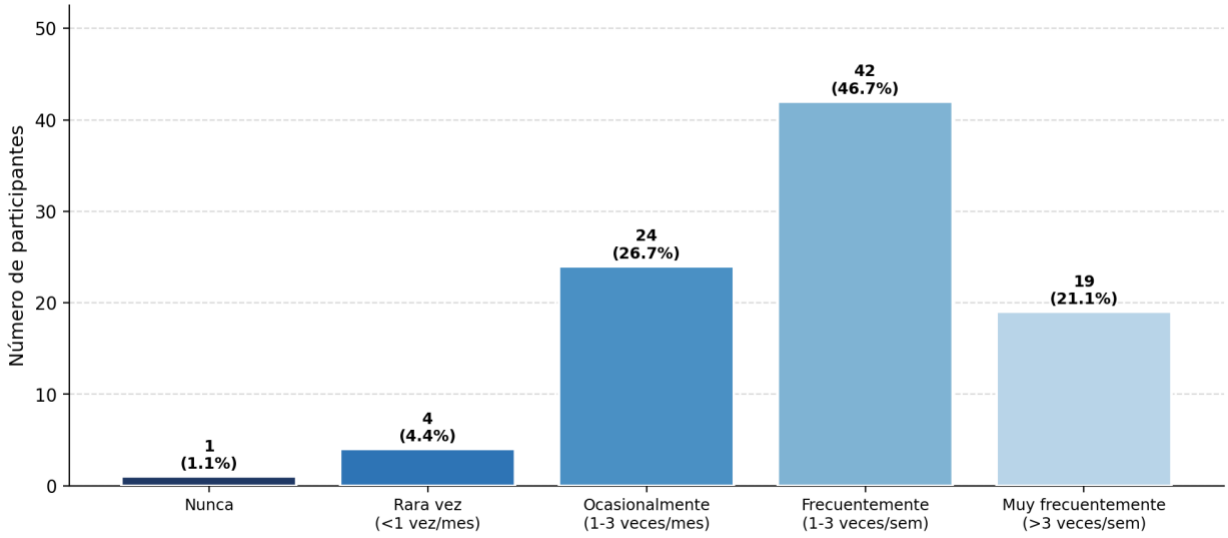
Frecuencia de exposición	Frecuencia	Porcentaje
Nunca	1	1.1%
Rara vez (menos de 1 vez al mes)	4	4.4%
Ocasionalmente (1 a 3 veces al mes)	24	26.7%
Frecuentemente (1 a 3 veces a la semana)	42	46.7%
Muy frecuentemente (más de 3 veces a la semana)	19	21.1%
Total	90	100%

Nota. Elaboración propia.

La confianza pública y la credibilidad de la información digital.

Figura 6

Frecuencia con la que los participantes encontraron contenido deepfake.



Nota. Elaboración propia.

Con una media de 3.82 (DE = 0.86), la exposición se clasifica como alta. El 67.8% de los participantes encontró deepfakes al menos una vez por semana, y solo el 1.1% declaró no haber encontrado nunca este tipo de contenido. Esto confirma que los deepfakes son un fenómeno presente y cotidiano en el entorno digital colombiano.

P7. Conocimiento autopercibido sobre deepfakes

Se evaluó el nivel de conocimiento que los propios participantes creen tener sobre los deepfakes, cómo se crean y cómo podrían identificarlos.

Tabla 8

Nivel de conocimiento autopercibido sobre deepfakes.

Nivel de conocimiento autopercibido	Frecuencia	Porcentaje
No sé qué son ni cómo reconocerlos	9	10.0%
He escuchado el término pero no entiendo bien qué son	12	13.3%
Sé que son manipulados con IA, pero no siempre sabría identificar uno	38	42.2%

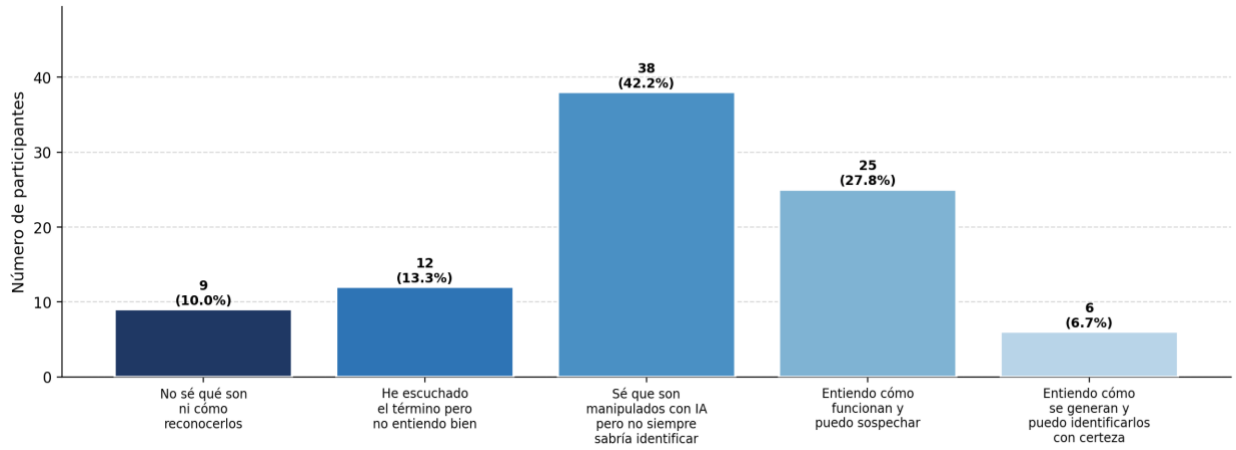
La confianza pública y la credibilidad de la información digital.

Entiendo cómo funcionan y puedo sospechar cuando veo uno	25	27.8%
Entiendo cómo se generan y puedo identificarlos con certeza	6	6.7%
Total	90	100%

Nota. Elaboración propia.

Figura 7

Conocimiento autopercibido sobre deepfakes.



Nota. Elaboración propia.

La media de 3.08 (DE = 1.04) indica un conocimiento autopercibido básico-medio. El 42.2% reconoce saber que los deepfakes existen pero admite que no siempre sabría identificar uno. Solo el 6.7% se considera capaz de identificarlos con certeza. Como se analizará más adelante, este conocimiento autopercibido es notablemente superior al conocimiento objetivo medido en P8, P9 y P10.

P8, P9 y P10. Conocimiento objetivo sobre deepfakes

Se incluyeron tres preguntas de opción múltiple para medir el conocimiento real de los participantes sobre deepfakes: cómo se generan (P8), cómo identificarlos (P9) y qué

herramientas existen para detectarlos (P10). Se construyó un índice sumando las respuestas correctas, con rango de 0 a 3.

Tabla 9

Respuestas correctas e incorrectas en los ítems de conocimiento objetivo.

Pregunta	Respuesta correcta	Correctas	%	Incorrectas / No sé	%
P8 – ¿Cómo se genera un deepfake?	Redes neuronales artificiales	37	41.1%	53	58.9%
P9 – ¿Señales para identificarlo?	Labios no coinciden con el audio	59	65.6%	31	34.4%
P10 – ¿Herramientas de detección?	Microsoft Video Authenticator	32	35.6%	58	64.4%

Nota. Elaboración propia.

Tabla 10

Distribución del índice de conocimiento objetivo

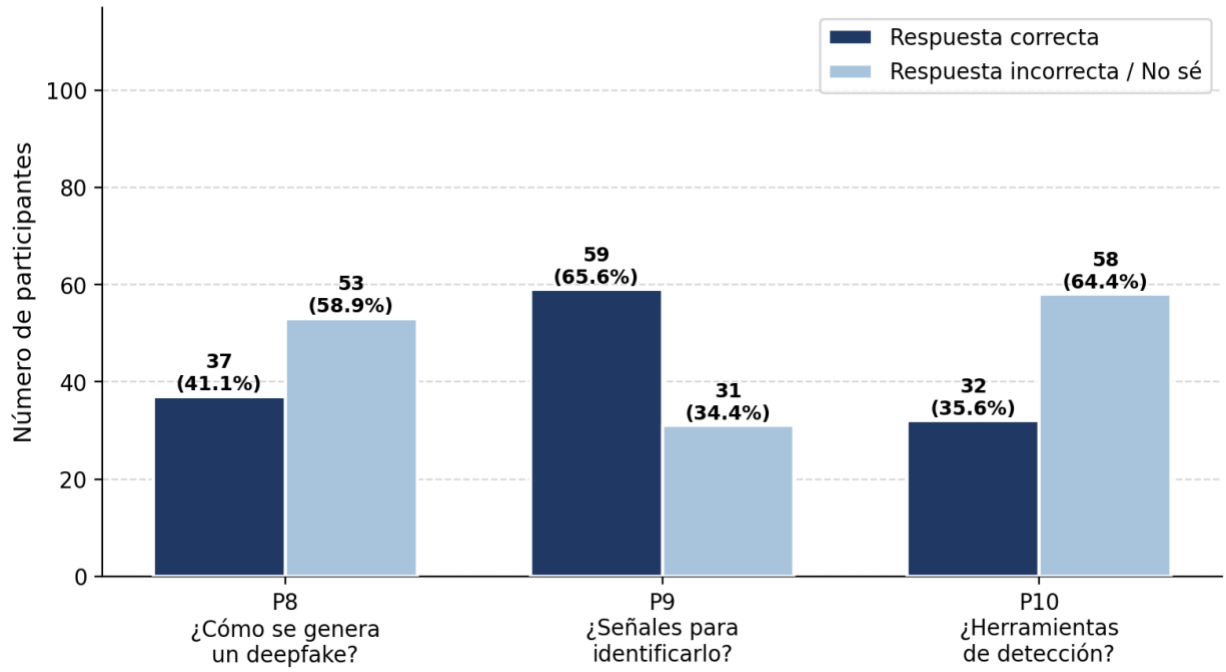
Puntaje en el índice (0–3)	Frecuencia	Porcentaje
0 respuestas correctas	21	23.3%
1 respuesta correcta	30	33.3%
2 respuestas correctas	19	21.1%
3 respuestas correctas	20	22.2%
Total	90	100%

Nota. Elaboración propia.

La confianza pública y la credibilidad de la información digital.

Figura 8

Comparación de respuestas correctas e incorrectas en los ítems de conocimiento objetivo.



Nota. Elaboración propia.

El índice de conocimiento objetivo arrojó una media de 1.42 sobre 3 (DE = 1.08), lo que indica un nivel bajo-básico. El 60% de los participantes desconoce herramientas de detección de deepfakes (P10), y el 40% no sabe cómo se genera este tipo de contenido (P8). La pregunta P9 tuvo el mayor acierto (65.6%), posiblemente porque la señal de los labios que no coinciden con el audio es intuitiva y se ha popularizado en medios. Esta brecha entre el conocimiento autopercibido (media = 3.08) y el objetivo (media = 1.42) confirma que los usuarios sobrestiman su capacidad real para detectar deepfakes.

P11. Estrategias de verificación ante contenido dudoso

Se consultó a los participantes qué hacen cuando dudan de si un contenido digital es real o manipulado, antes de creerlo o compartirlo.

La confianza pública y la credibilidad de la información digital.

Tabla 11

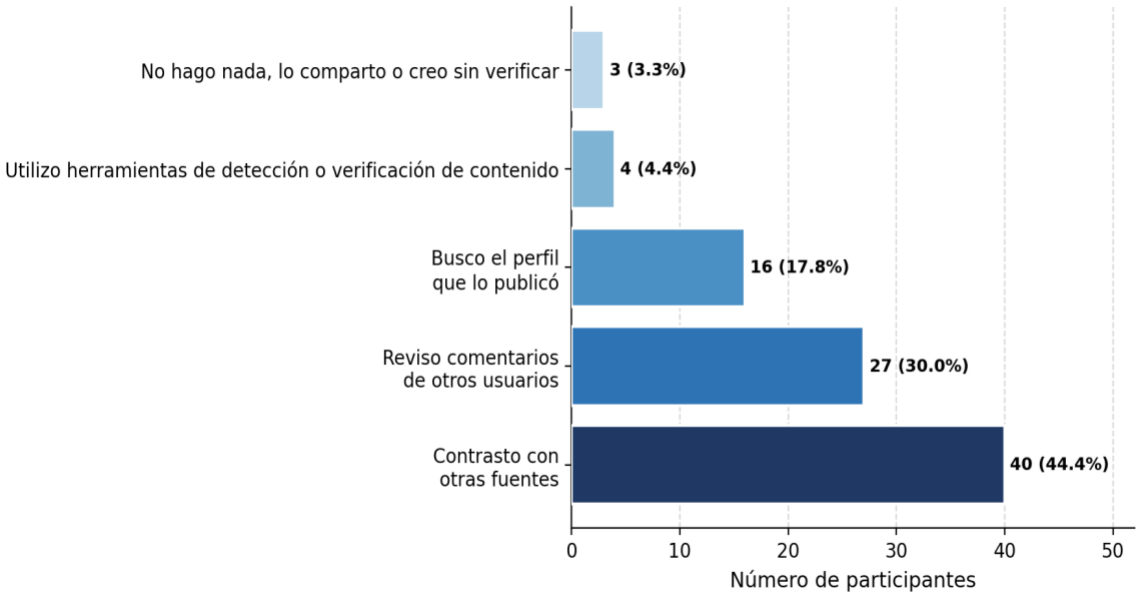
Estrategias de verificación empleadas ante contenido digital dudoso.

Estrategia de verificación	Frecuencia	Porcentaje
Contrasto con otras fuentes o medios de comunicación	40	44.4%
Reviso los comentarios o reacciones de otros usuarios	27	30.0%
Busco el nombre del medio o perfil que lo publicó	16	17.8%
Utilizo herramientas de detección o verificación	4	4.4%
No hago nada, lo comparto o creo sin verificar	3	3.3%
Total	90	100%

Nota. Elaboración propia.

Figura 9

Estrategias de verificación utilizadas por los participantes ante contenido dudoso



Nota. Elaboración propia.

La confianza pública y la credibilidad de la información digital.

El 44.4% contrasta con otras fuentes, que es la estrategia más robusta reportada. Sin embargo, el 30% se apoya en los comentarios de otros usuarios, lo cual no garantiza verificación real. Solo el 4.4% utiliza herramientas especializadas de detección, y el 3.3% no hace nada antes de creer o compartir. Estos datos evidencian que la mayoría de los usuarios depende de estrategias informales que no son suficientes para detectar deepfakes sofisticados.

P12. Plataforma con mayor exposición a deepfakes

Se indagó en qué plataforma los participantes han encontrado con mayor frecuencia contenido que sospechan o saben que es un deepfake.

Tabla 12

Plataforma digital con mayor exposición a deepfakes reportada por los participantes.

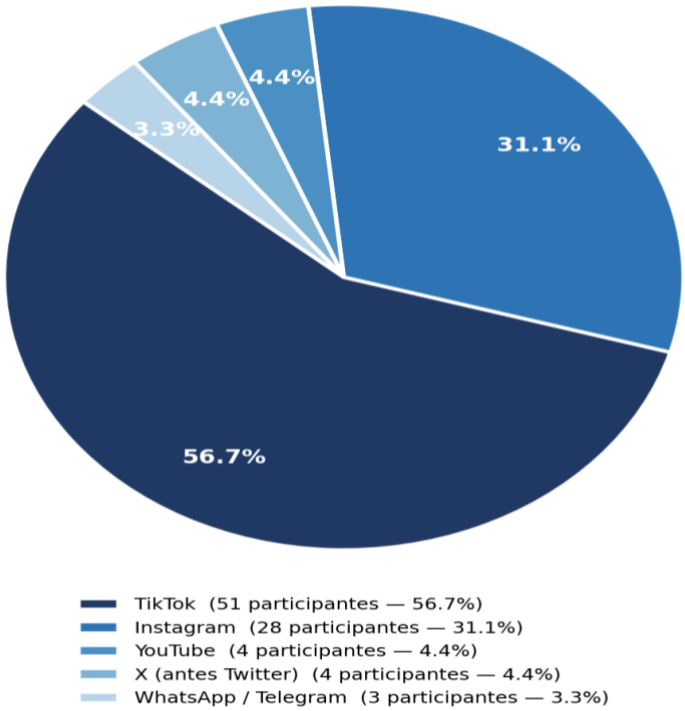
Plataforma	Frecuencia	Porcentaje
TikTok	51	56.7%
Instagram	28	31.1%
YouTube	4	4.4%
X (antes Twitter)	4	4.4%
WhatsApp / Telegram	3	3.3%
Total	90	100%

Nota. Elaboración propia.

La confianza pública y la credibilidad de la información digital.

Figura 10

Distribución porcentual de la plataforma con mayor exposición a deepfakes.



Nota. Elaboración propia.

TikTok concentra el 56.7% de la exposición reportada, seguido de Instagram con el 31.1%. Juntas, estas dos plataformas representan el 87.8% del total, lo que es coherente con la literatura revisada sobre el consumo masivo de contenido breve y dinámico en estas plataformas, especialmente entre usuarios jóvenes que no siempre cuentan con las herramientas críticas para evaluar la autenticidad del contenido que consumen.

12.2 Análisis Correlacional — Coeficiente de Spearman (ρ)

Se aplicó el coeficiente de correlación de Spearman (ρ) para establecer las relaciones entre variables ordinales, tal como se definió en el diseño metodológico. Se analizaron cuatro pares de variables relacionadas con los objetivos del estudio.

La confianza pública y la credibilidad de la información digital.

Tabla 13

Resultados del análisis correlacional de Spearman entre variables del estudio.

Relación analizada	ρ (rho)	p-valor	Resultado
Exposición a deepfakes (P6) → Confianza en medios digitales (P4)	-0.095	0.375	No significativa
Exposición a deepfakes (P6) → Credibilidad percibida (P5)	-0.078	0.468	No significativa
Conocimiento objetivo (P8–P10) → Confianza en medios digitales (P4)	0.213	0.044	Significativa ($p < 0.05$)
Conocimiento objetivo (P8–P10) → Credibilidad percibida (P5)	0.006	0.956	No significativa

Nota. Elaboración propia.

La exposición a deepfakes no mostró correlación estadísticamente significativa con la confianza ($\rho = -0.095$, $p = 0.375$) ni con la credibilidad percibida ($\rho = -0.078$, $p = 0.468$), ya que ambos p-valores superan el nivel de significancia $\alpha = 0.05$. Esto indica que la frecuencia de exposición, por sí sola, no determina el nivel de confianza o credibilidad de los usuarios colombianos en los medios digitales.

La única correlación estadísticamente significativa fue la encontrada entre el índice de conocimiento objetivo y la confianza en medios digitales ($\rho = 0.213$, $p = 0.044$). Esta correlación es positiva y de magnitud débil a moderada: a mayor conocimiento real sobre deepfakes, mayor tiende a ser la confianza en los medios digitales. Una interpretación posible es que los usuarios con mayor conocimiento técnico tienen mayor capacidad para discriminar el contenido manipulado del auténtico, lo que les permite mantener niveles de confianza más estables. La relación entre conocimiento objetivo y credibilidad percibida ($\rho = 0.006$, $p = 0.956$) no fue significativa.

La confianza pública y la credibilidad de la información digital.

Conclusiones

A partir del análisis descriptivo y correlacional de las respuestas obtenidas de 90 usuarios colombianos de medios digitales, se presentan las conclusiones del estudio:

Primera**.** La confianza y la credibilidad en los medios digitales son moderadas y tendientes al escepticismo. Las medias de 2.54 (confianza) y 2.79 (credibilidad) sobre una escala de 5 puntos ubican a los participantes en el rango medio definido en el estudio, con una inclinación hacia la desconfianza: el 42.3% indicó poca o ninguna confianza en la información digital, y ningún participante reportó confianza plena. Esto es coherente con lo señalado por Parra y Pinela (2025), quienes documentan que los deepfakes aceleran la pérdida de credibilidad en los medios digitales al minar el pensamiento crítico de los usuarios.

Segunda**.** La exposición a deepfakes es alta y frecuente, concentrada en TikTok e Instagram. El 67.8% de los participantes encontró este tipo de contenido al menos una vez por semana, y el 21.1% más de tres veces por semana. TikTok e Instagram concentraron el 87.8% de dicha exposición. Sin embargo, la frecuencia de exposición no presentó correlación estadísticamente significativa con la confianza ni con la credibilidad ($p > 0.05$), lo que indica que la mera exposición no explica por sí sola el nivel de desconfianza de los usuarios.

Tercera**.** Existe una brecha significativa entre el conocimiento autopercibido y el conocimiento objetivo sobre deepfakes. Si bien el conocimiento autopercibido fue moderado (media = 3.08), el índice de conocimiento objetivo apenas alcanzó una media de 1.42 sobre 3 puntos posibles. El 60% de los participantes no conoce herramientas de detección de deepfakes, y el 40% desconoce cómo se genera este tipo de contenido. Esta discrepancia evidencia que los usuarios colombianos sobrestiman su capacidad real para identificar deepfakes, lo que representa una vulnerabilidad frente a la desinformación.

Cuarta**.** El conocimiento objetivo sobre deepfakes actúa como factor protector de la confianza en los medios digitales. La correlación de Spearman entre el índice de conocimiento objetivo y la confianza fue la única estadísticamente significativa del estudio ($\rho = 0.213$, $p = 0.044$). Los usuarios con mayor conocimiento real sobre deepfakes tienden a mantener niveles de confianza más altos, posiblemente porque su alfabetización mediática les permite distinguir el contenido auténtico del manipulado con mayor precisión. Este hallazgo responde directamente al segundo y tercer objetivo específico del estudio.

Quinta**.** Las estrategias de verificación empleadas por los usuarios colombianos son predominantemente informales. El 44.4% contrasta con otras fuentes y el 30% revisa los

La confianza pública y la credibilidad de la información digital.

comentarios de otros usuarios, mientras que apenas el 4.4% utiliza herramientas especializadas de verificación o detección, y el 3.3% declaró no hacer nada antes de creer o compartir contenido dudoso. Esto confirma que los usuarios no cuentan con las competencias ni los recursos necesarios para enfrentar de manera efectiva la desinformación generada por deepfakes en el entorno digital colombiano.

Recomendaciones

Con base en los hallazgos del estudio, se formulan las siguientes recomendaciones para los distintos actores del ecosistema digital colombiano:

Para el sistema educativo

El estudio confirmó una brecha significativa entre el conocimiento autopercibido (media = 3.08) y el conocimiento objetivo real (media = 1.42 sobre 3) sobre deepfakes. Se recomienda incorporar contenidos de alfabetización mediática en todos los niveles educativos, con énfasis en la identificación práctica de contenido manipulado con inteligencia artificial y el uso de herramientas de verificación disponibles. El 60% de los encuestados desconoció la existencia de detectores especializados de deepfakes, lo que evidencia la urgencia de incluir esta información en los programas de competencias digitales, especialmente para los grupos de 18 a 35 años, que representan el 58.9% de la muestra y son los usuarios con mayor exposición a estas plataformas.

Para las plataformas digitales

TikTok e Instagram concentraron el 87.8% de la exposición a deepfakes reportada en el estudio. Se recomienda que estas plataformas implementen mecanismos de detección automatizada y etiquetado visible de contenido generado o modificado con inteligencia artificial, junto con acciones sostenidas de educación digital dirigidas a sus usuarios. El 46.7% de los encuestados indicó encontrar deepfakes entre 1 y 3 veces por semana en estas plataformas, lo que confirma que la intervención a este nivel tendría un impacto directo y medible sobre la población colombiana.

Para el sector regulatorio

La confianza pública y la credibilidad de la información digital.

Los resultados evidencian que los usuarios colombianos están expuestos frecuentemente a deepfakes sin contar con mecanismos suficientes de protección ni herramientas de detección accesibles. Se recomienda desarrollar un marco normativo específico sobre deepfakes en Colombia, tomando como referencia el Reglamento de Inteligencia Artificial de la Unión Europea (AI Act), que obliga al etiquetado de contenido sintético generado por IA y entró en vigor el 2 de agosto de 2025, y la Online Safety Act del Reino Unido, que establece obligaciones para las plataformas en relación con la eliminación de contenido ilegal generado mediante deepfakes. Asimismo, se recomienda establecer protocolos de coordinación entre plataformas digitales y entidades regulatorias para la detección y reporte oportuno de este tipo de contenido.

Para futuras investigaciones

Este estudio presenta limitaciones propias del muestreo no probabilístico por conveniencia, que no garantiza representatividad estadística de la totalidad de la población colombiana. Se recomienda replicar el estudio con un diseño de muestreo probabilístico y una muestra de mayor tamaño que permita generalizar los resultados. Dado que la exposición a deepfakes no mostró correlación significativa con la confianza ni la credibilidad, futuras investigaciones podrían explorar variables mediadoras o moderadoras no consideradas en este estudio, como el nivel de educación mediática previa, la frecuencia de uso de plataformas específicas o la motivación para verificar información. Asimismo, la brecha entre conocimiento autopercibido y objetivo identificada sugiere que metodologías cualitativas complementarias, como entrevistas en profundidad o grupos focales, podrían aportar una comprensión más rica de las razones detrás de esta discrepancia.

La confianza pública y la credibilidad de la información digital.

Referencias

Asobancaria. (2025, 21 de abril). Deepfake: una amenaza latente para el sistema financiero (Banca & Economía, n.° 1469). https://www.asobancaria.com/wp-content/uploads/2025/04/1469-BE.pdf

Ballesteros Aguayo, L. y Ruiz del Olmo, F. J. (2024). Vídeos falsos y desinformación ante la IA: el deepfake como vehículo de la posverdad [Fake videos and disinformation before the AI: deepfake as a post-truth vehicle]. *Revista de Ciencias de la Comunicación e Información*, 29, 1-14. https://doi.org/10.35742/rcci.2024.29.e294

Caldera-Serrano, J. (2025). Identidades audiovisuales post mortem. Entre la preservación de la memoria y la amenaza de la manipulación. *Información, Cultura y Sociedad*, 53. https://doi.org/10.34096/ics.i53.16755

Chandra, N. A., Murtfeldt, R., Qiu, L., Karmakar, A., Lee, H., Tanumihardja, E., Farhat, K., Caffee, B., Paik, S., Lee, C., Choi, J., Kim, A. y Etzioni, O. (2025). Deepfake-Eval-2024: A multi-modal in-the-wild benchmark of deepfakes circulated in 2024. *arXiv,* 1(4). https://doi.org/10.48550/arXiv.2503.02857

Cotino Hueso, L. (2022). Quién, cómo y qué regular (o no regular) frente a la desinformación. *Teoría y Realidad Constitucional*, 49, 199-238. https://doi.org/10.5944/trc.49.2022.33849

Deloitte Center for Financial Services. (2024). Deepfake banking fraud risk on the rise. Deloitte. https://www.deloitte.com/us/en/insights/industry/financial-services/deepfake-banking-fraud-risk-on-the-rise.html

EFE. (2023, 25 de abril). FundéuRAE: "ultrafalso", alternativa a "deepfake". ProQuest Wire Feeds. https://www.proquest.com/wire-feeds/fundéurae-ultrafalsoalternativa-deepfake/docview/2805403299/se-2

Euronews. (2025, 23 de diciembre). Video of coup in Paris: How an AI-generated video caused Macron a major headache. Euronews. https://www.euronews.com/my-europe/2025/12/19/video-of-coup-in-paris-how-an-ai-generated-video-caused-macron-a-major-headache

Garimella, K. y Eckles, D. (2020). Images and misinformation in political groups: Evidence from WhatsApp in India. *Harvard Kennedy School Misinformation Review*, 1(5). https://doi.org/10.37016/mr-2020-030

Garriga, M., Ruiz-Incertis, R. y Magallón-Rosa, R. (2024). Inteligencia artificial, desinformación y propuestas de alfabetización en torno a los deepfakes. *Observatorio (OBS*),* 18(5), 175-194. https://doi.org/10.15847/obsOBS18520242445

González Arencibia, M. y Martínez Cardero, D. (2024). Soluciones educativas frente a los dilemas éticos del uso de la tecnología deep fake. *Revista Internacional de Filosofía Teórica y Práctica*, 1(1), 99-126. https://doi.org/10.51660/riftp.v1i1.22

La confianza pública y la credibilidad de la información digital.

Khalil, M. (2025, 8 de septiembre). Deepfake statistics 2025: AI fraud data & trends. DeepStrike. https://deepstrike.io/blog/deepfake-statistics-2025

McAfee. (2024, 21 de agosto). McAfee lanza el primer detector de "deepfake" automático y con inteligencia artificial del mundo, exclusivo para las nuevas PC Lenovo con IA. Business Wire en Español. https://www.proquest.com/wire-feeds/mcafee-lanza-el-primer-detector-de-deepfake/docview/3095011236/se-2

Mirsky, Y. y Lee, W. (2022). La creación y detección de deepfakes: una encuesta. *ACM Computing Surveys,* 54(1), 1-41. https://doi.org/10.1145/3425780

Muñoz Medina, L. (2025, 11 de febrero). Ciberseguridad en Colombia: aumento de deepfakes pornográficos genera alarma. Infobae Colombia. https://www.infobae.com/colombia/2025/02/11/ciberseguridad-en-colombia-aumento-de-deepfakes-pornograficos-genera-alarma/

Parra, N. K. S. y Pinela, Á. V. S. (2025). Impacto de los deepfakes en la confianza pública y la credibilidad de los medios de comunicación: una revisión sistemática. Arandu UTIC, 12(3), 2098-2112. https://dialnet.unirioja.es/servlet/articulo?codigo=10572397

Petrino, G. (2025, 4 de diciembre). 2024 deepfakes guide and statistics. Security.org. https://www.security.org/resources/deepfake-statistics/

Poireault, K. (2025, 25 de noviembre). AI and deepfake-powered fraud skyrockets amid identity fraud stagnation. Infosecurity Magazine. https://www.infosecurity-magazine.com/news/ai-deepfake-fraud-skyrockets/

Regula Forensics. (2024, 31 de octubre). Deepfake fraud costs the financial sector an average of $600,000 for each company. https://regulaforensics.com/news/deepfake-fraud-costs/

Sánchez-Acedo, A., Carbonell-Alcocer, A., Gertrudix, M. y Rubio-Tamayo, J. L. (2024). Retos de la alfabetización mediática e informacional en la ecología de la inteligencia artificial: deepfakes y desinformación. *Communication & Society,* 37(4), 223-239. https://doi.org/10.15581/003.37.4.223-239

Serrano, J. Y. M., De la Toba Noriega, C. E., Zatarain, S. C., Contreras, L. Y. A. y Lucas, G. Á. D. (2025). Deepfakes: Revisión sistemática de tecnologías, impacto y estrategias de detección. *Revista de Investigación en Tecnologías de la Información,* 13(29), 19-37. https://riti.es/index.php/riti/article/view/335

Suastegui, M. I. V., Pico, L. A. V., Pinargote, M. G. C., Romero, D. L. R. y Gorozabel, O. A. L. (2026). La incidencia de la inteligencia artificial generativa en los deepfakes y sus repercusiones en la confianza social. *Ciencia y Educación,* 7(1.1), 677-685. https://www.cienciayeducacion.com/index.php/journal/article/view/2304/3020

La confianza pública y la credibilidad de la información digital.

Sumsub. (2025, 25 de noviembre). Identity fraud report 2025-2026: The sophistication shift. https://sumsub.com/fraud-report-2025/

Surfshark. (2025, 15 de abril). Deepfake statistics in early 2025: how frequently are famous people targeted? Surfshark Research. https://surfshark.com/research/study/deepfake-statistics

Surfshark. (2025, 29 de julio). Global map: Election-related deepfakes reached 3.8B people. https://surfshark.com/research/chart/election-related-deepfakes

Thorn. (2025, marzo). Deepfake nudes & young people: Navigating a new frontier in technology-facilitated nonconsensual sexual abuse and exploitation. https://info.thorn.org/hubfs/Research/Thorn_DeepfakeNudes&YoungPeople_Mar2025.pdf

Tolosana, R., Vera-Rodríguez, R., Fierrez, J., Morales, A. y Ortega-García, J. (2020). DeepFakes and beyond: A survey of face manipulation and fake detection. *Information Fusion*, 64, 131-148. https://doi.org/10.1016/j.inffus.2020.06.014