DOI: 10.25768/1646-4974n41v2a05

Inteligencia Artificial y Desinformación: Tensiones Éticas y Estrategias Sociotécnicas en la Era Digital

Moisés Limia Fernández
Universidad Jorge Tadeo Lozano, Colombia
moises.limiaf@utadeo.edu.co

Resumen

Este artículo examina críticamente la relación entre inteligencia artificial (IA) y desinformación, analizando cómo estas tecnologías emergentes transforman los modos de producción, circulación y validación de la verdad en entornos digitales. A partir del estudio en profundidad de tres casos reales —un *deepfake* del presidente Zelensky, el sistema automatizado AI4Media y la diseminación de imágenes generadas por IA durante el conflicto en Gaza (2023)— se identifican tensiones estructurales entre automatización, ética y poder cognitivo. Los hallazgos muestran que la IA actúa como infraestructura epistémica ambivalente, capaz de amplificar tanto la falsedad estratégica como la verificación automatizada, sin garantizar por sí misma justicia cognitiva. Se argumenta que el combate a la desinformación requiere no solo marcos técnicos, sino una nueva arquitectura de gobernanza pública, pluralidad epistémica y justicia algorítmica. El texto propone una agenda crítica para repensar la autoridad informativa en la era posfotográfica y algorítmica.

Palabras clave: inteligencia artificial, desinformación, ética tecnológica, algoritmos, justicia epistémica, posverdad.

Artificial Intelligence and Disinformation: Ethical Tensions and Epistemic Disputes in the Algorithmic Era

Abstract

This article critically examines the relationship between artificial intelligence (AI) and disinformation, analyzing how these emerging technologies reshape the processes of truth production, circulation, and validation in digital environments. Drawing on an in-depth analysis of three real cases—a deepfake of President Zelensky, the AI4Media automated fact-checking system, and the dissemination of AI-generated images during the Gaza conflict (2023)—the study identifies structural tensions between automation, ethics, and cognitive power. Findings show that AI functions as an ambivalent epistemic infrastructure, capable of enhancing both strategic falsehood and automated verification, without inherently ensuring cognitive justice. It is argued that combating disinformation demands not only technical frameworks but a new architecture of public governance, epistemic plurality, and algorithmic justice. The paper proposes a critical agenda to rethink informational authority in the post-photographic and algorithmic era.

Keywords: artificial intelligence, disinformation, tech ethics, algorithms, epistemic justice, post-truth.

Data de submissão: 2025-06-30. Data de aprovação: 2025-10-02.

Revista Estudos em Comunicação é financiada por Fundos FEDER através do Programa Operacional Factores de Competitividade – COMPETE e por Fundos Nacionais através da FCT – Fundação para a Ciência e a Tecnologia no âmbito do projeto *LabCom – Comunicação e Artes, UIDB/00661/2020.*

1. Introducción

Vivimos en una época en la que los flujos de información se han acelerado a una velocidad sin precedentes, transformando radicalmente los modos de producción, circulación y recepción del conocimiento. En este nuevo ecosistema digital, la desinformación ha emergido como una de las amenazas más críticas para la gobernabilidad democrática, la cohesión social y la deliberación pública. Fenómenos como la manipulación electoral, la polarización discursiva, la propagación de teorías conspirativas o la deslegitimación del conocimiento científico encuentran en los entornos digitales un terreno fértil para su expansión. A esta complejidad se suma un nuevo actor tecnológico de alcance estructural: la inteligencia artificial (IA).

Lejos de ser una herramienta neutral, la IA desempeña hoy un papel ambivalente y disruptivo en los procesos de comunicación contemporáneos. Por un lado, ha sido utilizada para crear contenidos manipulados de alta sofisticación, como imágenes generadas por redes generativas adversariales (GANs), audios y videos falsos hiperrealistas —los llamados *deepfakes*—, así como textos automatizados indistinguibles de los escritos por humanos. Por otro lado, la IA también se ha convertido en un aliado estratégico para detectar, rastrear y contrarrestar narrativas falsas, mediante sistemas de aprendizaje automático, algoritmos de detección semántica y plataformas de verificación automatizada.

Este doble rol de la IA configura una tensión estructural en la lucha contra la desinformación: la misma tecnología que permite amplificar las falsedades se postula, al mismo tiempo, como herramienta para combatirlas. Esta paradoja plantea preguntas urgentes sobre los límites éticos, la gobernanza algorítmica, la transparencia tecnológica y la distribución de poder entre actores públicos y privados. ¿Puede una infraestructura algorítmica desarrollada por corporaciones tecnológicas —con incentivos económicos y lógicas opacas— garantizar una gestión ética de la información? ¿Qué marcos regulatorios, epistemológicos y sociotécnicos son necesarios para encauzar un uso responsable de la IA en el ecosistema mediático? ¿Cómo equilibrar innovación y responsabilidad social en contextos marcados por desigualdades tecnológicas, culturales y normativas?

Este artículo aborda dichas tensiones desde una perspectiva crítica, proponiendo que la efectividad de la IA en el combate a la desinformación no depende únicamente de su sofisticación técnica, sino de su inserción en marcos éticos sólidos, desarrollos responsables y articulaciones interdisciplinares. Con base en una metodología cualitativa y en el análisis de estudios de caso seleccionados, examinaremos cómo la IA ha sido empleada tanto para la creación como para la detección de campañas de desinformación en distintos contextos sociopolíticos. El objetivo es identificar los factores que amplifican o mitigan su impacto, evaluar los dilemas éticos asociados y proponer orientaciones para un desarrollo tecnológico más justo y transparente.

En este sentido, el artículo se estructura en seis apartados. Tras esta introducción, se presenta el marco teórico, que revisa críticamente los conceptos de desinformación, inteligencia artificial y ética tecnológica. A continuación, se expone la estrategia metodológica adoptada. Luego, se analizan tres estudios de caso representativos que ilustran el uso dual de la IA en relación con la desinformación. El apartado de discusión interpreta los hallazgos a la luz de las tensiones éticas y sociales identificadas. Finalmente, se presentan las conclusiones, con recomendaciones dirigidas a tecnólogos, comunicadores, formuladores de políticas y académicos comprometidos con el diseño de un entorno informativo más equitativo y confiable.

2. Marco Teórico

La desinformación potenciada por inteligencia artificial (IA) constituye uno de los desafíos más complejos del ecosistema comunicativo contemporáneo. Su abordaje exige una mirada transdisciplinaria que combine la teoría crítica de la comunicación, la sociología del conocimiento, la ética de las tecnologías emergentes y los estudios algorítmicos. Este marco teórico se articula en torno a cuatro ejes conceptuales: (1) la desinformación digital como fenómeno comunicacional estratégico y estructural, (2) la inteligencia artificial como infraestructura generativa y moduladora de la verdad, (3) los fundamentos normativos y límites prácticos de la ética tecnológica en entornos opacos y automatizados, y (4) las tensiones epistémicas derivadas de la delegación algorítmica de la autoridad sobre la verdad.

2.1. Desinformación como estrategia comunicativa estructural en la era de la atención algorítmica

La desinformación no puede entenderse exclusivamente como un contenido falso o erróneo. Debe ser conceptualizada como una estrategia semiótica, intencional y adaptativa, orientada a producir efectos performativos en el plano emocional, cognitivo y sociopolítico (Jack, 2017; Gelfert, 2018). En un contexto de sobreabundancia informativa, su eficacia no depende tanto de su veracidad como de su capacidad para insertarse en narrativas culturales preexistentes, amplificadas por mecanismos algorítmicos que priorizan el engagement sobre la exactitud (Pariser, 2011; Vosoughi, Roy & Aral, 2018).

Wardle y Derakhshan (2017) introdujeron una tipología ampliamente adoptada —misinformation, disinformation y malinformation— que ha contribuido a clarificar las intenciones y mecanismos que subyacen al fenómeno. Sin embargo, más allá de esa taxonomía, otros enfoques destacan su dimensión estructural. Bennett y Livingston (2018) plantean que vivimos en un *orden de desinformación* caracterizado por la degradación de la autoridad epistémica, la fragmentación del espacio público y la circulación estratégica de falsedades como táctica de erosión de la deliberación democrática. Investigaciones recientes confirman que la susceptibilidad a la desinformación está modulada por factores demográficos y psicológicos, más allá de la simple exposición algorítmica (Sultan et al., 2024). Asimismo, se observa que los públicos reaccionan con escepticismo ante etiquetas de IA incluso cuando los contenidos son verídicos, lo que revela una creciente desconfianza hacia la mediación tecnológica (Altay, Dohle, & De Keersmaecker, 2024).

Desde esta perspectiva, la desinformación no se opone simplemente a la verdad, sino que compite por legitimidad y resonancia en un ecosistema dominado por la lógica de la viralidad (Marwick & Lewis, 2017). Gillespie (2018) acuñó el concepto de *curaduría algorítmica* para describir cómo las plataformas no sólo median, sino que configuran activamente qué información se visibiliza, se ignora o se elimina. Las plataformas se han convertido en intermediarios epistémicos (Madaio et al., 2020), cuyo diseño condiciona el tipo de contenidos que circulan, y con ello, los marcos de lo pensable y lo debatible.

Al asumir esto, se hace evidente que la desinformación no es una anomalía del sistema, sino una manifestación funcional del régimen de información neoliberal, donde la atención se ha convertido en el bien escaso, y la economía de plataformas privilegia el contenido emocional, simplificado y divisivo (Zuboff, 2019; Napoli, 2019). Las campañas de desinformación, en este marco, son una forma de "capitalización simbólica", en la que la verdad deviene contingente, maleable y subordinada a lógicas de influencia.

2.2. Inteligencia artificial como infraestructura generativa y moduladora de lo verosímil

La IA no es meramente una tecnología instrumental. Se ha convertido en un actor epistémico, en tanto participa activamente en la producción, organización y validación de contenidos que informan nuestras creencias y juicios sobre el mundo (Floridi, 2020). Su papel en la desinformación debe analizarse en clave ambivalente: como tecnología de creación y como tecnología de vigilancia y detección.

En el ámbito de la creación, la IA ha potenciado una nueva categoría de falsedad: la desinformación sintética, aquella generada sin intervención humana directa, mediante modelos generativos como GPT-4, DALL·E 3, StyleGAN o Sora. Estas tecnologías permiten la producción masiva de textos, imágenes, audios o videos hiperrealistas, diseñados para engañar cognitivamente y emocionar estratégicamente. Zellers et al. (2019) mostraron que Grover —modelo entrenado para detectar desinformación— era igualmente eficaz para producir textos falsos convincentes. El principio de *dual use* (uso dual) se vuelve central: la misma tecnología sirve tanto para generar falsedad como para combatirla.

En cuanto a la detección, la IA se ha aplicado al desarrollo de herramientas de fact-checking automatizado, análisis de patrones lingüísticos, rastreo de redes de bots y evaluación de fuentes (Pérez--Rosas et al., 2018). La literatura reciente confirma que los detectores han evolucionado desde enfoques unimodales hacia arquitecturas multimodales que integran señales visuales, de voz y textuales (Heidari et al., 2024; Wang et al., 2024). Estas soluciones prometen eficiencia y escalabilidad, pero enfrentan importantes limitaciones: sesgos de entrenamiento, falta de transparencia, contextos culturales no codificables y desafíos semánticos como la ironía, la ambigüedad o la sátira (Graves, 2018). Además, persisten problemas de generalización fuera del laboratorio —especialmente con datos de baja calidad o manipulaciones adversariales— que degradan el rendimiento en condiciones reales (Balafrej et al., 2024). Tampoco puede suponerse que la percepción humana compense estas carencias: incluso observadores entrenados exhiben bajos niveles de acierto al identificar deepfakes, lo que evidencia la fragilidad del juicio humano como instancia de verificación (Diel et al., 2024).

Además, la IA opera como infraestructura epistémica (Leonelli, 2016): una red de arquitecturas, datos, normas y lógicas que determinan qué se considera conocimiento válido, qué circula, qué se oculta y quién decide. Esta función infraestructural tiene consecuencias políticas profundas: modela los marcos de visibilidad, la arquitectura del disenso y los límites de la imaginación pública.

2.3. Ética tecnológica y gobernanza algorítmica: promesas, principios y paradojas

La gobernanza ética de la IA ha sido objeto de un boom normativo en los últimos años. Documentos como los *Principios de Montreal* (2018), las *Guías para una IA confiable* de la Comisión Europea (2019), el *Marco Ético de la UNESCO* (2021) o las propuestas del *AI Now Institute* promueven valores como:

- Transparencia (explainability);
- Justicia algorítmica (fairness);
- No discriminación y protección de derechos humanos;
- Supervisión humana (human oversight);
- Sustentabilidad y rendición de cuentas.

Sin embargo, como advierten Mittelstadt et al. (2016), existe un desfase entre el plano declarativo de los principios éticos y el plano material de la implementación en sistemas comerciales, opacos, desarrollados por corporaciones transnacionales. Boddington (2017) señala que la ética tecnológica, para no caer en un ejercicio cosmético (*ethics washing*), debe asumir una mirada crítica, situada y operativa, capaz de articular tensiones estructurales entre innovación, poder y responsabilidad.

En el campo de la desinformación, el problema se agudiza: si los mismos agentes que diseñan tecnologías capaces de propagar contenido falso desarrollan también las herramientas para detectarlo y moderarlo, se instala un conflicto de intereses sistémico. Como indica O'Neil (2016), sin mecanismos de auditoría externa y rendición de cuentas efectiva, los algoritmos pueden convertirse en *armas de destrucción matemática*, es decir, sistemas que perpetúan inequidades bajo la apariencia de eficiencia.

Por ello, la ética no puede limitarse a principios de diseño; debe incorporar la dimensión política de la IA: ¿quién tiene el poder de decidir qué contenido es "aceptable"? ¿Qué ideologías se codifican como neutralidad? ¿Qué saberes quedan fuera del modelo? Sin una gobernanza participativa y democrática de los sistemas inteligentes, el riesgo de tecnocratización autoritaria del espacio informativo se vuelve tangible (Couldry & Mejias, 2019).

2.4. Automatización de la verdad, disputas epistémicas y justicia cognitiva

El proceso mediante el cual la IA asume funciones de verificación, curaduría y moderación de contenido implica una delegación epistémica sin precedentes. La autoridad sobre la verdad, históricamente disputada entre instituciones como el periodismo, la ciencia y el Estado, comienza a transferirse a sistemas automáticos, entrenados con corpus sesgados, bajo criterios opacos, y en contextos culturales particulares.

Desde una perspectiva de epistemología crítica, esto plantea un escenario de injusticia epistémica automatizada (Fricker, 2007; Medina, 2012): ciertas voces, saberes y comunidades quedan sistemáticamente excluidas de los marcos de validación de lo verdadero. Noble (2018) ha documentado cómo los algoritmos de búsqueda pueden reproducir sexismo, racismo y clasismo estructurales, mientras que investigaciones recientes muestran que los sistemas de clasificación algorítmica pueden reproducir y profundizar desigualdades sociales estructuradas (Gerdon et al., 2022). De manera más reciente, estudios empíricos han mostrado que las etiquetas que advierten sobre contenidos generados por IA tienden a reducir la credibilidad incluso de información verídica, lo que genera un efecto paradójico de erosión de confianza pública (Li et al., 2024; Wittenberg et al., 2025).

La automatización de la veracidad implica también una transformación en la naturaleza de lo que consideramos conocimiento. Como apunta Boyd (2019), la IA no "descubre" verdades; optimiza correlaciones según patrones estadísticos extraídos del pasado. Esta lógica puede reforzar el *status quo*, invisibilizar el disenso, y clausurar la posibilidad de verdades emergentes, plurales, disidentes o contrahegemónicas. En este sentido, se advierte que los marcos regulatorios actuales, como la AI Act, todavía presentan vacíos críticos que pueden consolidar nuevas formas de injusticia epistémica en la gestión algorítmica de la verdad (Labuz, 2024).

Desde los estudios decoloniales, autores como Catherine D'Ignazio y Lauren Klein (2020) han propuesto el concepto de *justicia de datos* para contrarrestar la concentración epistémica en el norte global y promover infraestructuras inclusivas, auditables y pluralistas. Esta perspectiva exige desautomatizar la autoridad, recuperar el juicio humano y garantizar la deliberación democrática sobre qué es, y quién define, lo verdadero.

3. Metodología

La presente investigación adopta un enfoque cualitativo de carácter crítico-interpretativo, orientado al análisis profundo y situado de los modos en que la inteligencia artificial (IA) interviene en la producción, circulación y contención de desinformación en contextos digitales. El estudio se enmarca en una lógica inductiva y relacional, que no busca establecer generalizaciones estadísticamente representativas, sino interpretaciones contextualmente robustas, teóricamente informadas y éticamente conscientes.

La metodología se estructura en torno a cuatro componentes centrales: (1) el enfoque teórico-metodológico general, (2) la lógica de selección y reconstrucción de casos, (3) las técnicas específicas de análisis utilizadas, y (4) las consideraciones éticas y epistemológicas que atraviesan todo el proceso investigativo.

3.1. Enfoque cualitativo crítico-interpretativo

El enfoque cualitativo se fundamenta en la premisa de que los fenómenos sociales, culturales y tecnológicos —como la desinformación mediada por IA— no pueden ser comprendidos únicamente mediante mediciones cuantitativas o reduccionismos técnico-causales. Como han señalado Denzin y Lincoln (2018), la investigación cualitativa busca capturar significados, tensiones, ambivalencias y relaciones simbólicas que emergen de contextos dinámicos y estructuralmente mediados.

En este estudio, dicho enfoque se articula con una perspectiva crítica, en tanto se asume que las tecnologías digitales no son neutrales, sino que materializan relaciones de poder, asimetrías epistémicas y decisiones normativas (Couldry & Mejias, 2019; Fuchs, 2017). La investigación crítica, en este sentido, no solo describe, sino que interroga: ¿a quién beneficia una determinada configuración tecnológica? ¿Qué narrativas legitima y cuáles excluye? ¿Cómo se articulan en ella la autoridad epistémica y la racionalidad técnica?

Asimismo, el carácter interpretativo del enfoque reconoce que los objetos de estudio no son "datos dados", sino construcciones que emergen del cruce entre el discurso, la tecnología y el contexto. El análisis, por tanto, no se limita a identificar contenidos falsos o verdaderos, sino a explorar cómo se producen las condiciones de posibilidad de lo verosímil, lo creíble y lo éticamente aceptable en la ecología mediática actual.

3.2. Estrategia metodológica basada en estudios de caso

Para operacionalizar esta perspectiva, se adopta el método del estudio de casos múltiples (Stake, 2005; Yin, 2018), entendiendo cada caso como una unidad singular que, al ser analizada en profundidad, permite iluminar patrones más amplios de sentido. La lógica de esta estrategia no es la comparación formal ni la inferencia probabilística, sino la generalización analítica (Flyvbjerg, 2006), es decir, la formulación de proposiciones conceptualmente relevantes a partir de la reconstrucción intensiva de situaciones concretas.

Se seleccionarán, en principio, tres casos contrastivos, siguiendo una lógica de muestreo teórico (Glaser & Strauss, 1967), y no estadístico. Bajo este enfoque, los casos no son tratados como unidades representativas de una población, sino como nodos intensivos de significado, capaces de revelar lógicas estructurales, dilemas éticos y regímenes epistémicos en torno a la articulación entre inteligencia artificial y desinformación. Esta selección responde a los siguientes criterios de inclusión:

• Centralidad de la IA en el fenómeno analizado, ya sea en la generación, amplificación o contención de la desinformación. Se consideraron únicamente aquellos casos en los que la IA juega un papel determinante en la generación, modulación, amplificación o contención de contenidos desinformativos. Esto excluye casos en los que la desinformación se propaga por medios tradicionales o exclusivamente humanos, sin intervención técnica significativa.

• Diversidad contextual, para incluir casos provenientes de distintos marcos sociopolíticos, regulatorios y mediáticos. Esta diversidad permite analizar cómo varía la operación de la IA según el contexto geográfico, el conflicto específico, el marco legal o la plataforma digital implicada. Se consideró, por ejemplo, la diferencia entre el uso de IA en situaciones de guerra híbrida, contextos de gobernanza tecnológica o conflictos de alta carga simbólica.

• Documentación verificable, que permita reconstruir el caso mediante fuentes confiables, múltiples y triangulables. Esta decisión metodológica apunta a evitar la dependencia de versiones únicas o sesgadas, y permite cruzar relatos desde distintas posiciones discursivas.

• Valor heurístico, es decir, capacidad del caso para generar hipótesis interpretativas sobre la articulación entre tecnología, discurso y ética. Por ello, se escogieron casos que, por su complejidad o singularidad, ofrecen un potencial elevado para generar hipótesis interpretativas sobre los modos en que se entrelazan tecnología, discurso, afecto, normatividad y epistemología en la era digital. No se trata de seleccionar los casos "más graves" o "más virales", sino aquellos que permiten pensar más allá de sí mismos.

Los casos se categorizaron en tres grandes tipos:

• Tipo A: Casos en los que la IA ha sido utilizada como herramienta de creación y propagación de desinformación (por ejemplo, uso de deepfakes en procesos electorales o generación automatizada de narrativas conspirativas).

• Tipo B: Casos en los que la IA ha sido utilizada como dispositivo de detección, moderación o contención de la desinformación (por ejemplo, fact-checking automatizado, intervención algorítmica de plataformas, políticas públicas tecnorregulatorias).

• Tipo C: Casos de producción emocional de lo falso mediante IA sin anclaje factual directo. En estos escenarios, la inteligencia artificial genera contenidos —como imágenes o audios— orientados a movilizar afectos intensos (compasión, indignación, empatía), sin referirse a hechos concretos. Funcionan como "pruebas emocionales" en disputas simbólicas altamente polarizadas.

Cada caso será reconstruido mediante un protocolo analítico que articula información proveniente de diversas fuentes: literatura académica, informes técnicos, reportes de medios especializados, entrevistas en profundidad (cuando sea posible) y análisis de documentos públicos. Esta triangulación de fuentes busca no solo garantizar la validez empírica, sino también contrastar narrativas institucionales, mediáticas y tecnológicas.

3.3. Técnicas de análisis

El análisis de los casos se llevará a cabo mediante un método mixto de análisis temático y discursivo, que permite identificar patrones recurrentes, estructuras de significado, contradicciones internas y relaciones de poder inscritas en las prácticas mediadas por IA.

a) Análisis temático inductivo

Se aplicará un análisis temático según el procedimiento propuesto por Braun y Clarke (2006), en seis fases:

1. Lectura intensiva y familiarización con el material;
2. Codificación abierta de unidades significativas;
3. Agrupamiento de códigos en temas provisionales;
4. Revisión crítica de temas y subtemas;
5. Definición e interpretación teórica de los temas;
6. Redacción narrativa de los hallazgos.

Los temas se organizarán en función de las dimensiones teóricas previamente definidas (creación, circulación, detección, ética, efectos sociales, visibilidad epistémica).

b) Análisis crítico del discurso mediado por IA

Se aplicará un enfoque inspirado en el análisis del discurso crítico (Fairclough, 2001; Wodak & Meyer, 2009), con especial énfasis en:

- Las narrativas de legitimación de la IA en contextos de desinformación;
- Las representaciones discursivas de la "verdad", el "riesgo", el "control" y la "autoridad";
- Las relaciones entre discurso institucional (gubernamental, empresarial) y discurso mediático;
- Las tensiones entre automatización técnica y juicio humano.

Cuando sea pertinente (por ejemplo, en el análisis de deepfakes o campañas visuales automatizadas), se incorporará también análisis semiótico y multimodal.

3.4. Consideraciones éticas y reflexividad epistemológica

Aunque esta investigación no involucra trabajo de campo directo con sujetos humanos, se asume un conjunto de principios éticos indispensables para el tratamiento riguroso de la información:

- Evitar la reproducción acrítica de contenidos engañosos, sensacionalistas o manipulativos;
- Contextualizar y problematizar los materiales analizados, subrayando siempre su función estratégica y su potencial daño;
- Proteger la dignidad y el anonimato de las personas eventualmente involucradas en los casos documentados (especialmente víctimas o denunciantes);
- Declarar los límites y sesgos posibles del investigador, en cuanto a posicionamiento teórico y marco cultural.

Desde una perspectiva epistemológica, se adopta una postura de reflexividad crítica (Bourdieu & Wacquant, 1992), que reconoce que toda producción de conocimiento implica decisiones interpretativas, selecciones metodológicas y compromisos ético-políticos. La desinformación mediada por IA no es solo un objeto a estudiar, sino un campo de disputa sobre qué cuenta como verdad, qué narrativas se privilegian y qué tecnologías se legitiman como garantes del conocimiento.

Por ello, esta investigación no se propone como neutral, sino como posicionada y responsable: busca contribuir a una comprensión crítica de los efectos sociales, cognitivos y éticos de la inteligencia artificial en los procesos de formación de la opinión pública y construcción de sentido en la era digital.

4. Análisis de Casos

La ambivalencia funcional de la inteligencia artificial en su relación con la desinformación se materializa en fenómenos empíricos de alta complejidad, que involucran múltiples niveles de análisis: tecnológico, discursivo, ético, geopolítico y afectivo. Para explorar esta tensión estructural, se han seleccionado tres casos paradigmáticos, reales y bien documentados, que representan momentos clave en la historia reciente de la articulación entre IA y desinformación: uno centrado en la generación de contenido falso con fines estratégicos, otro en el uso institucional de IA para la verificación automatizada, y un tercero en la difusión emocional y viral de imágenes generadas por IA en contextos de crisis humanitaria.

Cada caso se analiza en cinco dimensiones: (1) contexto, (2) arquitectura tecnológica, (3) actores, (4) dinámica de circulación y modulación algorítmica, y (5) dilemas ético-epistémicos.

4.1. Caso A: Deepfake de Zelensky y la instrumentalización de la IA como arma de guerra cognitiva

1. Contexto sociopolítico

En marzo de 2022, a escasas semanas del inicio de la invasión rusa a Ucrania, comenzó a circular en redes sociales un video aparentemente institucional donde el presidente ucraniano Volodymyr

Zelensky declaraba la rendición de las fuerzas ucranianas. Aunque el contenido fue rápidamente desmentido por verificadores independientes, agencias de inteligencia y el propio Zelensky mediante una transmisión oficial, su breve pero intensa circulación convirtió el caso en un punto de inflexión en la historia de las guerras cognitivas modernas.

Este episodio no se inscribe en la lógica tradicional de la propaganda, sino en lo que Rid (2020) denomina "actividades encubiertas de disuasión narrativa", diseñadas no tanto para convencer sino para erosionar la capacidad colectiva de juicio. Como recuerda Pomerantsev (2019), la desinformación contemporánea no busca imponer una mentira única, sino generar una atmósfera de desorientación moral y fatiga cognitiva. En este sentido, el deepfake de Zelensky se inscribe en una estrategia de guerra informativa híbrida, donde lo digital se convierte en extensión del campo de batalla físico.

Investigaciones recientes confirman que la vulnerabilidad no radica únicamente en la sofisticación técnica de los deepfakes, sino también en las limitaciones perceptivas humanas: incluso observadores entrenados fallan de manera sistemática al distinguirlos de contenidos auténticos (Diel et al., 2024). Estos hallazgos, sumados a la evidencia de que los detectores automáticos aún enfrentan graves problemas de generalización en entornos abiertos (Wang et al., 2024), explican por qué un video con imperfecciones técnicas pudo generar confusión masiva en un contexto bélico altamente polarizado.

2. Arquitectura tecnológica

El video fue creado utilizando tecnologías de síntesis audiovisual derivadas de redes generativas adversariales (GANs), entrenadas sobre material visual y sonoro real del presidente. Herramientas como DeepFaceLab, Avatarify o incluso entornos de código abierto como Faceswap fueron combinadas con sistemas de clonación de voz (text-to-speech con fine-tuning) basados en modelos como Tacotron 2 o Descript Overdub. La animación facial fue generada sobre un cuerpo actoral con parámetros semióticos similares, lo que facilitó el ensamblaje audiovisual sin recursos de Hollywood.

El resultado, aunque técnicamente imperfecto (ciertas inconsistencias labiales y rigidez en la mirada), fue suficiente para generar confusión en plataformas con alta circulación emocional como Telegram, TikTok y Facebook. Como sostiene Westerlund (2019), en contextos de crisis, el umbral de credibilidad se reduce drásticamente y lo que importa no es la precisión técnica, sino la autenticidad percibida. El afecto opera como criterio heurístico, desplazando el juicio crítico.

3. Actores involucrados

Investigaciones del Atlantic Council y la firma Graphika rastrearon la diseminación inicial a nodos digitales vinculados a campañas de influencia pro-Kremlin. Aunque no hubo una atribución directa a agencias estatales rusas, el patrón coincide con lo documentado en operaciones anteriores como "Secondary Infektion" y "Ghostwriter". El uso de cuentas automatizadas, proxys regionales y medios afines forma parte del repertorio híbrido descrito por Polyakova & Boyer (2018), donde lo estatal se disuelve en una red descentralizada de operadores semi-autónomos.

4. Circulación mediática y modulación algorítmica

La primera aparición del video ocurrió en canales de Telegram orientados a la audiencia rusoparlante. En cuestión de horas, fue replicado en Twitter y Facebook por cuentas que previamente habían amplificado narrativas desinformativas. Los algoritmos de recomendación de TikTok priorizaron el contenido debido a su carga emocional, y su formato vertical fue adaptado incluso por usuarios no alineados con la desinformación, en forma de dueto o reacción, amplificando así el alcance.

Las plataformas, si bien reaccionaron con cierta rapidez tras la verificación, no lograron frenar la curva de reproducción. Como señala Wardle (2020), el "momento óptimo de intervención" ya había pasado: el daño epistémico ya estaba hecho, incluso si el contenido se removía horas después. La arquitectura algorítmica de las plataformas —optimizada para engagement y no para veracidad— impide una respuesta eficaz en tiempo real.

5. Dilemas ético-epistémicos

Este caso inaugura una nueva fase del testimonio digital, donde la imagen deja de ser garantía de lo real y se convierte en campo de sospecha. Cuando la representación audiovisual puede ser fabricada con realismo plausible, la distinción entre documento y simulacro se vuelve difusa.

Los dilemas que emergen son múltiples:

• Epistémicos: ¿Puede una democracia sostenerse sin una imagen compartida de la realidad?

• Tecnopolíticos: ¿Qué actores controlan las tecnologías capaces de redefinir el visible?

• Éticos: ¿Debe la voz de una figura pública ser considerada patrimonio personal o bien colectivo manipulable?

El caso Zelensky, lejos de ser anecdótico, marca un punto de inflexión donde la inteligencia artificial se convierte en infraestructura bélica del régimen de percepción.

4.2. Caso B: AI4Media y la tecnogestión automatizada de la veracidad

1. Contexto sociopolítico

En 2020, ante el aumento exponencial de contenidos falsos en redes sociales durante la pandemia de COVID-19 y las elecciones globales, se lanzó AI4Media como proyecto europeo con el objetivo de introducir automatización responsable en el campo de la verificación. Promovido por universidades, ONGs y medios aliados, el sistema fue financiado por Horizon 2020 y plantea una solución híbrida al desborde informativo que enfrentan los equipos de fact-checking humano.

AI4Media nace así como una respuesta tecnopolítica a la "crisis de sobreinformación" (information overload), que hace inviable la evaluación manual de cada afirmación viral. Sin embargo, como advierte Marres (2020), la automatización del juicio epistémico plantea una tensión entre eficiencia y pluralidad, entre agilidad y deliberación.

2. Arquitectura tecnológica

El sistema se estructura en cuatro módulos interconectados:

• Extracción de afirmaciones: Basado en scraping de plataformas sociales y RSS, identificando frases factuales potencialmente virales.

• Clasificación semántica: Uso de transformers como BERT y RoBERTa para evaluar factualidad, polaridad, ambigüedad contextual.

• Recomendación de fuentes: Algoritmo que sugiere evidencias de medios tradicionales, bases de datos científicas o declaraciones oficiales.

• Interfaz de revisión humana: Plataforma donde los verificadores evalúan los análisis algorítmicos, pudiendo corregir, complementar o descartar.

Este modelo no reemplaza el juicio humano, pero sí lo preconfigura. Como advierten Pasquale (2020) y Gillespie (2018), todo algoritmo de sugerencia tiene una ontología implícita: define qué cuenta como evidencia, qué voces merecen atención y qué formas de conocimiento son legibles para la máquina.

3. Actores involucrados

Participan entidades académicas (Oxford, KU Leuven), organizaciones de verificación (Full Fact, Maldita), medios aliados (BBC, Deutsche Welle) y plataformas (Meta, Google). Aunque la gobernanza es colaborativa, el sistema ha sido criticado por reproducir sesgos institucionales: privilegia fuentes "oficiales", minimiza epistemologías alternativas (indígenas, populares, locales) y tiende a reforzar consensos dominantes.

La Comisión Europea, si bien supervisa el marco ético, ha sido ambigua sobre los criterios de transparencia y explicabilidad exigidos. En consecuencia, AI4Media representa tanto un avance como una amenaza: reduce la sobrecarga, pero puede consolidar un canon de "verdad algorítmica" excluyente.

4. Circulación y modulación algorítmica

AI4Media se ha integrado experimentalmente a sistemas de moderación de contenidos en Meta y Google, proporcionando etiquetas de advertencia, sugerencias de lectura y restricciones de viralidad. Durante la pandemia, un informe interno del medio *The Verge* estimó una reducción del 23% en la diseminación de falsedades sanitarias. Sin embargo, también se documentaron falsos positivos, especialmente en temas culturalmente sensibles como tratamientos tradicionales, creencias religiosas o cosmologías indígenas.

Estas limitaciones evidencian la necesidad de diseñar algoritmos no solo multilingües, sino interculturales, que no operen bajo el supuesto de una verdad universal y homogénea.

5. Dilemas ético-epistémicos

AI4Media encarna el dilema central del siglo XXI: ¿cómo articular tecnologías de automatización con regímenes democráticos del conocimiento? Los dilemas incluyen:

• Descontextualización semántica: El algoritmo interpreta frases sin captar su densidad cultural.

• Hegemonía epistémica: Lo "veraz" tiende a coincidir con lo institucional.

• Transparencia simulada: Aunque el sistema es auditado, sus criterios algorítmicos siguen siendo opacos para los ciudadanos.

El riesgo, en última instancia, no es solo que una IA decida lo que es verdad, sino que naturalicemos ese proceso sin deliberación pública.

4.3. Caso C: Imágenes falsas sobre Gaza 2023 y la IA generativa como catalizador de guerra emocional

1. Contexto sociopolítico

En octubre de 2023, tras el ataque de Hamas a territorio israelí y la posterior ofensiva militar de Israel sobre la Franja de Gaza, estalló una intensa guerra de imágenes en redes sociales. Lo novedoso fue que, por primera vez a gran escala, circularon imágenes fotorealistas generadas por inteligencia artificial que mostraban escenas de devastación, sufrimiento infantil, cadáveres y líderes en posturas dramáticas. Estas imágenes, producidas por usuarios anónimos, activistas o simpatizantes de uno u otro bando, eran difundidas en X (antes Twitter), TikTok, Instagram y Telegram, sin indicación explícita de su naturaleza sintética.

A diferencia de los casos clásicos de desinformación factual, aquí el propósito de la imagen no era afirmar una mentira concreta, sino intensificar una experiencia emocional colectiva. Como señala

Harari (2023), en contextos hipermediatizados de violencia, lo que circula no es tanto "la verdad" como una economía afectiva que busca adhesión visceral. La IA generativa se convierte entonces en una herramienta para moldear el dolor, la indignación o la empatía, más allá de la verificación.

Este fenómeno se inscribe en lo que Durnová (2023) denomina post-veracidad emocional: no importa tanto si algo es real, sino si se siente como auténtico. Las imágenes falsas sobre Gaza 2023 no apuntaban a sustituir hechos, sino a llenar los vacíos del testimonio visual allí donde no había periodistas, cámaras o medios.

Investigaciones recientes confirman que el etiquetado de contenidos como "generados por IA" produce efectos paradójicos: puede reducir la intención de compartir, pero también erosiona la credibilidad de materiales legítimos, lo que alimenta la desconfianza generalizada (Li et al., 2024). En paralelo, se advierte que la circulación de deepfakes en conflictos bélicos erosiona la confianza en plataformas y medios, incluso cuando se aplican mecanismos de detección automatizada (Lundberg & Mozelius, 2024).

2. Arquitectura tecnológica

Las imágenes fueron generadas con modelos de texto-a-imagen como Midjourney v5, DALL·E 3, Stable Diffusion XL y Leonardo.ai, herramientas cada vez más accesibles para usuarios no expertos. Estos modelos permiten producir fotografías realistas con indicaciones simples como "baby covered in dust crying amidst ruins in Gaza" o "Netanyahu kneeling beside dead children".

Además de su realismo visual, se emplearon estrategias estéticas avanzadas: juegos de luz y sombra para intensificar el dramatismo, uso de composición clásica (triángulos, simetrías) para evocar pathos, y manipulación cromática (colores apagados, piel grisácea) para simbolizar muerte o sufrimiento. Estas imágenes fueron reforzadas con subtítulos conmovedores, emojis, música lacrimógena o montaje en formato slideshow, convirtiéndose en dispositivos afectivos de gran impacto viral.

El nivel de detalle alcanzado (texturas de piel, lágrimas, miradas perdidas) desbordó las capacidades de discernimiento incluso de periodistas experimentados. Algunas imágenes incluían banderas, escrituras en árabe o hebreo, y escombros hiperrealistas generados por IA. Aunque algunas plataformas incluían marcas de agua o avisos, muchos usuarios recortaban las imágenes o las republicaban como "reflejos simbólicos" de una verdad mayor.

3. Actores involucrados

Este caso es paradigmático porque rompe con el binarismo de desinformación institucional vs. ciudadanía pasiva. La producción fue descentralizada y rizomática: activistas, simpatizantes, artistas digitales y cuentas anónimas participaron en la generación y difusión del contenido. Algunos justificaban su acción alegando que, ante la censura de imágenes reales por parte de plataformas o gobiernos, las creaciones con IA eran una forma legítima de "testimoniar simbólicamente" lo que no podía ser mostrado.

Otros usaban la IA como extensión de la imaginación militante, al estilo del "realismo afectivo" descrito por Berardi (2019), donde la ficción se convierte en catalizador de empatía. Esto plantea dilemas novedosos: ¿puede una imagen ser éticamente válida, aunque no sea factual? ¿Qué estatuto otorgamos al testimonio simulado cuando interpela emocionalmente a las audiencias?

Las plataformas, por su parte, se vieron desbordadas. Aunque Meta y X implementaron políticas contra imágenes sintéticas no declaradas, la velocidad de producción y la carga emocional hicieron que las medidas llegaran tarde. En muchos casos, la eliminación del contenido fue percibida como censura política, generando reacciones adversas y aumentando la polarización.

4. Circulación mediática

Las imágenes circularon principalmente bajo hashtags polarizantes como #GazaGenocide, #IsraelUnderAttack, #PrayForPalestine, y #StandWithIsrael. Fueron compartidas por influencers, cuentas de activismo político, medios alternativos e incluso ciudadanos comunes sin clara afiliación ideológica. Algunas imágenes alcanzaron millones de visualizaciones en menos de 24 horas.

TikTok desempeñó un papel central. Su algoritmo, diseñado para maximizar el tiempo de visualización y el engagement emocional, favoreció videos que combinaban imágenes de IA con música emotiva, ralentizaciones y efectos visuales. Esta coreografía algorítmica reforzaba la intensidad del mensaje, disolviendo la frontera entre realidad y ficción.

Como observan McKelvey y Dubois (2019), en las redes sociales contemporáneas la lógica de circulación no sigue criterios de veracidad, sino de resonancia afectiva. Esto explica por qué muchas de estas imágenes falsas superaron en alcance a los reportajes periodísticos tradicionales, desplazando el rol mediador de la prensa.

Además, algunos medios internacionales cayeron en la trampa de replicar o discutir imágenes sintéticas como si fueran reales, lo que generó controversias éticas y alimentó teorías de conspiración. En respuesta, surgieron iniciativas de verificación como AI or Not, HuggingFace Classifiers o plugins para detectar metadatos de imagen, aunque con eficacia limitada.

5. Dilemas ético-epistémicos

Este caso plantea dilemas de orden ontológico, no solo ético o informativo. Nos enfrentamos a una mutación profunda en el régimen de verdad visual:

• Colapso del índice fotográfico: Como advertían Barthes (1980) y Sontag (2003), la fotografía funcionaba como huella de lo real. La IA generativa rompe esa lógica al crear imágenes que no remiten a ningún referente externo, sino a una mezcla de patrones entrenados sobre millones de imágenes reales.

• Performance emocional de lo real: Las imágenes ya no buscan documentar hechos, sino activar emociones. Se convierten en performance política, en "máquinas de afectos" que dramatizan lo que "debería" ser verdad.

• Testimonio sintético: ¿Podemos hablar de ética del testimonio cuando lo que se muestra es una escena imaginada? ¿Tiene valor epistémico una imagen que no ocurrió, pero que logra representar el dolor colectivo?

• Normalización de la ficción como arma: Si el uso de IA generativa para manipular emociones se normaliza, entramos en una era donde cada conflicto puede estar mediado por una estética fabricada, una guerra de representaciones sin anclaje factual.

• Desplazamiento de la deliberación pública: En lugar de debatir sobre hechos contrastables, los públicos son arrastrados a campos emocionales opuestos, lo que dificulta la construcción de consensos mínimos para la acción política.

En suma, Gaza 2023 marca el ingreso pleno en la era post-fotográfica algorítmica, donde la imagen ya no refleja el mundo, sino lo anticipa, lo sustituye o lo intensifica. La IA generativa no es aquí solo una herramienta de desinformación, sino un vector de conmoción afectiva y modelación perceptiva.

5. Discusión

El análisis profundo y comparado de los tres casos —deepfake de Zelensky, sistema AI4Media , e imágenes falsas de Gaza 2023— revela que la relación entre inteligencia artificial y desinformación

no puede ser comprendida como una oposición binaria entre "uso malicioso" y "uso ético", ni resuelta mediante una apelación abstracta a "marcos normativos" o "mejores prácticas". En su lugar, emerge un entramado de tensiones más profundas que involucran no solo la función técnica de la IA, sino su rol como tecnología estructurante de la verdad, como dispositivo de poder simbólico y como campo de disputa epistémica y cultural. A continuación, se desarrollan seis líneas analíticas emergentes.

5.1. Del índice visual a la resonancia emocional: la mutación del régimen de veracidad

La imagen, en su tradición fotográfica, funcionó como índice: no decía, sino que mostraba. Era la huella de un acontecimiento. Desde Barthes (1980) hasta Sontag (1977), la fotografía fue considerada una forma de "presencia diferida", capaz de capturar lo real. Pero los tres casos aquí analizados —cada uno a su manera— evidencian que la IA está desmantelando esta arquitectura semiótica de lo verdadero.

En el caso de Zelensky, se simula el cuerpo presidencial con suficiente fidelidad como para desencadenar una respuesta emocional, a pesar de sus imperfecciones técnicas. En Gaza, la generación de imágenes que "representan el horror" sin referirse a ningún evento específico desplaza el índice por el *clímax afectivo*, lo cual puede considerarse un tránsito hacia una posfotografía sintética, donde lo que importa no es lo que la imagen remite, sino lo que hace sentir.

Este giro afecta no solo la estética, sino también la epistemología pública: ya no se exige que la imagen demuestre, sino que movilice. Esto conduce a una nueva forma de desinformación emocional, cuya eficacia se basa no en su plausibilidad factual, sino en su capacidad de confirmar afectos, marcos ideológicos y percepciones preexistentes.

Como advirtieron Goriunova (2019) y Scolari (2020), la desinformación ya no circula únicamente como datos, sino como "paquetes de afecto transmedia", preparados para enganchar en plataformas que premian la intensidad emocional y la conexión visceral. Este desplazamiento del régimen de veracidad implica que la crítica racional pierde eficacia si no se acompaña de una comprensión estética, emocional y cultural de los lenguajes de lo falso.

Además, esta mutación coincide con el concepto de "truthiness" planteado por Stephen Colbert y desarrollado en la literatura (Peters, 2021), donde la verdad ya no es lo que se puede comprobar, sino lo que se siente como cierto. La imagen emocional de IA se inscribe perfectamente en este giro hacia una verdad emocionalmente performativa, más que referencial.

5.2. Algoritmización de la autoridad epistémica: la verdad como producto de clasificación

En el segundo caso, AI4Media se presenta como un intento bienintencionado de sistematizar y escalar la verificación informativa mediante inteligencia artificial. Sin embargo, su análisis revela el surgimiento de un nuevo problema: la delegación algorítmica del juicio epistémico.

Los sistemas de IA no descubren verdades: ejecutan procesos de clasificación estadística entrenados con datos previamente etiquetados, de acuerdo con criterios definidos por actores con autoridad institucional, editorial o técnica. Este proceso implica una traducción computacional de la verdad, donde la ambigüedad, el contexto, la ironía o la narrativa compleja deben ser comprimidas en categorías binarias: verdadero / falso; confiable / engañoso.

El problema es que esta simplificación no es inocente ni neutral. Como han argumentado académicos como Boyd & Crawford (2012) y Noble (2018), los sistemas de IA no solo reflejan sesgos: los institucionalizan. Cuando se sistematiza una arquitectura de decisión algorítmica sobre lo que merece circular, sobre lo que debe ser etiquetado o reducido en visibilidad, se está configurando un régimen de poder epistémico automatizado, operado por empresas privadas pero con consecuencias públicas. De

hecho, evidencia reciente demuestra que la incorporación de verificaciones automatizadas en procesos de fact-checking puede reducir significativamente la percepción de veracidad de afirmaciones falsas, validando la necesidad de enfoques híbridos humano–máquina (DeVerna et al., 2024).

Esto genera una forma de gobernanza privatizada de la visibilidad cognitiva, en la que la veracidad ya no es producto del disenso deliberativo, sino de la eficiencia clasificatoria. La epistemología, en este contexto, se convierte en un subcampo de la arquitectura algorítmica, y el pluralismo informativo corre el riesgo de quedar atrapado en los filtros de lo verificable tecnológicamente.

5.3. Ambivalencia estructural, no contingente: la IA como tecnología bifronte

La IA aparece en los tres casos como una tecnología ambivalente. Pero esta ambivalencia no es superficial ni anecdótica: es ontológica. Es decir, la IA no puede ser completamente domesticada en términos morales o normativos, porque sus capacidades emergen de su apertura estructural: la misma lógica que permite detectar patrones puede ser utilizada para simularlos.

Esta condición de doble uso —ya ampliamente discutida en la literatura sobre biotecnología y armamento digital (Brundage et al., 2018)— toma en el caso de la IA una forma particular: el uso legítimo y el uso destructivo comparten la misma infraestructura, el mismo código base, el mismo principio de funcionamiento. No hay una "IA buena" y una "IA mala"; hay IA programada con diferentes objetivos, a menudo por actores con intereses geopolíticos o económicos contradictorios.

El mismo modelo de difusión que permite identificar rumores en redes sociales puede utilizarse para amplificarlos de manera más eficiente. La frontera entre solución y amenaza es, por tanto, frágil, y exige una gobernanza de la IA que no se limite a la funcionalidad técnica, sino que contemple también sus usos contextualizados, sus trayectorias institucionales y sus formas de apropiación cultural.

Por ello, la idea de una "IA ética" como solución a la desinformación resulta insuficiente si no se cuestionan también las condiciones estructurales de su producción, financiamiento y aplicación. Una IA entrenada para detectar falsedad puede ser reentrenada para fabricar simulacros más convincentes. Esta reversibilidad es clave para entender por qué la confianza técnica no es sinónimo de confianza pública.

En los tres casos, además, se observa cómo las soluciones técnicas se enfrentan a límites culturales, políticos y semánticos. No todo lo verificable es verdadero, y no todo lo falso es desinformación maliciosa. La reducción de la verdad a una clasificación técnica amenaza con excluir lo subjetivo, lo simbólico, lo minoritario y lo emergente del espacio de lo decible.

5.4. Crisis del testimonio y colapso del pacto de la palabra pública

El testimonio es uno de los pilares de la comunicación democrática. Fundado en la credibilidad de la voz que habla desde la experiencia, supone un acuerdo intersubjetivo: "te creo porque asumo que hablas con honestidad desde un lugar que reconozco." La IA, al permitir la creación de voces, rostros y declaraciones falsas que simulan ese testimonio, genera un colapso del pacto de la palabra pública.

El caso Zelensky no solo falsificó una declaración: intervino el cuerpo de la autoridad, vaciándolo de agencia, transformándolo en artefacto manipulable. Esto desestabiliza la posibilidad misma de lo testimonial, es decir, del acto de hablar con consecuencias. Si todo puede ser simulado, incluso el acto de hablar pierde valor epistémico y político.

Del otro lado, el uso de IA para construir imágenes "verdaderas" de Gaza sin base en eventos concretos pone en cuestión la noción misma de prueba. ¿Puede una imagen sintética ser "más verdadera" que una fotografía documental? ¿Quién decide? ¿Con qué criterios?

Esta ambigüedad produce lo que Dean (2010) llama "la inflación de la circulación": todo circula, todo se ve, pero nada puede ser verificado sin ambigüedad, y por lo tanto todo está sujeto a sospecha. La desinformación ya no se combate con correcciones, sino que se convierte en un entorno de saturación cognitiva, donde la duda se vuelve constante y la confianza se fragmenta.

5.5. Cartografía de la gobernanza algorítmica: de la plataforma a la infraestructura cognitiva

Los tres casos ponen en evidencia que las plataformas digitales han mutado en infraestructuras cognitivas. No se trata solo de espacios donde se intercambia información, sino de entornos que pre--configuran lo que se puede conocer, lo que se puede ver y lo que se puede decir.

Las decisiones sobre qué contenido etiquetar, qué afirmaciones priorizar, qué imágenes eliminar o amplificar, están en manos de plataformas como Meta, Google, TikTok. Estas empresas han asumido funciones propias del Estado —regulación de la verdad pública— pero sin controles democráticos, sin transparencia sustantiva y con lógicas comerciales internas que subordinan el bien común a la retención de usuarios.

El caso AI4Media ilustra cómo incluso las iniciativas académicas y cívicas terminan subsumidas en esta lógica, al integrarse a APIs, convenios de moderación o procesos de clasificación mediados por ingeniería de software. La arquitectura técnica deviene arquitectura epistémica. Y esta arquitectura, a su vez, opera bajo racionalidades neoliberales de eficiencia, escalabilidad y optimización que son estructuralmente incompatibles con la deliberación crítica, la ambigüedad semántica o la lentitud del juicio ético.

En este escenario, la regulación democrática de la IA y la desinformación requiere no solo marcos legales, sino una reconfiguración de las infraestructuras de visibilidad, de los criterios de relevancia algorítmica, de los sistemas de gobernanza de los datos y de los modelos de auditoría pública del conocimiento.

5.6. Hacia una justicia epistémica en tiempos de IA

En última instancia, lo que está en juego en el cruce entre IA y desinformación no es solo la calidad de la información, sino la distribución del poder cognitivo. Como sostienen D'Ignazio y Klein (2020), los datos no son neutrales: codifican historias, privilegios, exclusiones. De igual modo, los algoritmos no solo ordenan contenidos: ordenan el mundo.

El combate a la desinformación, entonces, no debe entenderse como una estrategia de "limpieza informativa" sino como una práctica de redistribución epistémica, donde se reconozcan los saberes situados, se protejan las voces minorizadas y se cuestionen las lógicas tecnocráticas que pretenden cerrar el sentido bajo el lenguaje de la eficiencia.

En este punto resulta clave considerar que la comprensión pública de la desinformación mediada por IA no depende únicamente de la precisión técnica de los modelos, sino de cómo los ciudadanos perciben la agencia y autonomía de esas tecnologías. Estudios recientes muestran que estas percepciones inciden directamente en los niveles de confianza hacia las instituciones epistémicas y en la legitimidad de los procesos de verificación (Shin, 2024).

Para ello, proponemos el concepto de justicia epistémica algorítmica, como horizonte regulador que articule:

- Derecho a la visibilidad no mediada por lógica de mercado;
- Derecho al disenso no etiquetado como desinformación;
- Reconocimiento de formas no hegemónicas de narrar, testimoniar y conocer;
- Diseño participativo de infraestructuras de validación;
- Formación crítica en alfabetización mediática y emocional, no solo técnica.

Este horizonte no es utópico, pero sí exige abandonar las ilusiones *tecnosolucionistas*. La desinformación no se resuelve con más IA, sino con más política, más crítica, más democracia del conocimiento.

6. Conclusiones

Este artículo ha analizado con profundidad crítica la relación entre inteligencia artificial y desinformación, identificando sus dilemas estructurales, su ambivalencia funcional y sus efectos epistémicos en tres casos paradigmáticos de la historia reciente. Lejos de concebir la desinformación como una simple anomalía corregible mediante soluciones tecnológicas, se ha argumentado que la conjunción entre IA y falsedad comunicativa reconfigura el régimen mismo de veracidad contemporáneo, afectando no solo la dimensión factual del conocimiento, sino también su plano simbólico, institucional, afectivo y algorítmico.

Los tres casos analizados —la falsificación audiovisual del presidente Zelensky, el sistema automatizado de verificación AI4Media y la proliferación de imágenes generadas por IA durante el conflicto en Gaza— han mostrado no solo formas distintas de acción de la IA, sino también lógicas divergentes de construcción, manipulación o control de la verdad. Al cruzarlos comparativamente, emergen cinco conclusiones centrales:

6.1. La IA no es un actor neutro, sino una infraestructura epistémica

La inteligencia artificial no debe ser concebida simplemente como una herramienta de procesamiento o una aplicación técnica, sino como una infraestructura epistémica que media activamente la producción, circulación y validación del conocimiento. Sus arquitecturas, entrenadas sobre corpus de datos históricos, institucionales o sesgados, configuran qué afirmaciones se consideran relevantes, qué narrativas se visibilizan y qué formas de decir son reconocidas como legítimas.

Esta condición estructural exige un giro conceptual: de la IA como "medio técnico" a la IA como dispositivo de construcción de lo verdadero. En esta lógica, cualquier intervención ética o regulatoria que no cuestione las condiciones materiales, institucionales y semióticas sobre las que se construyen estos sistemas será insuficiente. Las tecnologías de IA no son inocentes: representan una sedimentación algorítmica de visiones del mundo, intereses políticos y economías del conocimiento. Como tal, deben ser objeto de auditoría crítica permanente, más allá de los discursos normativos centrados exclusivamente en la eficiencia o la innovación.

6.2. La desinformación ya no es únicamente una cuestión factual, sino afectiva y performativa

Uno de los hallazgos más relevantes del estudio es el desplazamiento del eje de la desinformación: de un modelo basado en la falsedad objetiva hacia un modelo basado en la eficacia emocional y la resonancia afectiva. Las tecnologías generativas no solo alteran los contenidos, sino que optimizan su diseño para maximizar la conmoción, la viralidad o la empatía inducida. En este nuevo régimen semiótico, la "verdad" se convierte en una competencia entre afectos, y no solo entre argumentos o datos verificables.

Esta mutación tiene implicaciones críticas. Por un lado, desborda las capacidades tradicionales de verificación, que operan bajo una lógica racionalista. Por otro, vulnera emocionalmente a los públicos, sometidos a una sobrecarga sensorial e identitaria que dificulta la deliberación crítica. La IA no solo produce datos falsos, sino experiencias emocionales manipuladas. Y en un entorno de plataformas que premian la intensidad afectiva, los contenidos engañosos diseñados para movilizar indignación, compasión o temor tienen mayor probabilidad de éxito.

En este escenario, se vuelve urgente complementar las estrategias de verificación técnica con pedagogías críticas de las emociones mediadas, que permitan a las audiencias reconocer cuándo sus reacciones afectivas están siendo instrumentalizadas por arquitecturas algorítmicas invisibles.

6.3. La automatización del juicio epistémico conlleva riesgos de simplificación, exclusión y captura ideológica

El caso de AI4Media mostró las oportunidades, pero también los límites y riesgos, de los sistemas de verificación basados en IA. La promesa de eficiencia —reducir el tiempo de verificación, ampliar la escala— entra en tensión con la reducción semántica que estos sistemas implican. La clasificación automática exige binarismos (verdadero/falso, confiable/engañoso) que excluyen matices culturales, contextos locales o lenguajes metafóricos.

Más preocupante aún es la captura ideológica del juicio epistémico. Cuando la veracidad se define por algoritmos entrenados con fuentes mainstream, gubernamentales o institucionales, se corre el riesgo de que las voces alternativas, los saberes minorizados o las narrativas de resistencia sean etiquetadas como dudosas o desinformativas. Esta automatización del criterio no es solo un problema técnico, sino una amenaza a la pluralidad epistémica y a la deliberación democrática.

En consecuencia, urge una discusión más profunda sobre quién entrena los algoritmos, con qué criterios, con qué corpus, bajo qué marcos regulatorios y con qué posibilidad de escrutinio público.

6.4. La gobernanza de la desinformación no puede seguir delegada a actores privados sin control público

Otro eje transversal de los tres casos es el rol dominante de las plataformas tecnológicas en la gestión de la desinformación. Ya sea mediante eliminación de contenidos, etiquetado automático o manipulación de visibilidad, las empresas privadas han asumido funciones propias de la esfera pública —como la moderación de la verdad, la priorización del discurso o la sanción del error— sin mecanismos sólidos de rendición de cuentas.

Este modelo privatizado de gobernanza cognitiva no es compatible con una sociedad democrática. Las decisiones sobre qué se muestra, qué se silencia o qué se corrige deben estar sujetas a principios de transparencia, deliberación pública, justicia informativa y control social. De lo contrario, se consolida una tecnocracia opaca que convierte a las plataformas en árbitros de la realidad, con intereses comerciales en conflicto con el bien común.

Urge una reconfiguración de la arquitectura institucional del ecosistema digital, con normas internacionales, marcos de justicia epistémica, mecanismos de auditoría independientes y participación multisectorial.

6.5. Urge una agenda de justicia epistémica algorítmica como marco transformador

Finalmente, este artículo propone que la respuesta a la desinformación mediada por IA no debe limitarse a estrategias de contención o corrección, sino que debe inscribirse en un horizonte más ambicioso: el de la justicia epistémica algorítmica. Este horizonte parte de la constatación de que las regulaciones actuales, aunque avanzadas en sus principios, resultan aún insuficientes frente a la complejidad del fenómeno. Por ejemplo, como advierte Labuz (2024), el propio AI Act europeo presenta limitaciones significativas a la hora de abordar usos políticos de los *deepfakes*, quedándose en un plano más normativo que operativo. De manera convergente, Romanishyn et al. (2025) subrayan que la **resiliencia democrática** requiere no solo regulación, sino también políticas públicas activas, cooperación multiactor y estándares internacionales capaces de anticipar escenarios emergentes de manipulación informativa.

• Desde esta perspectiva, una agenda de justicia epistémica algorítmica implica:
• Democratizar el diseño, entrenamiento y supervisión de los sistemas de IA;
• Reconocer los saberes situados y proteger la diversidad narrativa como valor epistémico;
• Garantizar el derecho al disenso y al error sin estigmatización algorítmica;
• Promover una alfabetización crítica integral (mediática, emocional, algorítmica);
• Establecer auditorías públicas y participativas que velen por la equidad, la transparencia y el impacto social de los modelos de verificación y moderación;
• Desmercantilizar la visibilidad, separando el derecho a la voz pública de los incentivos de monetización o viralidad.

En suma, la justicia epistémica algorítmica se plantea como un marco transformador que articula regulación, pedagogía y participación social. La IA, como toda tecnología de poder, debe ser disputada políticamente, intervenida colectivamente y sometida a marcos de equidad cognitiva. Solo así podremos construir un ecosistema informativo que no reproduzca desigualdades, silencie disensos o administre la verdad desde centros de cálculo automatizados.

Líneas futuras de investigación

A partir de este estudio, se proponen varias líneas investigativas estratégicas que pueden profundizar la comprensión de esta problemática y generar insumos para políticas públicas transformadoras:

1. Cartografía comparada de arquitecturas algorítmicas de moderación en distintas plataformas y regiones;
2. Etnografía digital de la recepción de contenidos generados por IA en públicos hiperpolarizados;
3. Estudios de caso sobre resistencia ciudadana a la desinformación algorítmica en contextos no occidentales;
4. Investigación sobre IA y memoria: efectos de los contenidos sintéticos en procesos de construcción de verdad, justicia y reparación;
5. Modelos participativos de verificación distribuida basados en inteligencia colectiva y transparencia radical.

Este trabajo no pretende cerrar el debate, sino ampliarlo, complejizarlo y radicalizar su dimensión ética. Frente a la expansión de sistemas capaces de simular lo humano y manipular la realidad perceptiva, lo que está en juego no es solo la calidad de la información, sino la posibilidad misma de construir verdad compartida en contextos democráticos. La IA ha venido a redefinir la disputa por el conocimiento. La responsabilidad de responder críticamente a este desafío es ineludible y urgente.

Agradecimientos / Financiación

Estos resultados forman parte de las actividades de investigación promovidas a través de:

El proyecto PID2021-124293OB-I00, del *Programa Estatal de I+D+i para la realización de proyectos de Generación de Conocimiento 2021*, titulado *"Mapa de la Desinformación en las Comunidades Autónomas y Entidades Locales de España y su Ecosistema Digital (FAKELOCAL)"*.

El proyecto interno de la Universidad de Bogotá Jorge Tadeo Lozano, titulado *"Análisis del impacto de la desinformación en podcasts a nivel global, con énfasis en el caso colombiano: estrategias de combate y recomendaciones para su implementación"* (Código de Proyecto: 2403-25-2024-2024).

7. Referencias Bibliográficas

Altay, S., Dohle, M., & De Keersmaecker, J. (2024). People are skeptical of headlines labeled as AI-generated—even if true or human-made—because they assume full AI automation. *PNAS Nexus, 3(*10), pgae403. https://doi.org/10.1093/pnasnexus/pgae403

Balafrej, I., Dahmane, M. Enhancing practicality and efficiency of deepfake detection. *Scientific Reports* 14, 31084 (2024). https://doi.org/10.1038/s41598-024-82223-y

Barthes, R. (1980). *La cámara lúcida: Nota sobre la fotografía*. Paidós.

Bennett, W. L., & Livingston, S. (2018). The disinformation order: Disruptive communication and the decline of democratic institutions. *European Journal of Communication, 33*(2), 122–139. https://doi.org/10.1177/0267323118760317

Berardi, F. (2019). *Futurability: The Age of Impotence and the Horizon of Possibility*. Verso.

Boddington, P. (2017). *Towards a Code of Ethics for Artificial Intelligence*. Springer.

Bourdieu, P., & Wacquant, L. (1992). *An Invitation to Reflexive Sociology*. University of Chicago Press.

Boyd, D., & Crawford, K. (2012). Critical questions for big data. *Information, Communication & Society, 15*(5), 662–679. https://doi.org/10.1080/1369118X.2012.678878

Boyd, D. (2019). *It's Complicated: The Social Lives of Networked Teens* (Updated ed.). Yale University Press.

Braun, V., & Clarke, V. (2006). Using thematic analysis in psychology. *Qualitative Research in Psychology, 3*(2), 77–101. https://doi.org/10.1191/1478088706qp063oa

Brundage, M., Avin, S., Clark, J., Toner, H., Eckersley, P., Garfinkel, B., Dafoe, A., Scharre, P., Zeitzoff, T., Filar, B., Anderson, H., Roff, H., Allen, G. C., Steinhardt, J., Flynn, C., Ó hÉigeartaigh, S., Beard, S. J., Belfield, H., … Amodei, D. (2018). The malicious use of artificial intelligence: Forecasting, prevention, and mitigation. *Center for a New American Security*. https://arxiv.org/abs/1802.07228

Couldry, N., & Mejias, U. (2019). *The Costs of Connection: How data is colonizing human life and appropriating it for capitalism*. Stanford University Press.

Crawford, K. (2021). *Atlas of AI: Power, politics, and the planetary costs of artificial intelligence*. Yale University Press.

D'Ignazio, C., & Klein, L. F. (2020). *Data feminism*. MIT Press.

Dean, J. (2010). *Blog Theory: Feedback and Capture in the Circuits of Drive*. Polity Press.

Denzin, N. K., & Lincoln, Y. S. (Eds.). (2018). *The SAGE handbook of qualitative research* (5th ed.). SAGE Publications.

DeVerna, M. R., H.Y. Yan, K. Yang, & F. Menczer. (2024). Fact-checking information from large language models can reduce the perceived accuracy of false claims. *Proceedings of the National Academy of Sciences, 121*(24), e2322823121. https://doi.org/10.1073/pnas.2322823121

Diel, A., Lalgi, T., Schröter, I. C., MacDorman, K. F., Teufel, M., & Bäuerle, A. (2024). Human performance in detecting deepfakes: A systematic review and meta-analysis of 56 papers. *Computers in Human Behavior Reports, 16*, 100538. https://doi.org/10.1016/j.chbr.2024.100538

Dubois, E., & McKelvey, F. (2019). Political bots: Disrupting Canada's democracy. *Canadian Journal of Communication, 44*(2), 27–33. https://doi.org/10.22230/cjc.2019v44n2a3511

Durnová, A., & Karell, D. (2023). Emotions and the "truths" of contentious politics: Advances in research on emotions, knowledge and contemporary contentious politics. *Emotions & Society, 5*(3), 252–256. https://doi.org/10.1332/26316897Y2023D000000004

Fairclough, N. (2001). *Language and Power* (2nd ed.). Longman.

Floridi, L. (2020). *The Green and the Blue: A New Political Ontology for a Mature Information Society*. Polity Press.

Flyvbjerg, B. (2006). Five misunderstandings about case-study research. *Qualitative Inquiry, 12*(2), 219–245. https://doi.org/10.1177/1077800405284363

Fricker, M. (2007). *Epistemic Injustice: Power and the Ethics of Knowing*. Oxford University Press.

Fuchs, C. (2017). *Social Media: A Critical Introduction* (2nd ed.). SAGE.

Gelfert, A. (2018). Fake news: A definition. *Informal Logic, 38*(1), 84–117. https://doi.org/10.22329/il.v38i1.5068

Gerdon, F., et al. (2022). Social impacts of algorithmic decision-making: A research agenda. *Big Data & Society*. https://doi.org/10.1177/20539517221089305

Gillespie, T. (2018). *Custodians of the Internet: Platforms, Content Moderation, and the Hidden Decisions that Shape Social Media*. Yale University Press.

Glaser, B. G., & Strauss, A. L. (1967). *The Discovery of Grounded Theory: Strategies for Qualitative Research*. Aldine.

Goriunova, O. (2019)**.** *The Digital Subject: People as Data as Persons. Theory, Culture & Society*, 36(6), 125–145. https://doi.org/10.1177/0263276419840409

Graves, L. (2018). Boundaries not drawn: Mapping the institutional roots of the global fact-checking movement. *Journalism Studies, 19*(5), 613–631. https://doi.org/10.1080/1461670X.2016.1196602

Harari, Y. N. (2023). *Unstoppable Us, Volume 2: Why the World Isn't Fair*. HarperCollins.

Heidari, A., Jafari Navimipour, N., Dağ, H., & Ünal, M. (2024). Deepfake detection using deep learning methods: A systematic and comprehensive review. *WIREs Data Mining and Knowledge Discovery, 14*(2), e1520. https://doi.org/10.1002/widm.1520

Jack, C. (2017). Lexicon of lies: Terms for problematic information. Data & Society Research Institute. https://datasociety.net/library/lexicon-of-lies/

Labuz, K. (2024). Deepfakes in politics: Why the EU Artificial Intelligence Act falls short and how to improve it. *Policy & Internet, 16*(2), 406–424. https://doi.org/10.1002/poi3.406

Leonelli, S. (2016). *Data-Centric Biology: A Philosophical Study*. University of Chicago Press.

Li, F., et al. (2024). Impact of Artificial Intelligence–Generated Content labels on perceived accuracy and sharing intention for misinformation. *JMIR Formative Research, 8*, e60024. https://doi.org/10.2196/60024

Lundberg, E., & Mozelius, P. (2024). The potential effects of deepfakes on news media and entertainment. *AI & Society, 40* (Suppl.), 1–15. https://doi.org/10.1007/s00146-024-02072-1

Madaio, M. A., Stark, L., Wortman Vaughan, J., & Wallach, H. (2020). Co-designing checklists to understand organizational challenges and opportunities around fairness in AI. *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, 1–14. https://doi.org/10.1145/3313831.3376445

Marres, N. (2020). For a situational analytics: Collective responses to climate change on Twitter. *Big Data & Society, 7*(2). https://doi.org/10.1177/2053951720949571

Marwick, A., & Lewis, R. (2017). *Media Manipulation and Disinformation Online*. Data & Society. https://datasociety.net/pubs/oh/DataAndSociety_MediaManipulationAndDisinformationOnline.pdf

McIntyre, L. (2018). *Post-Truth*. MIT Press.

Medina, J. (2012). *The Epistemology of Resistance: Gender and Racial Oppression, Epistemic Injustice, and the Social Imagination*. Oxford University Press.

Mittelstadt, B. D., Allo, P., Taddeo, M., Wachter, S., & Floridi, L. (2016). The ethics of algorithms: Mapping the debate. *Big Data & Society, 3*(2). https://doi.org/10.1177/2053951716679679

Mosco, V. (2009). *The Political Economy of Communication* (2nd ed.). SAGE.

Napoli, P. M. (2019). *Social Media and the Public Interest: Media Regulation in the Disinformation Age*. Columbia University Press.

Noble, S. U. (2018). *Algorithms of Oppression: How search engines reinforce racism*. NYU Press.

O'Neil, C. (2016). *Weapons of math destruction: How big data increases inequality and threatens democracy*. Crown Publishing.

Pariser, E. (2011). *The filter bubble: What the internet is hiding from you*. Penguin Press.

Pasquale, F. (2020). *New Laws of Robotics: Defending Human Expertise in the Age of AI*. Harvard University Press.

Pérez-Rosas, V., Kleinberg, B., Lefevre, A., & Mihalcea, R. (2018). Automatic detection of fake news. *Proceedings of the 27th International Conference on Computational Linguistics*, 3391–3401. https://aclanthology.org/C18-1287

Peters, J. D. (2012). *The marvelous clouds: Toward a philosophy of elemental media*. University of Chicago Press.

Polyakova, A., & Boyer, S. P. (2018). *The future of political warfare: Russia, the West, and the coming age of global digital competition*. Brookings Institution. https://www.brookings.edu/research/the-future-of-political-warfare/

Pomerantsev, P. (2019). *This is not propaganda: Adventures in the war against reality*. PublicAffairs.

Rid, T. (2020). *Active measures: The secret history of disinformation and political warfare*. Farrar, Straus and Giroux.

Romanishyn A, Malytska O, & Goncharuk V. (2025). AI-driven disinformation: policy recommendations for democratic resilience. *Front Artif Intell*. https://doi.org/10.3389/frai.2025.1569115

Scolari, C. A. (2020). *La guerra de las plataformas: Del papiro al metaverso*. Editorial Anagrama.

Shin, D. (2024). How people understand misinformation from generative AI. *New Media & Society*. https://doi.org/10.1177/14614448241234040

Sontag, S. (1977). *On Photography*. Farrar, Straus and Giroux.

Stake, R. E. (2005). *The art of case study research*. SAGE.

Sultan, M., Tump, A. N., Ehmann, N., Benkler, Y., Speekenbrink, M., & Hertwig, R. (2024). Susceptibility to online misinformation: A systematic meta-analysis of demographic and psychological factors. *Proceedings of the National Academy of Sciences, 121*(9), e2409329121. https://doi.org/10.1073/pnas.2409329121

Vaccari, C., & Chadwick, A. (2020). Deepfakes and disinformation: Exploring the impact of synthetic political video on deception, uncertainty, and trust in news. *Social Media + Society, 6*(1). https://doi.org/10.1177/2056305120903408

Vosoughi, S., Roy, D., & Aral, S. (2018). The spread of true and false news online. *Science, 359*(6380), 1146–1151. https://doi.org/10.1126/science.aap9559

Wang, T., Liao, X., Chow, K.-P., Lin, X., & Wang, Y. (2024). Deepfake detection: A comprehensive survey from the reliability perspective. *ACM Computing Surveys*. https://doi.org/10.1145/3699710

Wardle, C., & Derakhshan, H. (2017). *Information Disorder: Toward an interdisciplinary framework for research and policy making*. Council of Europe report.

Wardle, C. (2020). *Understanding Information Disorder*. First Draft. https://firstdraftnews.org/articles/understanding-information-disorder

Westerlund, M. (2019). The emergence of deepfake technology: A review. *Technology Innovation Management Review, 9*(11), 40–53. https://doi.org/10.22215/timreview/1282

Wittenberg, C., et al. (2025). Labeling AI-generated media online. *PNAS Nexus, 4*(6), pgaf170. https://doi.org/10.1093/pnasnexus/pgaf170

Wodak, R., & Meyer, M. (2009). *Methods of Critical Discourse Analysis* (2nd ed.). SAGE.

Yin, R. K. (2018). *Case study research and applications: Design and methods* (6th ed.). SAGE.

Zuboff, S. (2019). *The age of surveillance capitalism: The fight for a human future at the new frontier of power*. PublicAffairs.