DESINFORMACIÓN
Y FACT-CHECKING
en la era de la IA

CIESPAL | EDICIONES CIESPAL

Periodística Y Nuevas Culturas Informativas 8

Editor
David García-Marín

Desinformación y *fact-checking*
en la era de la IA

Desinformación y *fact-checking* en la era de la IA

CIESPAL | EDICIONES CIESPAL

2025

Desinformación y fact-checking en la era de la IA

Colección Periodística y Nuevas Culturas Informativas No. 8

Editor
David García-Marín

Autores
Manuel Álvarez Rufs
Roberto Aparici
Fernando Bordignon
David García-Marín
Leonardo Murolo

ISBN Digital: 978-9978-55-237-7

Edición General
Gissela Dávila Cobo
Gestión editorial
Diego S. Acevedo A.

CIESPAL
Centro Internacional de Estudios Superiores de Comunicación para América Latina
Av. Diego de Almagro N32-133 y Andrade Marín • Quito, Ecuador
Teléfonos: (593 2) 254 8011
www.ciespal.org
https://ediciones.ciespal.org/

Ediciones Ciespal, 2025

Índice

Prólogo

El 18 de mayo de 2025 el secretario general de las Naciones Unidas convocó a un acto multitudinario a todos los propietarios de los medios de comunicación, redes sociales y empresas tecnológicas, además de sus representantes y periodistas especializados. Este evento mundial iba a tener lugar en el estadio deportivo más grande que existe hasta el momento, el estadio Narendra Modi, construido en la ciudad de Ahmedabad, en la India.

Era un acto único desde que se redactara el informe MacBride, igualmente conocido como *Voces múltiples, un solo mundo.* Este informe fue publicado por la UNESCO en 1980 en el marco de una investigación desarrollada por la Comisión Internacional para el Estudio de los Problemas de la Comunicación, presidida en aquel entonces por Sean MacBride. El objetivo principal de esa investigación era "estudiar la totalidad de los problemas de la comunicación en las sociedades modernas".

En este actual evento, lo que se iba a discutir es la necesidad de actuar desde todos los medios de comunicación con la mayor velocidad posible, y con el objetivo principal de que los medios y las redes sociales informen de manera veraz y contundente. Es decir, que los medios y las redes sociales sirvan para la información y para el entretenimiento, entre otras cosas, pero no para desinformar.

El secretario general llegó a la hora indicada para dar comienzo a este encuentro ante la expectación de todas las personas presentes en el estadio Narendra Modi, provisto de un aforo de 132.000 espectadores totalmente completo, obviando pasillos, puestos de restauración, salas de prensa y oficinas de trabajo. Se puede decir que no cabía un alfiler. Se escuchaba un ligero murmullo hasta que la voz del secretario general de las Naciones Unidas comenzó a resonar amplificada por los altavoces del estadio:

—Señoras y señores representantes de los medios de comunicación, empresas tecnológicas y de las redes sociales de todo el mundo, nos hemos reunido aquí para debatir un tema crucial para la democracia. Ese tema crucial para la democracia tiene que ver con la información. Sin información veraz no se pueden producir o construir estados democráticos. Tenéis en cada uno de vuestros asientos un dispositivo donde se indican algunas cuestiones, y a mis preguntas tenéis que apretar alguno de esos botones. Después de las conferencias y las referencias que hemos hecho al importante aporte que supuso para la comunicación el informe MacBride, nos toca ahora dilucidar cuál es la perspectiva que tienen los medios de comunicación y las redes sociales en este siglo XXI y con el apogeo de la inteligencia artificial. Bueno, la primera cuestión que yo como secretario general de Naciones Unidas quiero preguntaros es lo siguiente. ¿Estáis dispuestas, estáis dispuestos, a decir en vuestros medios y redes la verdad, toda la verdad y nada más que la verdad? Tenéis que marcar sí o no en vuestro dispositivo.

De forma unánime, la respuesta fue contundente. Se alzó una voz en el silencio de todo ese acto.

—Eh, sí, pero tenemos una observación que hacer. El deseo nuestro, yo, como representante de todos los propietarios de los medios. Yo, y todos mis compañeros y compañeras que estamos compartiendo en este estadio, deseamos la veracidad, pero también queremos ganar dinero. Y el dinero no se gana con la verdad o la veracidad. El dinero se gana desinformando, el dinero se gana con verdades a medias. Es triste reconocerlo, pero el informe *La difusión de noticias verdaderas y falsas en línea* que realizó el Instituto Tecnológico de Massachusetts, MIT, a finales de la segunda década del siglo

XXI[1], ya informaba que las audiencias, los públicos se interesaban mucho más por los actos no verdaderos, los actos no reales, que por la veracidad. Es un triunfo de las audiencias que amen las mentiras, las mentiras que además creamos y fomentamos nosotros desde los grandes medios.

Se hizo un silencio total, un silencio atronador y una calma primigenia que resonaba en el aire. Cualquier pequeño ruido se multiplicaba por cien, por mil, por un millón. Cualquier pequeño sonido se multiplicaba al infinito en el estadio más grande del mundo. El secretario general de las Naciones Unidas, asombrado ante esta respuesta, tomó aire, respirando profundamente, y empezó a hacer una propuesta.

—Señoras y señores, propietarios y propietarias de los medios, representantes de los medios de comunicación, periodistas especializados en medios de comunicación, redes sociales y tecnología, os traemos aquí una propuesta. Esta obra a la que podéis acceder en vuestros dispositivos, este manual de desinformación puede serviros a cada una de vosotras y a cada uno de vosotros para empezar a trabajar sobre la colonización en los medios, y a su vez, sobre los procesos de alfabetización en los medios de comunicación y en las redes sociales, estableciendo los fundamentos de la desinformación, exponiendo su relación con influencers y procesos tecnopolíticos, descubriendo las nuevas fábricas algorítmicas de la desinformación, navegando en un universo deepfake, aplicando técnicas de verificación digital y fact-checking, para construir una alfabetización mediática útil para combatir la desinformación.

En esta obra coordinada por el profesor David García Marín, y con la colaboración de los profesores, Manuel Álvarez Rufs, Roberto Aparici, Fernando Bordignon y Leonardo Murolo, encontraréis un manual, un manual único que además de dar un enfoque teórico también ofrece perspectivas prácticas. Cada uno de vosotros, cada una de vosotras, tenéis en estos momentos un ejemplar. Espero que ese ejemplar se distribuya gratuitamente en todo el mundo y que pueda ser una vía de alfabetización, una primera lectura, una primera profundización a nivel colectivo, a nivel global, sobre la desinformación. Este manual, *Desinformación y fact-*

1 Vosoughi, S., Roy, D., & Aral, S. (2018). The spread of true and false news online. *Science*, *359*(6380), 1146-1151.

checking en la era de la IA, nos habla desde diferentes perspectivas acerca de las temáticas que comprometen a las democracias de hoy en día.

Bueno, hacemos ahora una interrupción, y seguimos luego hablando sobre cómo poner en práctica estas cuestiones que planteamos y que se plantean en este manual de desinformación. Os sugiero que, en este pequeño intermedio, leáis la introducción y veáis qué aportes o qué dimensión o dimensiones pueden serviros para comenzar a discutir. Así que abrid el libro, y vamos a iniciar una lectura individual y colectiva de esta obra que está producida por el Centro Internacional de Estudios Superiores de Comunicación para América Latina, CIESPAL, el organismo de comunicación por excelencia de Latinoamérica.

Aquí se abre un nuevo periodo en nuestra formación y en nuestras vidas. Ojalá profundice y sea de importancia vital como lo ha sido para mí a la hora de hacer esta convocatoria única para transformar la comunicación y la información a lo que debe ser. Muchas gracias, señoras y señores[2].

Roberto Aparici.
Coordinador de la Cátedra de Educomunicación
Mario Kaplún "Hoy es mañana"
CIESPAL

2 Este es un acto ficticio. Es un acto que no fue real, pero todos los datos que se dan en este prólogo están basados en informaciones verdaderas, auténticas, como es el caso del estadio más grande del mundo y también de la necesidad de una convocatoria mundial acerca de la desinformación y sobre cómo construir la veracidad en estos tiempos complejos y oscuros.

Introducción

David García-Marín
Universidad Rey Juan Carlos
España

La noticia estaba en todos los medios. En las redes. En las tertulias televisivas. En las conversaciones de bar. En los grupos de WhatsApp. Las terribles inundaciones provocadas por una dana (fenómeno meteorológico que ocasiona precipitaciones muy intensas) en Valencia (España) en otoño de 2024 habían llenado de agua y barro el parking del centro comercial de Bonaire, situado en la localidad de Aldaia. Se trata de un aparcamiento subterráneo con varias plantas de 1200 metros cuadrados cada una, donde supuestamente se encontraban sepultadas bajo el agua y el lodo unas 700 personas. La lógica parece aplastante: en un parking subterráneo tan grande, a la hora de la riada, lo normal es que hubiera centenares de personas que no tuvieron escapatoria y fallecieron ahogadas.

No fue así. Días después, cuando los equipos de rescate entraron al parking y descubrieron que no había ningún fallecido en ninguna de las plantas, la noticia fue toda una sorpresa. Toda la sociedad española, de forma acrítica, había dado por hecho algo que, aunque parece lógico, no lo es. En realidad, la lógica opera al contrario: si había algún lugar donde era fácil no encontrar fallecidos, precisamente ese era un

parking como el de Bonaire. La razón es pura física: en una superficie tan grande, el agua se reparte y tarda más tiempo en subir lo suficiente como para que una persona de un tamaño normal se ahogue. Los (pocos o muchos) usuarios del parking que estaban allí en el momento de la inundación tuvieron tiempo suficiente para escapar. En un plazo no superior a 20 minutos, es posible desalojar a un millar de personas en una infraestructura como esa. Ante una riada, es más difícil sobrevivir en una habitación de 12 metros cuadrados que en una superficie de 1200. A nadie se le ocurrió y toda la sociedad española creyó fielmente las alarmistas noticias que vaticinaban la masacre.

Este ejemplo nos muestra derivadas interesantes sobre cómo operan algunos de los mecanismos de las noticias falsas en el mundo contemporáneo, donde las campañas desinformativas se configuran como entramados complejos con la intervención de una panoplia de agentes: instrumentos tecnológicos con mayor o menor grado de sofisticación, actores políticos, *influencers*, pseudomedios, plataformas digitales, medios de comunicación hiperpartisanos, agendas económicas, políticas e ideológicas y, finalmente, el procesamiento cognitivo de la información de cada uno de los usuarios.

En este contexto, el libro *Desinformación y fact-checking en la era de la IA* constituye un aporte fundamental al estudio crítico de la desinformación en el mundo postdigital. La obra pretende abordar con profundidad y rigor académico las múltiples dimensiones que configuran este fenómeno complejo, transversal y en constante mutación. A lo largo de sus capítulos, se despliega una cartografía de los mecanismos, actores, tecnologías y narrativas que intervienen en la producción, circulación y recepción de contenidos desinformativos, así como algunas de las estrategias de resistencia y alfabetización necesarias para contrarrestarlos.

Uno de los principales méritos de la obra radica en su enfoque interdisciplinario, que articula saberes provenientes de la comunicación, la filosofía, la sociología, la ciencia política, la informática y la pedagogía crítica. Esta perspectiva permite comprender la desinformación no

como un fenómeno aislado o meramente técnico, sino como una manifestación sintomática de las transformaciones estructurales que atraviesan nuestras sociedades contemporáneas: la crisis de los sistemas de mediación tradicionales, la erosión de la confianza institucional, la polarización ideológica, la precarización del periodismo, el auge de los movimientos neofascistas, la automatización de la comunicación y la emergencia de nuevas formas de subjetivación algorítmica.

La obra está configurada a partir de seis capítulos, con voluntad claramente pedagógica, lenguaje claro y accesible para cualquier lector, independientemente de su grado de especialización en la materia. Se configura, en realidad, como un manual para una iniciación básica al fenómeno de la desinformación en su configuración actual. Una obra para entender el final de la tercera década del siglo XXI y anticipar futuros posibles.

En el primer capítulo, de carácter introductorio, los profesores Roberto Aparici, Manuel Álvarez Rufs y Fernando Bordignon, abordan los siempre controvertidos conceptos de desinformación, posverdad y *fake news*, poniendo de manifiesto la complejidad del fenómeno a través de las múltiples manifestaciones que adquiere el contenido desinformativo en la sociedad actual. Analizan los precedentes de la desinformación que nos circunda en nuestros días, demostrando que éste no es un fenómeno nuevo, sino que se ha reconfigurado en la última década; no solo por cuestiones tecnológicas, sino también por la acción de nuestros sesgos cognitivos, de forma que la posverdad representa el triunfo de la forma en que nuestro cerebro procesa la información frente al razonamiento lógico y la realidad de los hechos. En este sentido, el recorrido conceptual que inaugura el libro ofrece un marco teórico sólido para comprender cómo las emociones, las creencias y los atajos mentales condicionan la forma en que las personas perciben la realidad. En este sentido, destaca la importancia de nociones como el *hedonismo cognitivo*, el *sesgo de confirmación*, el *efecto de contrafuego* o la *disonancia cognitiva*, que explican por qué las falsedades persisten incluso frente a evidencias empíricas que las refutan. Esta dimensión

psicológica y cultural de la desinformación es clave para entender su eficacia y su capacidad de penetración en el tejido social.

El profesor Leonardo Murolo, en el segundo capítulo, reflexiona sobre el actual escenario de los activismos y militancias políticas en el ámbito digital. Ahonda en la acción de la tecnopolítica como el tablero donde se dirime la batalla ideológica, vinculando este fenómeno a los movimientos neofascistas globales. Este capítulo ofrece algunas claves sobre cómo estos movimientos se sirven de las dinámicas propias de las redes sociales para diseñar campañas desinformativas a fin de favorecer sus posicionamientos ideológicos. En este escenario, los *influencers* se erigen como actores clave al promover de forma más o menos subliminal sus cosmovisiones mientras realizan su labor como prescriptores de productos y servicios. Estos *influencers*, muchas veces al servicio de los populismos globales de ultraderecha, mantienen, por tanto, un rol central en la sociedad de hoy, que permea diferentes industrias y esferas como la política, la publicitaria o la cultural.

El análisis sobre el actual contexto sociotécnico no estaría completo sin una aproximación a la convergencia entre inteligencia artificial (IA) y desinformación. En el tercer capítulo del manual, Fernando Bordignon coloca en tiempo presente la revolución algorítmica, marcada por los últimos desarrollos en inteligencia artificial generativa y sus efectos sobre el fenómeno de la desinformación a partir de la proliferación de nuevas fábricas posdigitales de (falsas) noticias. Bordignon presenta una acertada descripción de estas factorías algorítmicas de desinformación, caracterizadas por su capacidad de producción de falsedad a una velocidad y escalas sin precedentes. La emergencia de las *deepfakes*, la automatización de la producción noticiosa y la proliferación de portales sintéticos plantean desafíos inéditos para la verificación de la información, la preservación de la verdad y la defensa de la democracia. En este contexto, resulta urgente el desarrollo de marcos normativos, éticos y tecnológicos que regulen el uso de estas herramientas, así como la promoción de una ciudadanía crítica y alfabetizada digitalmente.

Como continuación y complemento, el profesor David García-Marín profundiza –en el capítulo cuarto– en la tipología de contenidos falsos generados con IA. Fundamenta su análisis en las últimas investigaciones realizadas en la confluencia entre sistemas de IA y desinformación y propone un mapa de las narrativas falsas más frecuentemente explotadas por los desinformadores que utilizan instrumentos algorítmicos con fines económicos, políticos e ideológicos. Establece, asimismo, una taxonomía de las *deepfakes*, los contenidos en vídeo o audio producidos con modelos de IA donde se muestra a determinados personajes realizando acciones o pronunciando discursos que jamás se produjeron. De acuerdo con las instituciones internacionales, el impacto potencial de estos productos desinformativos puede ser letal para nuestras democracias en términos sociales y económicos; pero también para los ciudadanos a nivel individual, puesto que son capaces de dañar la reputación, la imagen personal y la salud mental de forma profunda.

Estos cuatro primeros capítulos presentan un diagnóstico, un análisis sobre la desinformación en la contemporaneidad y de sus posibles desarrollos futuros. Plantean las coordenadas del momento actual en cuatro ejes: contenidos desinformativos, tecnologías de frontera (con la IA generativa a la cabeza), discursos de odio en el ámbito digital y la tecnopolítica protagonizada por los neofascismos globales. Aspectos que apenas han sido abordados en su convergencia en ninguna obra precedente, dada su emergencia y novedad.

Tras el establecimiento del mapa inicial de partida que caracteriza el presente sistema tecnológico e informativo, los dos capítulos restantes, el quinto y el sexto, abordan dos de las soluciones más efectivas y analizadas en la primera década de estudios contra la desinformación: (1) la verificación de contenidos y el *fact-checking,* y (2) la alfabetización mediática e informacional de las audiencias. Es evidente que no son los únicos diques de contención posibles, ya que existen otros abordajes al fenómeno de la desinformación desde otras ópticas como la regulación del ecosistema informativo o la imposición

de controles a las plataformas digitales para evitar la difusión de mensajes falsos o maliciosos, pero sí han sido las vías de combate contra la desinformación más investigadas durante los últimos años.

En esta línea, en el capítulo quinto, David García-Marín explica el método y las técnicas más comúnmente utilizadas en el proceso de verificación de contenidos. Se trata de un texto con un carácter más instrumental y aplicado, ya que pretende configurarse como una guía para todo aquel que quiera adentrarse en el conocimiento del periodismo de verificación y el *fact-checking*, no solo desde el ámbito profesional, sino como ciudadano comprometido y crítico contra las campañas desinformativas. La propuesta de García-Marín parte de la hipótesis de que todo ciudadano del siglo XXI debe actuar como verificador del inmenso caudal de mensajes falsos, engañosos, descontextualizados y maliciosos que recibe a diario y, por ello, las técnicas y las herramientas habitualmente utilizadas por los verificadores deben ser conocidas por la ciudadanía (al menos a un nivel básico), a fin de dotarnos de sociedades más resilientes contra la desinformación. Para ello, es fundamental comprender que el vídeo o la fotografía falsa son la cara más visible y evidente de la desinformación, que siempre oculta una capa invisible: las estrategias a cuyo servicio se ponen los contenidos falsos. En este sentido, el capítulo explica algunas de las tácticas desinformativas más frecuentes, apenas analizadas en ámbitos académicos y desconocidas para el gran público, tales como el *ataque mariposa (butterfly atrack)*, las *técnicas de supresión* (*doxing*), el blanqueo de información (*information laundering*) o el *astroturfing.*

Aunque el manual no pretende adoptar un carácter tecnocéntrico, el capítulo se complementa con un anexo a modo de repositorio de herramientas tecnológicas gratuitas y sencillas de dominar que la ciudadanía puede utilizar para diferenciar la información falsa de la verdadera. Esta práctica se erige como una necesidad perentoria, dado que casi seis de cada diez ciudadanos a nivel mundial (el 58%) muestra inquietud por su capacidad para distinguir qué es verdadero y qué es

falso al consumir noticias online, según el informe Digital News Report 2025 del Instituto Reuters y la Universidad de Oxford.

Finalmente, en el sexto capítulo, Roberto Aparici, Manuel Álvarez Rufs y Fernando Bordignon proponen un abordaje esencial para combatir la desinformación: la necesaria alfabetización de las audiencias desde el lado de la educomunicación. En concreto, la propuesta se centra en una dimensión apenas tratada en el ámbito académico hasta la fecha: la perspectiva ecológica. A partir del novedoso concepto de *eco-educomunicación*, este capítulo defiende la necesidad de incorporar en los procesos de alfabetización mediática la capa más invisible de la tecnología y del sistema tecnológico-informativo de hoy, que tiene que ver con su dimensión física: (1) servidores de alto impacto energético para el mantenimiento del entramado cibernético que nos rodea y para el entrenamiento de los algoritmos de IA, y (2) extracción de minerales necesarios para la fabricación de *hardware*. A la vez, el capítulo estudia las diferentes estrategias y dispositivos conceptuales para aplicar la eco-educomunicación a fin de combatir una de las manifestaciones donde más comúnmente se materializa la desinformación: los discursos de odio.

Este manual está especialmente dirigido a estudiantes de Comunicación y, en general, para cualquier titulación en el ámbito de las Ciencias Sociales. Puede resultar útil como una obra introductoria en actividades formativas esenciales relacionadas con la alfabetización mediática, informacional y algorítmica, incluso para abordar temas que anteriormente no se habían tratado dentro del ámbito de la educomunicación. El libro también interesará a docentes universitarios que buscan recursos didácticos para preparar sus clases en estas áreas. Asimismo, puede ser una herramienta valiosa para investigadores que deseen desarrollar marcos teóricos para sus estudios. Y, por supuesto, gracias a su lenguaje claro y enfoque pedagógico, resulta accesible para cualquier tipo de lector interesado en la materia.

1. Fundamentos de la desinformación

Roberto Aparici
UNED
España
Manuel Álvarez Rufs
UNED
España
Fernando Bordignon
UNIPE
Argentina

Introducción

Este primer capítulo tiene por intención acercar una serie de conceptos fundamentales para comenzar a trabajar y reflexionar en torno al tema de la desinformación como estrategia de control de masas a partir de instalar medias verdades, o incluso mentiras, desde el periodismo y otras plataformas como las redes sociales.

Para ello trabajaremos sobre tres conceptos, a saber: a) posverdad, como la emisión de mensajes que apelan a incentivar emociones y creencias tratando de disminuir y/o anular el razonamiento basado en hechos generando "nuevas realidades" basadas en "hechos alternativos"; b) desinformación, como táctica de "combate o guerrilla mediática", donde se manipula a los usuarios de información a partir

de manipular hechos, o incluso inventarlos, con la intención de generar sentidos que determinen distintos tipos de "ganancias" para grupos de poder determinados; y c) sesgos cognitivos, entendidos como una suerte de "atajos mentales" que a menudo colaboran en un procesamiento de situaciones mucho más eficiente en tiempos, pero por esos recortes o parcialidades, pueden llevarnos a cometer errores de juicio en nuestras apreciaciones personales.

Trabajar y reflexionar sobre la base de los conceptos anteriores proporciona un escenario más amplio acerca de cómo ciertos medios han estado cambiando sus narrativas en función de tratar de influir sobre las poblaciones, sembrando sentidos a partir de medias verdades o mentiras completas, que en general responden a intereses políticos y/o corporativos que distorsionan drásticamente su fin y aporte a la sociedad. Así, es importante que cada lector y lectora pueda hacer suyos los saberes aquí compartidos, resignificándolos y poniéndolos en diálogo con su entorno, es decir su propia realidad cotidiana.

Posverdad

El vocablo posverdad[3] aparece a finales del año 2017 en la versión electrónica del diccionario de la lengua española de la Real Academia (RAE) como un neologismo procedente del término en lengua inglesa *post-truth*. El término está formado por el prefijo *pos-* y *verdad*, y queda definido como "distorsión deliberada de una realidad, que manipula creencias y emociones con el fin de influir en la opinión pública y en actitudes sociales".

En este caso, hay que tener en cuenta que el significado del prefijo pos-, (*post-* en lengua inglesa), no se refiere a un tiempo después de que se produzca una situación o evento específico, sino que indica la pertenencia a un tiempo en el que el concepto especificado ha perdido

3 El primer uso del término en España se produce en el año 2003 en la obra de Luis Verdú "*El prisionero de las 21:30*".

la importancia o resulta irrelevante. Es decir, posverdad no significa algo que ocurre después de la verdad, sino algo que ocurre sin tener en cuenta la verdad. No se utiliza en un sentido temporal, "sino en el sentido de que la verdad ha sido eclipsada, de que es irrelevante" (McIntyre, 2018). El primer uso de la palabra *post-truth* se atribuye a Steve Tesich en el artículo *The Watergate Syndrome. A Government of Lies* publicado en The Nation en el año 1992:

> Las implicaciones son aún más aterradoras. Rápidamente, nos estamos convirtiendo en prototipos de personas que los monstruos totalitarios solo podrían babear en sus sueños. Todos los dictadores, hasta ahora, habían tenido que trabajar duro para suprimir la verdad. Nosotros, por nuestras acciones, estamos diciendo que esto ya no es necesario, que hemos adquirido un mecanismo espiritual que puede despojar a la verdad de cualquier significado. De una manera muy fundamental, nosotros, como personas libres, hemos decidido libremente que queremos vivir en un mundo posverdad. (Tesich, 1992, p.13)

Si atendemos a la definición original del *Oxford Dictionary* de *post-truth*[4], la posverdad aparece definida como un adjetivo que "se relaciona o denota circunstancias en las que los hechos objetivos son menos influyentes en la formación de la opinión pública que las apelaciones a emociones y a creencias personales". En el caso de *Cambridge Dictionary*, la posverdad (*post-truth*[5]) se define como un adjetivo "relacionado con situaciones en las que las personas son más propensas a aceptar un argumento basado en sus emociones y creencias, en lugar de uno basado en hechos".

Para McIntyre (2018), la posverdad no se trata de la realidad; se trata de la forma en que los humanos reaccionan a la realidad:

---

4 Post-truth: Relating to or denoting circumstances in which objective facts are less influential in shaping public opinion than appeals to emotion and personal belief. https://www.theguardian.com/books/2016/nov/15/post-truth-named-word-of-the-year-by-oxford-dictionaries

5 Post-truth: Relating to a situation in which people are more likely to accept an argument based on their emotions and beliefs, rather than one based on facts. https://dictionary.cambridge.org/dictionary/english-spanish/post-truth#google_vignette

> El primer paso para combatir la posverdad es comprender su génesis. El fenómeno cuenta con profundas raíces que nos llevan miles de años atrás, a la evolución de irracionalidades cognitivas que se comparten por liberales y conservadores por igual. [...] También tiene sus raíces en los debates académicos sobre la imposibilidad de acceder a la verdad objetiva que han sido usados para atacar a la autoridad de la ciencia. Y todo esto se ha visto exacerbado por los cambios recientes en el panorama de los medios. [...] Aunque se relaciona Brexit y elección de Trump con la posverdad, no son la causa, sino el resultado de ella. (McIntyre, 2018)

Hedonismo cognitivo

Ibáñez Fanés (2017, pp. 34-35) advierte que la posverdad fragmenta y contrae la lógica de un espacio libre y con una ciudadanía mínimamente competente en el manejo de la información "hasta regresar de nuevo al mundo oscuro de la secta, de la creencia privada y particular emancipada de toda idea de comunidad y de compartición de saberes e informaciones". Según el autor, "la posverdad no es el viejo perro de la propaganda y la mentira, o de la ocultación y el secreto propios de la política, armado ahora con un collar nuevo, sino que realmente estamos hablando de otra cosa". Ibáñez Fanés sitúa la clave de la posverdad en el "hedonismo cognitivo", entendido como el hecho de que la ciudadanía sólo lea y escuche lo que le gusta oír y escuchar, de manera que la clase política "no se preocupa de si lo que dice es verdad o mentira, sino sólo de si complacerá a los suyos, a su público" (p. 31-32).

Truthiness

En este punto, resulta interesante mencionar el término *truthiness* que es utilizado para referirse a la cualidad de parecer y sentirse como verdadero, incluso sin ser necesariamente cierto. Stephen Colbert popularizó el término al utilizarlo en el primer episodio de *The Colbert Report*[6] para referirse a ideas que "parecen verdad" o que "deberían serlo".

---

6 https://youtu.be/Ck0yqUoBY7M?si=Ap93824iHbCWFMii

Desinformación

La palabra desinformación proviene de la traducción literal del término ruso *dezinformatsia*, el cual aparece en el diccionario de la lengua rusa de S. Ojegov en el año 1949 donde se le define como "la acción de inducir a error mediante el uso de informaciones falsas" (Durandin, 1995). El diccionario de la lengua española de la RAE refiere que desinformar es "dar información intencionadamente manipulada al servicio de distintos fines", y también "dar información insuficiente u omitirla". Según Wardle y Derakhshan (2018), es posible distinguir entre:

- **Desinformación:** Contenido intencionalmente falso y que ha sido diseñado específicamente para hacer daño con la motivación de ganar dinero, tener influencia política, o simplemente causar problemas por el mero hecho de hacerlo.
- **Información errónea:** Se trata de una desinformación que se comparte sin que la persona se esté dando cuenta de que se trata de un contenido falso o engañoso.
- **Información maliciosa:** Utiliza información genuina que está basada en la realidad pero con la intención de causar un daño. En este caso se pueden incluir filtraciones, situaciones de acoso y discursos de odio.

Por otro lado, Wardle y Derakhshan definen siete categorías de "desorden informativo" o "trastorno de la información":

- **Sátira o parodia:** sin intención de causar daño, aunque tiene potencial para poder engañar.
- **Contenido impostor:** se suplantan las fuentes genuinas.
- **Contenido engañoso:** uso engañoso de la información para enmarcar problemas o individuos.
- **Conexión falsa:** los titulares, las imágenes o los subtítulos no se corresponden con el contenido.

- **Falso contexto:** el contenido genuino se comparte junto con información contextual falsa.
- **Contenido manipulado:** se manipula información o imágenes genuinas para engañar.
- **Contenido fabricado:** el nuevo contenido es cien por cien falso, y está diseñado para engañar y hacer daño.

Realidad-*Fake*-*Ficción*

Resulta interesante tener en cuenta las reflexiones de Rodríguez Ferrándiz (2018) sobre las paradojas de la etimología ante la ficción y los hechos, términos que parecen contradictorios, pero que se revelan próximos en la paranomasia entre *fictum* y *factum*, *res factae* y *res fictae*, cosas hechas y cosas fingidas o imaginadas. Según el autor, en lengua inglesa estaríamos hablando de *fact* y *fiction*, pudiendo encontrar un término que se sitúa entre los dos, "que es uno pretendiendo ser el otro", siendo este término *fake*, que procede directamente del *facere*, del hacer:

> Un sinónimo de *fake* es *counterfeit*. En español existe "contrahacer" y "contrahechura", pero ya desusados, en el sentido precisamente de imitación fraudulenta de alguna cosa, con propósito de engañar. Se emplea todavía el adjetivo "contrahecho", pero más como una deficiencia física o formal que moral, como sucede en inglés o en castellano antiguo. En estos últimos casos, precisamente su perfección formal es lo que lo condena, porque aspira a suplantar al original. (Rodríguez Ferrándiz, 2018)

"*Als ob…*": "Como si…"

Según Fuller (2018), Hans Vaihinger, fundador del *Kant Studien*, es un filósofo que "personifica la sensibilidad posverdad", y que ofrece orientación para navegar a través del ambiente intelectual posverdad mediante una visión completa del mundo en torno del uso repetido de la frase kantiana "*als ob…*", ("como si…").

El hecho de pensar o de actuar "como si" ciertas cosas fuesen verdaderas, a pesar de que tal vez nunca puedan llegar a ser demostradas, o incluso lleguen a ser completamente falsas, conduce a una visión del mundo resultante denominada "*ficcionalismo*" (Vaihinger, 1924). En el "*ficcionalismo*", la realidad se mueve entre el polo de la ficción, (donde no sabes que habitas un mundo falso), y el de la hipótesis, (donde sabes que no habitas en un mundo falso), sin llegar en ningún caso a alcanzar un sentido robusto de la noción de verdad.

Desinformación predigital

Según Ireton y Posetti (2020), "la desinformación es una vieja historia, impulsada por las nuevas tecnologías". Las autoras recuerdan un registro temprano de desinformación en la antigua Roma con el caso de Octaviano, enemigo político de Cleopatra que lanzó una campaña de desprestigio contra su rival Marco Antonio, mediante la inscripción de lemas cortos sobre las monedas en curso, lo cual le sirvió para convertirse en el primer emperador romano.

Por su parte, Illades (2018) nos recuerda la historia de la campaña egipcia de Napoleón Bonaparte como un posible caso de noticias falsas. La campaña resultó ser un fracaso rotundo, pero en Francia se tuvo otra idea muy distinta, debido a las noticias que afirmaban un enorme éxito del poderío francés. Napoleón, "al igual que en campañas previas como la de Italia, se alió con la prensa y los artistas de moda para convertir la expedición en algo glorioso". De esta forma, se escribieron notas sobre sus grandes hazañas, se pintaron cuadros de Napoleón en plena conquista y varios dramaturgos escribieron obras que exaltaban sus inexistentes triunfos.

En otro contexto, la cuidadosa edición del "telegrama de Ems" por parte de Otto von Bismarck convirtió un inocuo mensaje que informaba de la conversación del rey prusiano con el embajador francés en un mensaje lo suficientemente enardecedor, dado el contexto geopolítico

de la época, como para que Francia declarase la guerra a Prusia el 19 de julio de 1870 (Taylor, 1954).

McIntyre (2018) recuerda los días del "periodismo amarillo", cuando los magnates de medios informativos como William Randolph Hearst y Joseph Pulitzer estaban en guerra el uno con el otro por la venta de periódicos. McIntyre refiere que "en la década de 1890, plutócratas como Hearst y su *Morning Journal* utilizaron la exageración para ayudar a desencadenar la Guerra Hispanoamericana". Según Amorós (2018), Pulitzer y Hearst luchaban por erigirse en los amos del "cuarto poder". La noticia falsa de Hearst fue difundida por todos los medios estadounidenses, los cuales culparon a España. La opinión pública se agitó de tal manera que el gobierno se vio obligado a actuar, declarando la guerra a España:

> En 1898, el buque de guerra de la Marina de los Estados Unidos, el USS Maine, explotó mientras estaba fuera de La Habana, Cuba, matando a más de 250 estadounidenses. La causa nunca fue descubierta. Pero la prensa amarilla llegó a la conclusión de que los españoles lo hicieron deliberadamente. "Remember the Maine" se convirtió en el lema de la prensa amarilla, impulsando la opinión pública hacia la guerra. (Wolf, 2016)

El 30 de octubre de 1938, miles de estadounidenses escuchaban la transmisión del programa *Mercury Theater on Air*, creación del prodigio Orson Welles de apenas 23 años. Welles y un gran reparto de actores dramatizaban libros clásicos. Ese día se transmitió una adaptación de la novela de H. G. Wells *La guerra de los mundos* tan realista, que hizo pensar a muchos radioyentes que los marcianos realmente estaban invadiendo el planeta (Illades, 2018). Según Illades, la transmisión dejó una importante lección: "Siempre habrá quien acepte las cosas sin preguntar ni preguntarse sobre su veracidad, ya sea un programa de radio o televisión, una noticia de un medio con reputación intachable o un simple rumor disfrazado de verdad".

Fake News

Fake News (noticias falsas) es nombrada como palabra del año 2017[7] por el diccionario Collins, adquiriendo así cierta legitimidad, y quedando definida como "una información falsa, a menudo sensacionalista, divulgada bajo la apariencia de cobertura de prensa". Para Diego Rubio (2017), las noticias falsas son noticias en las que las falsedades aparecen por intención deliberada en lugar de por accidente o error.

Al hablar de noticias falsas, Ball (2017) utiliza "la definición original y muy estricta de historias que han sido totalmente inventadas, generalmente en un intento de alcanzar a una gran audiencia para una variedad de propósitos". McIntyre (2018) afirma que las noticias falsas "son un intento deliberado de lograr que las personas reaccionen ante la información errónea de cada uno, ya sea con fines de lucro o de poder. Pero, en cualquier caso, las consecuencias pueden ser nefastas".

Según Amorós (2018) "las *fake news* son informaciones falsas diseñadas para hacerse pasar por noticias con el objetivo de difundir un engaño o una desinformación deliberada para obtener un fin político o financiero". Para el autor, "las *fake news* son algo más que informaciones tendenciosas o manipuladas. Son mentiras".

El término *deepfake* se utiliza actualmente para describir medios fabricados mediante el uso de inteligencia artificial (en adelante, IA). Al sintetizar diferentes elementos de archivos de vídeo o audio existentes, la IA permite métodos relativamente fáciles para crear contenido "nuevo", en el que las personas parecen decir palabras y realizar acciones que no están basadas en la realidad (Wardle, 2018). De forma más concreta, el capítulo 4 analizará en profundidad este tipo de contenido elaborado con algoritmos de IA, caracterizado por ser altamente realista.

En marzo de 2022 se difundió un vídeo falso en el que aparecía el presidente ucraniano Volodímir Zelensky ordenando la rendición

7 https://theobjective.com/espana/2017-11-02/fake-news-palabra-del-ano-2017/

de su ejército y de su pueblo ante la invasión rusa, considerándose el primer *deepfake* utilizado en un conflicto armado para generar desinformación y desmoralizar a la población (Kardoudi, 2022). La IA permite la generación de contenidos hiperrealistas que se aprovechan del fenómeno de la posverdad y tratan de confundir emocionalmente a las audiencias. Ante un posible caso de *deepfake* podemos tratar de buscar incoherencias en la sincronización labial o utilizar herramientas específicas para el escaneo y detección de este tipo de contenidos, como puede ser Deepware.

Por otra parte, el desarrollo de la IA también puede ser utilizado de manera estratégica para generar dudas acerca de posibles evidencias que pueden resultar incómodas. Se trata de una dinámica denominada por los profesores de Derecho Bobby Chesney y Danielle Citron como *Liar's Dividend* o *dividendo del mentiroso*:

> A medida que el público internaliza que los vídeos y audios pueden falsificarse de manera convincente, algunos intentarán eludir la rendición de cuentas denunciando material auténtico como *deepfakes*. En pocas palabras: un público escéptico estará predispuesto a dudar de la autenticidad de pruebas reales en audio o vídeo. Este escepticismo se puede aplicar igualmente tanto a contenido auténtico como a contenido manipulado. (Chesney & Citron, 2019)

Según Goldstein & Lohn (2024) la teoría es sencilla: conforme las personas descubren que las *deepfakes* son cada vez más realistas, las afirmaciones falsas de que contenido real ha sido fabricado por IA también resultan más convincentes. Como ejemplo citamos el caso de un político opositor indio, quien se vio expuesto al difundirse dos grabaciones de audio en las que acusaba a miembros de su partido de corrupción al tiempo que elogiaba a su rival. Este político negó públicamente la evidencia de los audios y afirmó que las grabaciones habían sido generadas por inteligencia artificial. Finalmente, tras un análisis de las grabaciones por parte de diferentes compañías dedicadas a la evaluación y detección de *deepfakes*, se concluyó que al menos una

de las grabaciones era auténtica y que la otra podría estar manipulada (Cristopher, 2023).

Sesgos cognitivos

Los sesgos cognitivos actúan como atajos mentales que se utilizan para procesar la información a la que accedemos de manera más rápida, pero que pueden llevarnos a cometer errores de juicio en nuestras apreciaciones. En relación con el fenómeno de la posverdad, donde las emociones y creencias personales tienen más valor que los hechos objetivos, estos sesgos juegan un papel crucial. En un mundo saturado de información, especialmente en el contexto de las redes sociales, estos sesgos amplifican la difusión de noticias falsas y distorsionan la percepción de la realidad, debilitando el valor de la verdad objetiva y alejándonos de la realidad misma. Según Arias Maldonado (2017), la posverdad representa el triunfo de nuestros sesgos cognitivos sobre la realidad.

McIntyre (2018) nos advierte de un concepto central de la psicología humana, y es que las personas nos esforzamos por evitar la incomodidad psíquica. Esto se relaciona con la idea de hedonismo cognitivo que ya hemos visto anteriormente. Además, existe una disposición ancestral en las personas para decir mentiras y engañar a los demás, de manera que se puede hablar de una *deshonestidad situacional* entendida como la existencia de diferentes estándares de honestidad en las personas dependiendo de las diferentes configuraciones del contexto (Keyes, 2004). Según Keyes, existe un principio psicológico que está bien establecido y que consiste en que la mayoría de las personas operan en base a un *sesgo de verdad*. Esto quiere decir que las personas asumen que lo que alguien les diga tiene más probabilidades de ser verdadero que de ser falso. Sin embargo, al tratarse el engaño de algo cada vez más común, el prejuicio de la verdad podría dar paso a un *sesgo de mentira*, bajo el cual las personas llegan a cuestionar la veracidad de cualquier cosa que se les diga.

La resonancia emocional de la mentira hace que las falsedades se sigan repitiendo incluso años después de haber sido desacreditadas (Rabin-Havt, 2017). Se trata de una característica extraña de la cognición humana que implica que una vez que formamos una creencia o aceptamos un determinado reclamo, es muy difícil que las personas se desprendan de ello, incluso frente a una evidencia abrumadora y pruebas científicas de lo contrario (Levitin, 2017).

Por otro lado, la *disonancia cognitiva* implica un estado psicológico en el que creemos simultáneamente dos cosas que están en conflicto entre sí, lo cual genera una cierta tensión psíquica (McIntyre, 2018). La disonancia cognitiva describe la tensión que producen nuestras creencias cuando chocan con la realidad, ya que la mente nos autoengaña para obtener ciertos beneficios que considera más importantes que la aceptación de la realidad tal y como es (Amorós, 2018). La idea de *doble pensamiento*, entendida como el poder de mantener dos creencias contradictorias y aceptarlas ambas, es el antecesor directo de la posverdad (D'Ancona, 2017).

La *conformidad social* es la tendencia a estar de acuerdo con las creencias de las personas que nos rodean, incluso cuando la evidencia que se muestra a nuestros ojos nos dice todo lo contrario. Amorós (2018) nos recuerda que el cerebro siempre busca aliados, es decir, dentro de un mismo grupo social siempre es más fácil que se produzca una convergencia de la memoria. En esta línea, McIntyre (2018) se refiere al trabajo de Cass Sunstein, quien propone el término *efecto del grupo interactivo*, que se basa en la idea de que cuando las personas interactúan pueden alcanzar resultados que habrían eludido en el caso de haber actuado en solitario.

El *sesgo de confirmación* consiste en la tendencia a dar más peso a la información que confirma alguna de nuestras creencias preexistentes (McIntyre, 2018). Este efecto parece fortalecer las creencias con las que tenemos una conexión emocional mediante la búsqueda y retención de información que confirma nuestras ideas previas y la no aceptación de información que vaya en contra de ellas (Ball, 2017). Este sesgo se

relaciona con el *razonamiento motivado*, según el cual lo que esperamos que resulte ser cierto puede influir en nuestra percepción de lo que realmente es verdad, ya que en la mayoría de las ocasiones razonamos dentro de un contexto emocional, es decir, existe una tendencia a buscar información que respalde lo que queremos creer (McIntyre, 2018).

La *estrategia de prueba positiva* se refiere al hecho de buscar aquello que esperamos encontrar, mientras que la *asimilación sesgada* consiste en evaluar las ambigüedades a la luz de nuestras propias convicciones (D'Ancona, 2017).

El denominado *efecto de contrafuego* es un fenómeno psicológico en el que la presentación de información verdadera que entra en conflicto con las creencias erróneas de una persona hace que se mantengan y defiendan dichas creencias aún con más fuerza (McIntyre, 2018). Se relaciona con el hecho de que cuando se presenta evidencia en contra de una creencia profunda, en realidad sirve para reforzar esa creencia en lugar de desafiarla. El sesgo de confirmación nos protege cuando buscamos información de manera activa, mientras que el efecto de contrafuego nos defiende cuando nos llega la información que va en contra de nuestras creencias. En todo caso, existe un apego a las propias creencias en lugar de un cuestionamiento (Ball, 2017). El cerebro siempre quiere darnos la razón, para lo cual filtra la información que recibe destacando lo que se alinea con nuestras creencias e ignorando lo que nos contradice (Amorós, 2018).

El *efecto de Dunning-Kruger* es un fenómeno psicológico en el que nuestra falta de habilidad nos hace sobreestimar ampliamente nuestra habilidad real. Ocurre cuando individuos de bajas capacidades son incapaces de reconocer su propia ineptitud (McIntyre, 2018).

Según McIntyre (2018), un *silo de información* consiste en una tendencia a buscar información de fuentes que refuerzan nuestras creencias, bloqueando aquellas que no lo hacen. La *trampa del silo* se basa en la existencia de compartimentos estancos en la discusión pública, que impiden sacar conclusiones del conjunto de la realidad

(Estefanía, 2017). Estas tendencias, que alimentan la polarización social y la fragmentación del contenido de los medios (McIntyre, 2018), se ven reforzadas por los algoritmos de las redes sociales que impulsan el contenido agradable en la dirección del consumidor y eliminan las irritaciones, como los puntos de vista alternativos (Murphy, 2016).

El efecto *amnesia de la fuente* se produce cuando recordamos lo que leemos o escuchamos, pero no somos capaces de recordar si proviene de una fuente confiable (McIntyre, 2018). El autor también refiere que es más probable que creamos un mensaje si se nos ha repetido muchas veces. En este sentido, Amorós (2018) nos recuerda que el cerebro etiqueta con dificultad las noticias virales, produciéndose el *efecto de la verdad ilusoria*, según el cual nuestro cerebro etiqueta como más veraces las noticias que más se repiten.

Levitin (2017) advierte del error que se produce al trazar cosas que no están relacionadas:

- *Post hoc, ergo propter hoc* (después de esto, por lo tanto, debido a esto). Esta es una falacia lógica que surge de pensar que solo porque una cosa (Y) ocurre después de otra (X), esa X causó Y.
- *Cum hoc, ergo propter hoc* (con esto, por lo tanto, debido a esto). Esta es una falacia lógica que surge de pensar que solo porque dos cosas ocurren simultáneamente, una debe haber causado la otra.

Levitin (2017) también advierte que las estadísticas no son hechos sino interpretaciones, por lo que en el caso de porcentajes y promedios pueden producirse las siguientes falacias:

- *Falacia ecológica*: se produce cuando hacemos inferencias sobre un individuo en base a datos agregados (como la media de un grupo).
- *Falacia de excepción*: ocurre cuando hacemos inferencias sobre un grupo basadas en el conocimiento de unas pocas personas excepcionales.

Levitin (2017) también identifica los siguientes tipos de sesgo cognitivo: (1) *correlación ilusoria* (el cerebro es un gran detector de patrones, y busca extraer orden y estructura de lo que a menudo parecen ser configuraciones aleatorias), y (2) *efecto de encuadre* (las personas reaccionan a una elección en particular de diferentes maneras dependiendo de cómo se presenta la información).

Finalmente, según la Agencia Española de Protección de Datos (2024), otras debilidades o vulnerabilidades psicológicas y sesgos cognitivos persistentes y generalizados son:

- *Heurística del afecto*. El contenido que provoca emociones positivas influye significativamente en las decisiones de la persona usuaria.
- *Anclaje*. Las personas usuarias dependen demasiado de la primera información ofrecida (el ancla) al tomar decisiones.
- *Sesgo de automatización.* Las personas usuarias tienen a confiar de manera excesiva en sistemas automatizados o algorítmicos.
- *Efecto avestruz.* Se trata de la tendencia a evitar situaciones negativas obvias.
- *Statu quo*: Las personas suelen preferir una opción que no provoque ningún cambio, la tradicional.

Nótese el carácter necesariamenre conceptual de este primer capítulo de la obra. Resulta fundamental adentrarnos en el marco semántico de la desinformación para, a partir de aquí, comenzar a desarrollar cuestiones vinculadas con el actual fenómeno de la posverdad, como su confluencia con la IA, el papel de los medios de comunicación o la reacción de las instituciones educativas ante este desafío. También es básico comprender la esencia de los nuevos movimientos tecnosociales y cómo los creadores digitales se aprovechan de los sesgos cognitivos anteriormente explicados para lanzar campañas desinformativas con fines claramente ideológicos. Sobre estos espacios tecnopolíticos versa el siguiente capítulo.

Referencias

Agencia Española de Protección de Datos. (2024). *Informe sobre las implicaciones de los patrones adictivos en el tratamiento de datos personales*. AEPD. https://www.aepd.es/guias/patrones-adictivos-en-tratamiento-de-datos-personales.pdf

Arias Maldonado, M. (2017). Informe sobre ciegos: Genealogía de la Posverdad. En: Ibáñez Fanés, J. (ed.) (2017): *En la era de la posverdad. 14 ensayos*. Calambur.

Ball, J. (2017). *Post-Truth. How Bullshit Conquered the World*. Biteback Publishing.

Chesney, B., & Citron, D. (2019). Deep fakes: A looming challenge for privacy, democracy, and national security. *Calif. L. Rev.*, 107, 1753-1819.

Christopher, N. (5 de julio de 2023). *An Indian politician says scandalous audio clips are AI deepfakes. We had them tested*. Rest of World. https://bit.ly/3XNDdv3

D'Ancona, M. (2017). *Post-Truth. The new war on truth and how to fight back*. Ebury Press.

Durandin, G. (1995). *La información, la desinformación y la realidad*. Ediciones Paidós.

Estefanía, J. (2017). La mentira os hará eficaces. En: Ibáñez Fanés, J. (ed.) (2017): *En la era de la posverdad. 14 ensayos*. Calambur.

Fuller, S. (2018). What Can Philosophy Teach Us About the Post-truth Condition. En: Peters, M.A., Rider, S., Hyvönen, M., Besley, T. (2018): *Post-Truth, Fake News. Viral Modernity & Higher Education*. Springer.

Goldstein, J. A. & Lohn, A. (23 de enero de 2024). *Deepfakes, Elections, and Shrinking the Liar's Dividend*. Brennan Center for Justice. https://bit.ly/44dUi5k

Ibáñez Fanés, J. (ed.) (2017). *En la era de la posverdad. 14 ensayos*. Calambur.

Illades, E. (2018). *Fake News. La nueva realidad*. Grijalbo.

Ireton, C. & Posetti, J. (2020). *Periodismo, "noticias falsas" & desinformación: manual de educación y capacitación en periodismo*. UNESCO. https://unesdoc.unesco.org/ark:/48223/pf0000373349

Kardoudi, O. (5 de abril de 2022). *El primer 'deep fake' usado en un conflicto armado muestra a Zelenski rindiéndose*. El confidencial. https://bit.ly/4lhr9ML

Keyes, R. (2004). *The Post-Truth Era. Dishonesty and Deception in Contemporary Life*. St. Martin's Press.

Levitin, D. J. (2017). *Weaponized Lies. How to Think Critically in the Post-Truth Era*. Dutton.

McIntyre, L. (2018). *Post-Truth*. (The MIT Press Essential Knowledge series). MIT Press 2018.

Rabin-Havt, A. y Media Matters. (2016). *Lies Incorporated. The World of Post-Truth Politics*. Anchor Books.

Real Academia Española. (2024). Posverdad. En *Diccionario de la lengua española* (23ª ed.). https://dle.rae.es/posverdad?m=form

Rodríguez Ferrándiz, R. (2018). *Máscaras de la mentira. El nuevo desorden de la posverdad*. Ajuntament de València, Pre-textos. 2018.

Rubio, D. (2017). La política de la posverdad. *Estudios de política exterior 176*. pp. 58-67.

Taylor, A. J. P. (1954). *The struggle for mastery in Europe 1848–1918*. Clarendon Press.

Tesich, S. (1992). The Watergate Syndrome. A Government of Lies. *The Nation*, 12-14.

Vaihinger, H. (2009). (1925). *The philosophy of 'As if'. A system of the Theoretical, Practical and Religious Fictions of Mankind*. Martino Publishing.

Wardle, C. (2018). *Information Disorder: The Essential Glossary*. Shorenstein Center on Media, Politics, and Public Policy, Harvard Kennedy School.

Wardle, C., y Derakhshan, H. (2018). Thinking about 'information disorder': formats of misinformation, disinformation, and mal-information. *Journalism,'fake news'& disinformation*, 43-54.

Woolf, C. (2016). *Back in the 1890s, Fake News Helped Start a War*. Public Radio International, 8 de diciembre. Recuperado 15/10/24 de: https://bit.ly/3BKkyIB

2. Desinformación, *influencers* y tecnopolítica

Leonardo Murolo
Universidad Nacional de Quilmes
Argentina

Desde los orígenes: medios tradicionales y redes sociales

Sería una falacia sostener que las denominadas *fake news* solamente tienen lugar en los medios de comunicación digitales. El periodismo desde que es periodismo alberga en su práctica habitual diversas formas de desinformación. Sin arreglo a los lenguajes, los medios tradicionales como la prensa escrita, la radio y la televisión, pueden ostentar en perspectiva histórica notorios casos de información falsa, o como se dice en la jerga periodística "venta de pescado podrido". Desde el nacimiento hace años de una prensa amarilla hasta las actuales dinámicas de desinformación, *lawfare* o infodemia, el rol de la prensa está mucho más en debate público. Asimismo, el escenario actual que combina medios masivos con redes sociales se caracteriza por la sobreinformación y la ubicuidad de producción en los espacios digitales, diferenciador que potencia la aparición de contenidos falsos.

Como sea, es importante partir de cualquier conceptualización o análisis concibiendo a la información falsa o desinformación como una práctica intencionada y no como una equivocación o error. Tanto

por parte de medios de comunicación y periodistas entendidos como actores políticos, como de cualquier otro sujeto o perfil que produce contenidos, debemos comprender que la puesta en circulación de información falsa tiene la intención para nada ingenua de incidir en el debate público.

Hacia comienzos del siglo XXI, se sentenció la existencia de una web 2.0, un paso adelante de internet donde las audiencias se convertirían en usuarios que mediante la interactividad podrían alojar información al nuevo medio. Primero con la masificación de los blogs, un verdadero furor que mediante sus usos fue tildado hasta de "periodismo ciudadano", luego con la potencia de las wikis y con la irrupción de los usos masivos de las redes sociales virtuales: Facebook en 2004, YouTube en 2005, Twitter en 2006, Instagram en 2010. La cantidad de información que comenzó a circular en internet venía de la mano de una sociedad de la información donde el acceso y la participación se entendía que podrían transformar esa información en conocimiento. Es por ese entonces que a este andamiaje teórico se le agregaría la existencia de brechas digitales, las cuales se asentaron en primera instancia en variables económicas de acceso a *hardwares*, *softwares* y conexión, pero que más adelante mostrarían las dificultades en las brechas de uso. Una vez accediendo a la red de redes, la gran cantidad de información disponible requiere de habilidades para seleccionar la información precisa para luego transformarla en conocimiento. Un escenario celebrado pero poco novedoso de "comunidad masiva" (Morozov, 2016) o cultura participativa (Jenkins *et al.,* 2016) que retoma la lógica de internet descentralizada para que muchos desconocidos produzcan en conjunto contenidos convergentes. En ese escenario tiene lugar toda una dimensión destacable para la industria cultural, como son las labores de las comunidades de sentido transnacionales que se reúnen a producir conocimiento, comunicación y entretenimiento, intercambio de archivos, debates públicos, aprendizaje de idiomas y producción *fandom* en los diferentes lenguajes. Asimismo, afloran instancias más políticas que de consumo como el ciberactivismo y la puesta en agenda

de temáticas que los medios tradicionales no ponen en sus portadas y horarios centrales. En ese contexto también tienen lugar las dinámicas de la desinformación, la actividad de *trolls*, la implementación de *bots* y los discursos de odio que interpelan a usuarios de las redes.

Desde aquellos años también se popularizaron vocablos en inglés como *fake news* y otros neologismos ambiguos como posverdad para denotar el cúmulo de información falsa que circula en la web. Palabras clave que enfrentaron y aún enfrentan a investigadores de las ciencias sociales y periodistas por ser los primeros o más notorios en diferenciarlos conceptualizarlos y reconceptualizarlos[8].

La notoriedad, desconcierto y moda de estos términos tienen su base en que los medios digitales sociales son producto de la constante intervención de las audiencias o usuarios quienes en medio de las determinaciones del mercado crean dinámicas cada vez que habitan un nuevo espacio digital. Las redes sociales se configuran paradójicamente como espacios de gestión privada que erigimos como escenarios privilegiados para el debate público contemporáneo. Por definición son páginas en blanco que se presentan con reglas muchas veces tácitas sobre los límites de su habitabilidad. Son los usuarios quienes completan el contenido de Facebook, Twitter o Instagram quienes con pretendida libertad editorial inician dinámicas de producción y de consumo de contenidos que, si bien obedecen a contratos de lectura diferentes a los que establecemos con los medios tradicionales, pueden despertar lecturas similares.

Al mismo tiempo, estas redes sociales son habitadas por perfiles de medios de comunicación tradicionales, celebridades y políticos que a través de cuentas verificadas ponen a circular información bajo el mismo formato que los ignotos pero con mayor repercusión dada la cantidad de seguidores. La miscelánea entre cuentas verificadas, que tienen cierta responsabilidad en la veracidad de sus publicaciones, y

8 En el capítulo "La posverdad es mentira. Un aporte conceptual sobre fake news y periodismo" (2019) hemos propuesto una diferenciación entre fake news y posverdad en términos conceptuales y analíticos.

cuentas desconocidas o personales proponen un escenario confuso a la hora de la formación de opiniones con lo que allí puede leerse o verse. Este escenario, por lo tanto, es privilegiado para hacer circular información falsa en titulares llamativos, con imágenes elocuentes y comentarios disruptivos, ya que la lógica de consumo de las redes sociales se asienta en *scrollear* sin ingresar a todos los enlaces que nos llevan fuera de la propia plataforma. Hay usuarios que dicen informarse mediante redes sociales más que con medios de comunicación. En ese escenario se encuentra el problema de la información legitimada, ya que esa información leída en titulares de enlaces que en general no se leen en su totalidad forman opiniones sesgadas.

Aunque sepamos que los medios de comunicación tradicionales parten de una línea editorial, un manual de estilo y la defensa de intereses, depositamos en ellos una confianza que se halla en su rol social como instituciones creadas para informar. Hemos aprendido a leer los medios de comunicación, a reconocer las partes y los géneros discursivos de un periódico, las secciones y columnas de un noticiero televisivo, las gramáticas de un programa de radio. Desde allí, de manera habitual la desinformación retoma estas formas aprendidas ya sea encubiertas en los propios medios tradicionales o como formatos propios de las redes.

Los medios tradicionales, se supone, saben construir noticias y contarlas a la ciudadanía que a su vez las necesita para tomar decisiones. Estas decisiones pueden tener que ver con las dinámicas de consumo más banales, pero también mediante los medios de comunicación se accede a la publicidad de los actos de gobierno, a reconocer a los candidatos en elecciones y a ideas a las cuales adherir. En definitiva, para construir nuestras cosmovisiones desde ya que recurrimos a un *habitus* donde se encuentran las clases sociales, las sexualidades, las creencias religiosas, el nivel educativo, las tradiciones aprendidas de la familia, los consumos culturales. En esa trama, sin pretensión de determinar, es con los medios de comunicación que también se establecen líneas desde donde construir ideologías e identidades.

Tecnopolítica, ciberactivismo y movimiento red

En los últimos años, con la proliferación de la comunicación política en redes sociales, se ha publicado abundante bibliografía que nos permite reflexionar sobre el escenario de los activismos y militancias en los territorios virtuales. Un conjunto de prácticas que habitualmente se denominan *tecnopolítica* (Candón-Mena y Montero-Sánchez, 2021). Como primera definición, podríamos decir que se trata de la comunicación política, electoral, gubernamental, de crisis y de riesgo que producen instituciones, partidos y actores políticos en escenarios digitales. Es interesante pensar que en estas prácticas se desarrolla un lenguaje y un conjunto de narrativas específicas.

Más allá de las determinaciones técnicas de las plataformas, que permiten una cantidad de texto, imágenes o tiempo de contenido audiovisual, hay una forma de presentarse como sujeto político en las redes que por momentos difiere de otras formas, como el acto de masas, la representación por proximidad o los medios masivos (Murolo, 2023). Estas dinámicas son conocidas por los equipos de comunicación, quienes adiestran candidatos, planifican estrategias y recomiendan narrativas.

Por su parte, el concepto de *ciberactivismo* se refiere a un activismo o militancia digital que puede o no ser partidario. Por definición, como *matrioshkas*, entra en la categoría de tecnopolítica, pero se puede caracterizar desde su especificidad. La bibliografía habla de ciberactivismo cuando se trata de dinámicas que apelan a una perspectiva *hacker*, progresista, disruptiva o de revuelta: cuando desde redes y aplicaciones se desafía al *statu quo* y, por extensión, a los medios masivos. Podrían incluirse en esta idea temáticas que habitualmente no ocupan las agendas mediáticas de periódicos, televisión y radios *mainstream*, por ejemplo, demandas y reivindicaciones feministas, LGBTIQ+, multiculturalistas, ambientalistas, antirracistas, antiespecistas, anticapitalistas. En la práctica, se trata de intentar, a

través de estrategias comunicacionales, incluir en la agenda pública temas que no están más que en los bordes de los medios masivos.

Fabio Malini y Henrique Antoun (2017), en su libro *La internet y la calle. Ciberactivismo y movilización en las redes sociales*, realizan una genealogía de usos de internet y desarrollan la importancia de pensar el activismo digital en el ineludible cruce de dos instancias contiguas que refieren como la internet y la calle, como un diálogo virtual y presencial. En ese escenario pueden pensarse los fenómenos del ciberactivismo como contigüidad y correspondencia con la presencialidad.

Por otra parte, Marcela Fuentes, en *Activismos tecnopolíticos*, introduce la idea de *performance* al activismo, como una dimensión que puede verse en internet y en las calles. La autora puntualiza en algunos casos importantes: "El #MeToo (#YoTambién), *hashtag* de denuncia del acoso sexual, como fenómeno expandido y naturalizado, es un ejemplo (...)". "Otros *hashtags* se suman a esta práctica de tecnopolítica feminista, y transforman las redes sociales en espacios de discusión, posicionamiento y visibilización. Por ejemplo, en Estados Unidos, #WhyIStayed (#PorQuéMeQuedé) agrupa relatos que intentan responder a acusaciones de complicidad en el fenómeno de la violencia doméstica. Por su parte, #SayHerMe (#DiSuNombre) denuncia la violencia policial hacia las mujeres negras y a la vez cuestiona la centralidad de los varones como eje del duelo en movimientos como Black Lives Matter" (Fuentes, 2020, p. 200). La autora trabaja #NiUnaMenos como caso paradigmático argentino y luego latinoamericano que nace el 3 de junio de 2015 por impulso de colectivos feministas como un *hashtag* en redes, pero que pronto se convirtió en una dinámica militante que llevó a mujeres a llenar las plazas con proclamas.

Existen casos donde los activismos feministas coparon las calles bajo el recurso de la *performance*, como el colectivo chileno Las Tesis, las conferencias y puestas del colectivo Actrices Argentinas, o en forma de *cosplay* vistiendo como las subyugadas mujeres de Gilead en la novela y serie de plataforma *El cuento de la criada*. Es entonces cuando movimientos feministas salen a producir activismo tecnopolítico en

términos de *performance* habitando coreografías y ropas. Mientras hacen referencia a la cultura pop, interpelan políticamente a la sociedad.

Las denominadas *derechas alternativas* (*alt-right*), *nuevas derechas* o neofascismos, emergieron con una importante base en juventudes insatisfechas por la democracia y sin variantes en tanto las perspectivas de futuro. En tanto lo discursivo, lograron erigirse como sector "antisistema" aunque pregonan la tiranía del capitalismo y corrieron el umbral de lo decible en tanto poner en debate los derechos humanos, las libertades individuales o las políticas de memoria, desde intervenciones cargadas de violencia en el espacio público. Los medios masivos y las redes sociales comenzaron a recibir a sujetos autopercibidos como "libertarios", embanderados de ideas xenófobas, homófobas y racistas. En los foros de 4chan y Reddit a los que hacíamos mención, se desarrollaron dos lógicas necesarias para la proliferación de discursos de odio: la primera es el anonimato, al no tener que verificar la identidad de ningún modo y escudándose detrás de un *nickname* y un avatar; en segundo lugar, la dinámica gregaria o en horda, los usuarios se envalentonan y juegan a quién corre más los límites, en este caso, de la violencia. De este modo, se fueron abriendo y habitando canales, conversaciones y *subreddits* que lentamente construyeron un sentido común violento capaz de naturalizar la violencia. Este caldo de cultivo migró y se complementó con otras conversaciones similares en X y en otras redes sociales *mainstream*. Allí, los medios masivos nunca ingenuos, con sus líderes de opinión de horarios centrales también en la medida de lo posible participaron de esta germinación.

En *¿La rebeldía se volvió de derecha?*, Pablo Stefanoni (2021) rastrea primeramente el fenómeno no solamente en torno a dimensiones políticas y económicas, repasando los principales autores "libertarios" y sus postulados. Luego también explora imaginarios y escenarios mediáticos, con implicancias culturales, cuando estas perspectivas toman como puntos de ataque colectivos progresistas y conquistas recientes de derechos sociales. Allí, las juventudes son actores privilegiados en la construcción y adopción de estas

subjetividades políticas. El autor incluso plantea la emergencia del homonacionalismo, que se opone a las migraciones sobre todo a las árabes en Europa, y perspectivas ambientalistas de derecha, que recuperan la reivindicación de la propia tierra.

Un conjunto de sentidos se pone en juego en esta batalla cultural, los neofascistas en términos económicos pondrán de su lado "las ideas de la libertad", la disolución del Estado, la exaltación de la figura del empresario y la defensa primordial de la propiedad privada. Por otro lado, en términos estrictamente sociales y culturales atacarán al "marxismo cultural", "la ideología de género", "los guerreros de la justicia social", "generación de cristal", "*woke*" o "progres", quienes supuestamente detentan una hegemonía cultural.

En su libro *La era del conspiracionismo*, Ignacio Ramonet (2022) examina el giro hacia la derecha en la política contemporánea, centrando su análisis en las condiciones políticas, mediáticas y culturales que llevaron a Donald Trump a la presidencia de los Estados Unidos. Ramonet analiza cómo Trump, un famoso millonario y figura mediática, se subió a demandas de un sector descontento con la política, utilizó tanto líderes de opinión de los medios como *influencers* de redes sociales para la difusión de ideas y llegar al poder. El autor argumenta desde casos que parecen ser aislados, pero que guardan relación con una subjetividad de época, en donde se atraviesa una "desconfianza epistémica" por los hechos y se apela a la emocionalidad (Ramonet, 2002, p. 39). Los casos del Pizzagate, el surgimiento de QAnon en Reddit, durante la campaña, y el advenimiento de la pandemia y las teorías conspirativas que suscitó, son ejemplos de episodios que le sirvieron a Trump y sus seguidores para embarrar la cancha en contra de sus opositores políticos. Este caso, de un *influencer* conservador y reaccionario, nos sirve para evidenciar que las dinámicas propias de las redes como las *fake news*, la posverdad y el conspiracionismo son usadas por estos sectores para construir imagen pública.

En *Paisajes insurrectos. Jóvenes, redes y revueltas en el otoño civilizatorio*, Rossana Reguillo retoma el término *movimiento red* y

denomina como *superficie de inscripción* al "espacio social y digital en el que las personas inscriben, a través de palabras, imágenes o gestos, sus imaginaciones y deseos, sus miedos y esperanzas, sus odios y afectos" (2017, p. 88). Estos espacios, en la lógica del activismo político, pueden ser de forma presencial, como pancartas y grafitis, o digitales, como publicaciones en redes sociales. Por lo tanto, se despliegan instancias entre la internet y la calle como continuidad y retroalimentación. Desde allí trabaja, entre otros, los movimientos de #BlackLivesMatter, #NosFaltan43, #OccupyWallStreet, encontrando tanto puntos en común como particularidades.

Reguillo (2017, p. 105) define los *repertorios de la acción conectiva*, que lo integran aunque no lo agotan, en torno al *streaming*, la *memética*, el *micrófono humano* y el *hashtag*. El *streaming* es utilizado por militancias cuando los medios masivos no cubren sus reclamos en las calles, transmitiendo en vivo a través de dispositivos móviles y redes sociales. El *micrófono humano*, observado en Occupy Wall Street, es una forma de rudimentaria comunicación oral en cadena que los manifestantes ensayaron cuando les interrumpieron la electricidad y no podían cargar los celulares. La *memética* se basa en el humor, la ironía y el sarcasmo para abordar temas sociales y políticos, a menudo borrando las huellas de autoría y permitiendo correr los límites de lo decible. El *hashtag* indexa y etiqueta temas de agenda, permitiendo a las audiencias buscar información ordenada. En el ciberactivismo, el manejo de la indexación de temas es crucial para burlar los algoritmos que privilegian ciertas cuentas y temáticas. A fuerza de reiteración y posicionamiento, la temática del movimiento red se hace un lugar en las redes y en la agenda pública.

A partir de que los repertorios de la acción conectiva no se agotan en las prácticas señaladas, es pertinente proponer como una instancia más la irrupción de sujetos denominados *influencers* en redes sociales y plataformas de *streaming*. Desde ellos podríamos leer algunas dinámicas tecnopolíticas.

Influencers

En el presente capítulo nos interesa desarrollar la emergencia mediática de sujetos *influencers*, quienes actúan como mediadores para proponernos no solo habilidades y consumos, sino también ideas y cosmovisiones. Partimos de la premisa de que no son ingenuos en esta práctica ni que se equivocan cuando vierten opiniones políticas en medio de sus contenidos de temática variada. También es objetivo de nuestro trabajo considerar que estos repertorios, que fueron concebidos para describir los movimientos en red y los ciberactivismos, participan en instancias tecnopolíticas más amplias que las progresistas y pueden ser adoptados como narrativas en disputas por miradas conservadoras o reaccionarias.

En otros trabajos se han explorado las figuras de *influencers* o creadores de contenido como sujetos que intentan determinar comportamientos y prácticas de sus audiencias (Murolo y Del Pizzo, 2021). Desde allí, los denominados *influencers* de las redes sociales podrían leerse como signos rompecabezas. En la construcción de este rol participan los líderes de opinión, el sistema de estrellas y los modelos publicitarios. Si consideramos las teorías clásicas de la comunicación, hace un siglo una mirada ligada al Funcionalismo y a las formas administradas de la comunicación, entre otros conceptos fuertes, proponían la existencia de una comunicación en dos pasos, de líderes de opinión y de los efectos de los mensajes en los receptores. Estos conceptos fueron posteriormente cuestionados tanto por la Teoría Crítica como por los Estudios Culturales. Sin embargo, las teorías no son narrativas en las que una termina y otra empieza, sino que, siguiendo a Alcira Argumedo, se trata de *matrices de pensamiento* que conviven y desde diferentes epistemologías explican el fenómeno comunicacional. Por lo tanto, esas formas de comprender la comunicación siguen en pie en algunas apuestas de la publicidad y la propaganda.

En una sociedad compleja, los líderes de opinión juegan un papel crucial. Pueden ser figuras mediáticas que se convierten en

ídolos o personas cercanas con conocimientos específicos a quienes recurrimos para obtener información. Su importancia radica en que no podemos saberlo todo sobre todos los temas que nos interesan, por lo que necesitamos de su orientación. Incluso nosotros mismos podemos desempeñar este rol en algún ámbito. Según la teoría de los dos pasos, dependemos de otros para obtener información antes de llegar a la fuente primaria. Tradicionalmente, los líderes de opinión eran figuras como periodistas y políticos que aparecían en medios masivos y luego se destacaban en círculos sociales más cercanos. Por ejemplo, en la televisión argentina de las décadas de 1980 y 1990, los líderes de opinión más destacados eran periodistas políticos que transmitían una imagen de seriedad y respetabilidad. Aunque se presentan como neutrales, su narrativa refleja sus propias militancias, ideologías y visiones del mundo, al tiempo que descartan otras. En esencia, el concepto moderno de *influencer* tiene sus raíces en el tradicional líder de opinión de los medios de comunicación masivos.

Por otro lado, existe el sistema de estrellas o *star system*, que consiste en celebridades populares como la diva pop, el romántico latino o el galán de telenovela. A lo largo del siglo XX, cada generación tuvo sus propias celebridades, consideradas las personas más famosas y exitosas del mundo. En la actualidad, si observamos las redes sociales como indicador de popularidad, vemos que la cuenta más seguida en Instagram es la propia red, seguida por Cristiano Ronaldo y Lionel Messi. Esto indica que los futbolistas son actualmente las figuras más famosas a nivel mundial, en contraste con las décadas pasadas donde las divas pop o los cantantes románticos latinos dominaban la industria cultural. Estos estereotipos y roles, como actores, cantantes, músicos y deportistas, son fundamentales en la industria cultural y crean la relación necesaria entre ídolo y fanático. A diferencia de los líderes de opinión, el sistema de estrellas se basa en talentos y personalidades famosas que capturan la atención del público de una manera distinta.

Por su parte, el modelo publicitario es una figura característica del siglo XX. En los años 90, había un grupo selecto de modelos que

eran celebridades. La industria de la moda, con agencias de modelos, campañas publicitarias y desfiles, ha perdurado hasta hoy, pero ahora ve en los *influencers* una evolución. ¿En qué se parecen los *influencers* a los modelos? Ambos hacen publicidad de productos, servicios, moda y estilos físicos. Por ejemplo, un jugador de fútbol, además de su habilidad deportiva, aparece en anuncios y sirve como modelo a seguir para tendencias como cortes de cabello o tatuajes de moda.

Por último, los *influencers* de diseño ilustran cómo la industria cultural reconoce los estereotipos como una fórmula y crea robots que desempeñan esos roles. Figuras como Lil Miquela, Blawko y Bermuda Is Bae, creadas por inteligencia artificial de la empresa Brud, tienen una gran cantidad de seguidores en redes sociales. Estos *influencers* virtuales visten marcas como Diesel, Alexander McQueen y Chanel, y comparten narrativas, cantan canciones y visitan lugares populares. En resumen, Brud ha captado la esencia de ser *influencer* en la intersección entre el sistema de estrellas y los modelos tradicionales.

Los denominados *influencers* de redes sociales manifiestan sus talentos, pero en suspensiones narrativas presentan productos y servicios, y de manera subliminal presentan opiniones ideológicas. En este contexto, estos sujetos cuentan con una cantidad considerable y una calidad de seguidores, dos elementos importantes para la difusión de ideas. Esto es de interés tanto para la comunicación como para el marketing, ya que las marcas los buscan para crear contenido. La calidad de los seguidores se refiere a variables como intereses, edad, aspectos socioculturales y ubicación geográfica.

A modo de cierre: los antídotos

Como hemos advertido en nuestro desarrollo, todas las formas de la desinformación constituyen un verdadero problema para el pleno ejercicio de la democracia comunicacional. Se advierte en el ámbito científico académico, se implementan iniciativas estatales a escala global, y fundamentalmente afecta no solamente a la ciudadanía en

su relación con la información sino también al genuino ejercicio del periodismo profesional.

En el entramado de la tecnopolítica, nos interesan un conjunto de prácticas y acciones conectivas en donde hemos puesto la mirada en las lógicas de los *influencers*. Estos sujetos de la comunicación con umbral de enunciación privilegiadoso n muchas veces quienes replican y amplifican las desinformaciones.

Desde allí es propicio que reflexionemos acerca de la posibilidad de señalar y llevar a cabo prácticas, a modo de antídotos, contra este problema. En primera instancia podemos enunciar tres formas de contrarrestar, esquivar y combatir la desinformación a las que denominaremos de manera provisoria como: tácticas de detección, educación en medios y recursos alternativos.

Tácticas de detección

Las propias redes sociales virtuales y las empresas de mensajería instantánea presentan en sus plataformas un listado de consejos para que los usuarios incorporen en la habitabilidad de estos espacios. En ese sentido, tanto Facebook como YouTube y Whatsapp plantean que ante una información que nos interesa en lo puntual que antes de compartirla o formarnos una opinión al respecto que chequemos su verosimilitud. Este ejercicio se puede emprender verificando la fecha en la que esta información fue publicada y que no se trate en principio de una información extemporánea. En este caso no sería una información falsa, pero si se trata de una noticia vieja presentarla como actual oficiará como un dato falso para formar opiniones erróneas. También se aconseja leer con atención la dirección web o URL para detectar que no se trata de una web apócrifa o imitación de un sitio reconocido. Como en las marcas de ropa falsas, los sitios web que reconocemos y con los que nos informamos también pueden tener imitaciones que modifican una letra de su nombre para asemejarse a marcas reconocidas y desde allí enunciar en nombre de su prestigio y credibilidad a sus audiencias.

Asimismo, habría que prestar atención a si la noticia o información no se trata de una broma o chiste. Existen en varios países medios de comunicación que basan su información en la ironía. En Argentina son notorios los casos de *Revista Barcelona* o de la página de memes *EAMEO*, que trabajan con humor y sarcasmo con la agenda mediática. Si algunos usuarios no conocen estos medios, y por lo tanto sus contratos de lectura, podrían compartir estas informaciones como verdaderas.

En otro orden de lectura se encuentran las imágenes, desde las cuales se podría enunciar una información que no es la que refiere. Se conocen casos de desinformación que apelan a fotografías de otros lugares geográficos o momentos históricos para graficar una información actual. Los mecanismos de detección de imágenes de los buscadores mostrarían en pocos segundos el origen primero de esas imágenes.

Estas y otras dinámicas de detección de desinformación, que serán ampliadas en el capítulo 5 de este libro, constituyen por un lado un sujeto de la comunicación digital activo, que se procura información realizando una curaduría de contenidos y de sus consumos. Sin embargo, debemos señalar la subyacente injusticia en tales dinámicas dado que los medios de comunicación tienen por fin brindarnos información chequeada y valedera para formar nuestras opiniones ciudadanas. Desde allí que tengamos que realizar todas estas tareas para encontrarnos con información verdadera se torna tedioso.

Educación en medios

Es lógico que quienes se dedican a las ciencias sociales en general y al campo específico de la comunicación en particular cuenten con herramientas para reconocer las dinámicas de los medios de comunicación, sus narrativas, usos y formas de habitarlos por parte de las audiencias. Sin embargo, los medios de comunicación, tradicionales y digitales, forman parte de las dinámicas de información, comunicación y entretenimiento de toda la ciudadanía. Por lo tanto,

sería propicio reforzar una educación en medios en escuelas primarias y secundarias que pongan foco en temáticas concretas.

Son diversas las materias que tematizan los medios de comunicación en el ámbito escolar y en general se tratan desde el grado cero de la ideología. Se enseña sobre lenguajes, teorías, géneros discursivos y en otras materias se producen contenidos escritos, sonoros, audiovisuales, multimedia. En el mejor de los casos se pone en tensión el ejercicio del periodismo como rol social y se tematiza la ética de la profesión. Sin embargo, en los casos que conocemos no se profundiza en el estudio de los medios masivos concentrados como empresas diversificadas que se presentan con línea editorial, manual de estilo, y desde allí defienden intereses empresariales y económicos, ideológicos y de clase. En ese contexto se puede sentar las bases de un estudio del consumo crítico de medios que pueda advertir las dinámicas de la desinformación como un elemento más del juego de poder de los medios masivos en tanto actores políticos, tal como se analizará con mayor profusión en el capítulo 6.

Recursos alternativos

Una posibilidad de esquivar la desinformación de los medios masivos es la de procurarse otras vías de información para cotejar. En ese sentido, existen una cantidad de medios de comunicación tanto tradicionales como digitales que albergan otras dinámicas de producir y contar información relacionada a sus formas de gestión y presentación de las agendas.

Es allí cuando debemos mirar el desarrollo de medios públicos tanto nacionales, como provinciales y municipales, medios de comunicación de gestión cooperativa, medios universitarios, sindicales, alternativos y participativos que propugnan por la comunicación como derecho humano y le hablan a sujetos parecidos a ellos mismos. Esto es una comunicación local que comparte intereses, agendas y preocupaciones, además de interpelar desde la horizontalidad comunicacional. En

este sentido, la opción de buscar otros medios además de los masivos y concentrados redunda en mayor información, otras campanas en los temas de agenda y otras agendas que incomodan a la información hegemónica.

En definitiva, la complejidad de la desinformación en el escenario digital exige abordar el fenómeno desde múltiples dimensiones que se complementan y potencian mutuamente. No alcanza con adoptar una única práctica aislada, sino que es imprescindible conjugar tácticas de detección que promuevan un consumo crítico y activo de la información; la educación en medios, que permita comprender las estructuras y lógicas de los sistemas comunicacionales; y la búsqueda consciente de recursos alternativos que diversifiquen y enriquezcan las agendas informativas. Solo así se podrá fortalecer un ejercicio democrático y ciudadano pleno, donde las audiencias no sean meras receptoras pasivas, sino sujetos críticos capaces de navegar con autonomía y responsabilidad en un entramado mediático cada vez más complejo y tecnopolítico.

Referencias

Aparici, Roberto y García Marín, David (2019) "La posverdad: el software de nuestra era". En *La posverdad. Una cartografía de los medios, las redes y la política.* Madrid: Gedisa.

Arrabal, Victoria (2019) "La posverdad es la gran mentira. Entrevista a Leonardo Murolo". En Página 12, 11 de mayo de 2019. Disponible en: https://www.pagina12.com.ar/192992-la-posverdad-es-la-gran-mentira

Candón-Mena, Jose y Montero-Sánchez, David (2021). "Más allá del ciberactivismo. El complejo escenario de la tecnopolítica contemporánea". En José Candón-Mena y David Montero-Sánchez (eds.), Del ciberactivismo a la tecnopolítica. Movimientos sociales en la era del escepticismo tecnológico (pp. 23-46). Salamanca: Comunicación Social Ediciones y Publicaciones.

Fuentes, Marcela A. (2020). Activismos tecnopolíticos. Constelaciones de performance. Buenos Aires: Eterna Cadencia.

Jenkins, Henry (2008) *Convergence Culture. La cultura de la convergencia de los medios de comunicación.* Buenos Aires: Paidós.

Morozov, Evgeny (2016) *La locura del solucionismo tecnológico.* Buenos Aires: Capital Intelectual.

Murolo, Leonardo (2019) "La posverdad es mentira. Un aporte conceptual sobre fake news y periodismo". En Aparici, Roberto y García Marín, David, *La posverdad. Una cartografía de los medios, las redes y la política.* Madrid: Gedisa.

Murolo, Leonardo (2021) "La dimensión simbólica de las pandemias. Distopía e infodemia entre ficción y realidad". En Aparici, Roberto y Martínez-Pérez, Jorge, *El algoritmo de la incertidumbre: los códigos invisibles de nues*Malini, Fabio y Antoun, Henrique (2017). *La internet y la calle. Ciberactivismo y movilización en las redes sociales.* Guadalajara: ITESO.

Murolo, Leonardo y Del Pizzo, Ignacio (2021). *Cultura pop. Resignificaciones y celebraciones de la industria cultural del siglo XXI.* Buenos Aires: Prometeo Editorial.

Murolo, Leonardo (2023). "Los cuatro escenarios de la comunicación política". *Cuadernos de Coyuntura*, Vol. 8 número continuo, Facultad de Ciencias Sociales, Universidad Nacional de Córdoba. https://revistas.unc.edu.ar/index.php/CuadernosConyuntura/article/view/42769

Ramonet, Ignacio (2022). *La era del conspiracionismo.* Buenos Aires: Siglo XXI Editores.

Reguillo, Rossana (2017). *Paisajes insurrectos: jóvenes, redes y revueltas en el otoño civilizado.* Barcelona, España: NED Ediciones - Guadalajara, México: ITESO.

Sierra Caballero, Francisco (2021). *Ciberactivismo. Disrupciones, emergencias y procesos de remediación.* España: Tirant Lo Blanch.

Stefanoni, Pablo (2021). ¿Por qué la rebeldía se volvió de derecha? Buenos Aires: Siglo XXI.

3. Las nuevas fábricas algorítmicas de desinformación

Fernando Bordignon
UNIPE
Argentina

Introducción

Para Agamben (2003), el término postdigital hace referencia a un paradigma –fruto de cierto ingreso a un estado avanzado de relación cotidiana con las tecnologías digitales– que intenta describir la oportunidad actual de explorar las consecuencias de este momento. Bajo esta perspectiva, la IA, entendida como un conjunto de técnicas algorítmicas disruptivas, se conforma como un elemento central que está provocando sorpresa y cambios en una amplia gama del quehacer humano. En este sentido, la aparición del servicio ChatGPT en noviembre de 2022, trajo y repuso una diversidad de miradas y discusiones acerca de la IA y sobre todo de aquello que es una novedad en sí mismo, la inteligencia artificial generativa (IAG). Por otro lado, desde la mirada sociotécnica, casi nada nuevo, estuvieron los que celebraron acríticamente hasta quienes se posicionaron como detractores de la novedad. Con el correr de los meses el período de sobreexpectación se ha transitado y hoy podemos decir, podemos

pensar más tranquilamente, fuera de este "tsunami tecnológico" y observar las apropiaciones y usos del "algoritmo que aprendió a narrar".

Este capítulo tiene por objetivo tratar de recuperar lo que ha pasado y ponerlo en diálogo con un presente postdigital, que hasta el momento trae más incertidumbres que certezas en torno a la integración al cotidiano de los mencionados algoritmos, y más aún cuando se observan sus usos en entornos políticos, mediáticos, periodísticos. Este capítulo está organizado en cuatro partes. La primera aborda y reflexiona sobre la emergencia de nuevas herramientas denominadas agentes cognitivos. La segunda problematiza las fallas "por naturaleza" de los algoritmos de la IAG y cómo, en determinadas situaciones, se puede convertir en una "máquina imperfecta". La tercera parte se centra en el efecto generador y propagador de desinformación que estas "nuevas fábricas" postdigitales está trayendo. A continuación, se caracterizan a las nuevas fábricas algorítmicas de noticias y sus efectos en la sociedad. Finalmente, se integran las ideas centrales abordadas en un apartado donde se narran las consideraciones finales.

La inteligencia artificial generativa y los nuevos agentes cognitivos

Los nuevos algoritmos programados con técnicas de IAG están colaborando en cambiar, de a poco y de manera efectiva, la forma en que pensamos, creamos y usamos el *software*. Ahora, en esta transición, en lugar de limitarnos a programar las computadoras indicando qué deben hacer, les estamos delegando el cómo se hace algo en función de objetivos concretos que les suministramos en lenguaje natural. Así, la novedad aportada por la IAG es que ahora el *software* se comporta como un agente que utiliza los grandes modelos de lenguaje (LLM) como infraestructura para determinar qué hay que hacer, cómo se hace algo y luego realizar, por sí mismo, la tarea encomendada, sin programación extra alguna. Desde esta perspectiva, el *software* que se comporta como

un agente cognitivo está colaborando en que los servicios de usuario más usados en internet dejen de ser solo proveedores pasivos de respuestas (como los buscadores) y pasen a ser sistemas inteligentes que pueden asumir funciones tradicionalmente reservadas a las personas.

Lo fascinante de esta nueva generación de *software*, en particular lo observado en ChatGPT o el servicio DALL-E, es que ha provocado la sensación de que hemos dado un salto importante en la aproximación al lenguaje natural, tanto en su comprensión y consideración (en especial el trabajo sobre la memoria de lo precedente y las técnicas básicas de análisis semántico) como en la producción de contenidos. Para el filósofo Luciano Floridi estos agentes cognitivos son revolucionarios por al menos dos elementos. Por un lado, porque el usuario sólo expresa qué quiere y el agente sabe cómo decodificar el requerimiento y producir un contenido con la potencial respuesta y, por otro lado, se ha desarrollado una forma novedosa de agencia de la inteligencia. Basada en un modelo conversacional, donde a partir de modelos probabilísticos sofisticados se logra responder, en general, de manera satisfactoria sin razonar ni comprender como lo hace un humano (Floridi, 2023). Así, su operación nada tiene que ver con los procesos cognitivos que habitan en los cuerpos del mundo animal (especialmente en la mente humana) para gestionar con éxito los contenidos semánticos (Bishop, 2021). Sobre esto, en particular, algunos investigadores han comparado el comportamiento de los LLM con el de los loros, llamando a estos servicios "loros estocásticos" ya que repiten textos sin comprenderlos (Bender *et al*., 2021). Hoy estamos frente a una forma de agencia inédita, ya que tiene éxito y puede aprender y mejorar su comportamiento sin necesariamente ser inteligente para ello. Así, la novedad es que hemos diseñado y construido *agentes sin inteligencia*, y esto es inédito en la historia de la humanidad ya que:

> [H]emos pasado de estar en contacto permanente con agentes animales y lo que creíamos agentes espirituales (dioses y fuerzas de la naturaleza, ángeles y demonios, almas o fantasmas, espíritus buenos y malos) a tener que entender, y aprender a interactuar, con agentes artificiales creados

> por nosotros, como nuevos demiurgos de tal forma de agencia. Hemos desvinculado la capacidad de actuar con éxito de la necesidad de ser inteligente, entender, reflexionar, considerar o comprender algo. Hemos liberado la agencia de la inteligencia. (Floridi, 2023)

Cualquiera sea la aplicación de IAG que se use, la carencia de inteligencia tiene su origen en que en su interior continúa siendo un modelo de lenguaje y no de conocimiento (como en los humanos); pero con un robusto sistema de predicción que es superior a cualquiera conocido hasta el momento en varios órdenes de magnitud. Ya que se alimenta con la información disponible y accesible en la red internet, la cual sabemos que es mucha, dado que pertenece a la categoría de los grandes datos.

Máquinas imperfectas por naturaleza

En este corto tiempo en que se ha interactuado y evaluado las nuevas aplicaciones de IAG, ya se pueden vislumbrar diferentes tipos de fallas. Algunas de estas provienen de los datos de entrenamiento ya que: a) contienen sesgos (y pueden fácilmente conducir a respuestas que discriminan, dañan y ofenden a diferentes grupos de personas), b) algunos tienen derechos de autor (puede generar respuestas que se apropien de contenidos con propiedad intelectual) y c) otras reflejan los datos a una determinada fecha. Esto se agrava si tomamos en consideración que estas aplicaciones no pueden reconocer cuándo una de sus respuestas es (o puede ser) errónea o potencialmente problemática (Baeza-Yates y Villoslada, 2022), ya que no hay comprensión acerca de las relaciones internas que el modelo aprendió y cómo impacta lo que se está generando (como se ha dicho el modelo se construye con estadística pura y dura). En particular, esto dificulta detectar errores o sesgos en las respuestas. Las fallas (y sus implicancias) pueden ser diferentes según el tipo de aplicación y el ámbito en que se utilice la IAG. Cuando se trata de generar un objeto visual o auditivo

(una pintura o una canción) las fallas se pueden manifestar en aspectos conceptuales de la obra, por ejemplo, DALL-E ha pintado humanos con seis dedos, sin que parezca una acción *ex profeso*, o también en utilizar técnicas e ideas que provienen de un autor particular (con implicancias en los derechos de propiedad como tal). En tanto, cuando se generan respuestas textuales a consultas de los usuarios, las fallas pueden derivar en otros problemas. Un ejemplo es la creación (y potencial distribución) de contenidos imprecisos, inexplicables e, incluso, inexistentes. Un pseudo-autor fácilmente puede indicar que se escriba una noticia acerca de un hecho falso generando una *fake news*. Si bien es responsabilidad de quien lo indica, la disponibilidad y poder de la herramienta colabora con la generación y distribución de contenido dañino en algún sentido. En ciertas ocasiones la falla en la generación de contenidos se puede manifestar sin que el usuario la pueda advertir. Algunos textos producidos artificialmente expresan descripciones inexistentes llamadas alucinaciones (Alkaissi y McFarlane, 2023). Estudios como el de Woody (2023) sitúan la tasa de alucinaciones entre un 15% y 20%.

En este sentido, surgen algunos interrogantes (Bordignon *et al.*, 2023), por ejemplo, si la respuesta infringe derechos de autor y un usuario lo utiliza en un trabajo, ¿quién es el responsable: el usuario o la compañía que provee el servicio de IAG? Más aún, si se usa un servicio pago (ya existen), ¿qué tipo de contrato relaciona a los usuarios con estos respecto de estas obligaciones? De manera similar, si se revela información sensible. Es decir, algún nivel de validez ética en un marco de responsabilidad debería ser exigido al contenido generado por un modelo de IAG, lo que va a representar un desafío en el camino de monetizar esta tecnología. Si el escenario es sensible, el costo de un fallo puede tener consecuencias muy negativas o impacto en la vida de las personas. Uno de los casos donde se propone el uso de IAG es el ámbito de la salud, más precisamente, un lugar de atención alejado de centros urbanos donde la IA se utilice con la idea de que ayude a mejorar un diagnóstico y/o planificar un tratamiento. ¿Cómo impacta un error

en un diagnóstico médico? Se podría tratar a alguien por algo que no tiene o no tratarlo porque se asume que no tiene una enfermedad. Se puede argumentar que el diagnóstico lo debe determinar el profesional médico, pero si su conclusión es desviada por un análisis incorrecto (realizado por un sistema de IAG) sobre un conjunto de datos, entonces la situación es compleja. Incluso si falla solo el 0,1 % de las veces (el número es arbitrario), sigue siendo un problema en este ámbito. Otra propuesta está dirigida a maestros para brindar formación personalizada a sus diferentes estudiantes usando la IA para planificar las lecciones de acuerdo al nivel y características de cada uno (Gates, 2023). Nuevamente, esta es otra situación en la cual un sistema de IAG podría ayudar, pero también complicar. ¿Qué sesgos pueden incluirse en el material personalizado a cada estudiante? O incluso, ¿qué nivel de desarrollo del pensamiento crítico necesitan estos estudiantes para poder validar la información provista, detectando, sesgos, errores y alucinaciones?

En resumen, no están claros los límites de los potenciales perjuicios que pueden venir asociados al uso de IAG por lo que se requiere que los sistemas sean explicables (puedan mostrar cómo arribaron a la respuesta), transparentes (se conozcan detalles, como los datos usados para su entrenamiento), imparciales e inclusivos (minimicen/eliminen los posibles sesgos). En este sentido, la comunidad científica ha definido el concepto de *IA responsable* que persigue que el uso de estas tecnologías no perjudique a las personas ni a la sociedad. En esta dirección, se reconoce que la intervención humana es necesaria tanto en la fase de entrenamiento como en la de prueba de los modelos (*human in the loop*) con el fin de producir mejores respuestas sobre la base de un ciclo de retroalimentación.

Las nuevas fábricas algorítmicas de noticias

Es evidente que con el desarrollo de la red internet y las plataformas –como nuevos espacios donde es posible realizar casi toda actividad– la manera de ejercer el periodismo, desde la producción y difusión de noticias ha experimentado (y así continúa) profundas transformaciones. Cambios que tienen que ver, centralmente, con el aumento sustancial del volumen y la accesibilidad de las fuentes, así como con la posibilidad inmediata de acceso a las mismas. Por otro lado, desde los usuarios, la forma en que a estos les llegan las piezas de información, los múltiples formatos multimedia, la posibilidad de ampliar y comprender una noticia desde diversas fuentes a la vez hacen que la experiencia del "consumidor" de piezas periodísticas se vea enriquecida de manera significativa. Ahora, este momento único también trae aparejado algunos problemas y desafíos por resolver. Así, por ejemplo, es urgente la mejora de la calidad de la producción de noticias, en especial la autenticidad y fiabilidad de los contenidos que se difunden. Estamos en un momento donde se están dejando atrás los métodos tradicionales de producción de noticias, los cuales dependen en gran medida de la mano de periodistas. El nuevo volumen de noticias operado ha permitido que los algoritmos tomen en una buena parte el rol de humanos. Una nueva etapa se está forjando en las redacciones, caracterizada por la convivencia de personas y máquinas en busca de una sinergia para la producción informativa.

La generación automática de noticias se sustenta en una serie de técnicas algorítmicas que, combinadas, permiten emular el estilo y la estructura del lenguaje humano. En particular, podemos citar: a) modelos de generación de texto: pueden generar automáticamente piezas de información a partir de elementos de noticias dados mediante un amplio entrenamiento que incluye un *corpus* de noticias representativo; b) tecnología de extracción de información: algoritmos que extraen automáticamente información clave de textos (estructurados o no estructurados) y proporcionan insumos de valor

para la posterior generación de noticias; y c) tecnología de generación de resúmenes: programas que pueden extraer información clave de contenidos noticiosos con el fin de generar resúmenes para consumo de lectores.

Ya hemos abordado el tema de los algoritmos que narran todo tipo de contenidos. Esto ha sido gracias, en gran parte, al desarrollo de tecnologías avanzadas de procesamiento del lenguaje natural y su vinculación estrecha con tecnologías de aprendizaje automático. Así, este maridaje tecnológico ofrece un camino interesante y por ahora viable y prometedor para abordar el proceso de producción de noticias dado que los nuevos algoritmos emulan las capacidades lingüísticas humanas para producir textos bien estructurados con una forma de narrativa pasable. Hoy, los nuevos procesos automáticos de generación de noticias se basan en algoritmos avanzados que procesan y analizan extensos conjuntos de datos, y así identifican y toman hechos relevantes con la finalidad de sintetizarlos dándoles forma de contenido periodístico.

Como vimos, en el sector del periodismo, se ha dado un importante salto en el uso de algoritmos. En muy poco tiempo, se ha pasado de tener agentes automatizados que controlaban la sintaxis, la gramática y el estilo a contar con una nueva generación de agentes que generan contenido informativo, a una velocidad sorprendente, de carácter objetivo e imparcial.

Pero, más allá de los avances mencionados aún persisten problemas, algunos nuevos, que se dan en torno a estas tecnologías. Si bien, como vimos se pueden usar para facilitar la tarea de los redactores y las agencias informativas, también estos modelos generativos pueden entrenarse para generar noticias falsas y hasta discursos de odio con fines de manipulación política de masas. Con lo cual ahora, a la vez, hemos construido fábricas algorítmicas de desinformación con una capacidad de producción y difusión nunca conocida antes.

Paralelo a este problema, han surgido una serie de herramientas que intentan lidiar con el mismo. En particular, se está automatizando

las tareas de comprobación de hechos para la detección de noticias falsas o contenidos engañosos, tal como se explicará en el capítulo 5. Estos sistemas también emplean técnicas de IA y de procesamiento del lenguaje natural para evaluar la veracidad de la información, pudiendo compararla con diversas fuentes fiables. Así, se está tratando de mitigar el problema de la desinformación automatizada dado que estas herramientas de "combate" desempeñan un rol central en el mantenimiento de la integridad de la oferta informativa destinada a la población en general.

Antecedentes en relación con la producción y verificación automática de contenido periodístico

Las técnicas de generación automática de textos derivan de los estudios sobre sistemas de generación de lenguaje natural y se centran en la producción de texto coherente a partir de datos no lingüísticos (Reiter y Dale, 2000). Así, las primeras experiencias abordaron el desarrollo de textos descriptivos sencillos, como son los informes técnicos. El área de investigación ha sufrido avances significativos con la aparición de la denominada IA de aprendizaje profundo. En particular, las redes neuronales recurrentes (Sutskever *et al*., 2014) dieron pie a un avance importante sobre la capacidad de procesar secuencias de textos más complejas. En esta deriva, hubo una segunda mejora importante en el área a partir de la emergencia del modelo Transformer (Vaswani *et al.,* 2017), que agregó aspectos básicos de reconocimiento semántico y de enfoque de la atención, lo cual permitió una comprensión y generación de texto de mejor calidad. La mejora fue tal que, cuando se hicieron públicos los servicios de chat, éstos sorprendieron a los usuarios por su emulación del decir humano.

De manera más específica, cuando la generación de textos se aplica a ambientes periodísticos, se reconocen los trabajos de Graefe *et al*. (2016). A partir del sistema Quakebot –generador de noticias sobre

terremotos–, dieron una prueba acerca del potencial de los sistemas algorítmicos de producción de noticias y su dinamismo, cuando la producción y difusión debe ser realizada en tiempos muy acotados. Una razón de peso que tiene la industria informativa para adoptar el uso de sistemas de la generación automática de noticias es su capacidad de aumentar significativamente la producción de contenido periodístico, ya que estos servicios operan de manera continua, como *fábricas algorítmicas de noticias* produciendo contenidos a una velocidad y volumen nunca visto antes. Las primeras adecuaciones de las redacciones a estos sistemas se realizaron en áreas temáticas de fácil procesamiento, como informes de inversiones y financieros, noticias deportivas e información meteorológica.

Otra de las áreas donde se está experimentando es la generación automática de noticias a demanda o a medida de un perfil de usuario determinado. La idea general es poder ofrecer servicios informativos donde los contenidos se adapten a intereses particulares y se distribuyen las piezas informativas en tiempo real. Mucha experiencia de esta última área está en relación con el funcionamiento de los algoritmos de redes sociales y su generación de filtros burbujas (Pariser, 2017).

Desde el área de verificación de noticias también ha habido un camino marcado por un importante desarrollo. Se reconocen los trabajos de Vlachos y Riedel (2014), quienes usaron bases de datos para experimentar técnicas de verificación de afirmaciones sobre textos. Lo mismo que sucedió con las técnicas de generación de noticias pasó con la verificación, se vio notablemente enriquecida por los últimos avances en IA. Las nuevas técnicas de aprendizaje automático permitieron mejorar los sistemas de comprobación de hechos a partir de análisis mucho más profundos y flexibles. Así, también la generación y la comprobación se integraron; por ejemplo, Zellers *et al.* (2019) introdujeron un modelo que no solo genera contenido textual sino que también comprueba la exactitud de los hechos. Estas herramientas integradas son una manera de luchar contra la falsedad y la desinformación desde los propios escritorios del periodismo. En

base a lo planteado, parecería ser que, de la misma manera que sucede en la medicina con las vacunas, el mismo componente de un virus es la solución para detenerlo. De este modo, se espera que las técnicas de IA también sean efectivas en la verificación de hechos.

Fábricas algorítmicas de desinformación y de discursos de odio

Durante este último tiempo hemos podido observar –y a la vez vivenciar– la capacidad de las nuevas técnicas de la IAG con el propósito de generar hechos periodísticos falsos a gran escala. Así, debemos contar, por un lado, que las fábricas algorítmicas de desinformación son parte de nuestro devenir cotidiano y, por otro lado, que la aplicación del pensamiento crítico ya es un arma obligatoria para su defensa y combate.

Hoy es posible construir un ambiente de generación de noticias falsas a bajo costo. Hay demasiados servicios gratuitos y rentados (en general económicos) que uno puede tomar y entrenar para poner a punto fábricas algorítmicas personales de noticias falsas. Esto recién comienza, hay muchos intereses económicos y políticos en juego que ven en estas tecnologías instrumentos óptimos (por el costo reducido y alcance masivo e instantáneo) para obrar sobre la manipulación de consumidores y la opinión pública. Ejemplos de esta última situación hay varios y cercanos. En las elecciones de Estados Unidos, 2016 y 2024, hubo una notable generación y difusión de información falsa, principalmente a través de redes sociales, que de alguna manera, en la del año 2016, habría afectado las percepciones de los votantes y por ende el resultado electoral (Allcott y Gentzkow, 2017). Otro caso de desinformación de gran magnitud sobre la población se dio en el referéndum del Brexit en el Reino Unido. Como resultado de tal campaña se verificó que hubo influencias sobre la opinión pública que contribuyó a la polarización del electorado (Levy *et al*., 2016).

Un estudio pionero a destacar, para detectar noticias falsas generadas por algoritmos, fue el llamado modelo Grover (Zellers *et al*., 2019), realizado en el año 2019 en la Universidad de Washington. La idea es que, para detectar el contenido desinformativo, primero hay que estudiar en profundidad cómo se genera. El modelo Grover se basó en una versión modificada de GPT-2 donde la siguiente palabra a predecir no solo se basa en las anteriores sino también en otros fragmentos de metadatos (por ejemplo, un título o un autor). Así, Grover puede incluir en sus producciones de noticias falsas elementos del contexto que ayuden a construir y "fortalecer" la mentira. Grover puede generar texto a partir de diversos datos que funcionan como "gérmenes" una noticia. Por ejemplo, se le puede suministrar un dominio, un autor, una fecha y un título, y el modelo generará el cuerpo de la (falsa) información. O se le puede darle un dominio, un autor, una fecha y un título, y puede generar un título. De esta manera, produciendo noticias falsas y compilando versiones de noticias verdaderas los investigadores pudieron entrenar un modelo que sea capaz de dictaminar si un texto es una noticia falsa o no.

Los discursos de odio como las noticias falsas siempre han existido, la diferencia con el presente, como se ha comentado, es el poder de amplificación instantánea y masiva que poseen las plataformas de redes sociales, lo cual hace que estos contenidos agresivos conformen un verdadero problema social propio de este tiempo. Un discurso de odio (DO, en adelante) según la Organización de las Naciones Unidas es:

> Cualquier tipo de comunicación verbal, escrita o conductual, que ataca o utiliza lenguaje peyorativo o discriminatorio con referencia a una persona o un grupo sobre la base de quiénes son. En otras palabras, sobre la base de su religión, etnia, nacionalidad, raza, color, ascendencia, género u otro factor de identidad. (ONU, sf)

Los DO más allá de ser insultos principalmente destinados a minorías o grupos vulnerados y marginados, para el Fondo de las Naciones Unidas para la Infancia (FNUI) son instrumentos de

manipulación dado que tratan de provocar respuestas emocionales sobre los agredidos, por ejemplo angustia, intimidación, miedo, aislamiento, entre otros. El efecto que tratan de promover los DO está relacionado con el establecimiento de un espacio o clima de intolerancia y odio que permita controlar y dirigir fuerzas sociales de ataque, vinculadas con la sociedad civil, sobre los destinatarios de la comunicación y así provocar conflictos que involucren prácticas agresivas y segregacionistas (LEDA, 2021).

La automatización de la generación de discursos de odio es un proceso semejante al de la creación de noticias falsas. Solamente hay que entrenar modelos de IAG con un *corpus* representativo de tales discursos para luego, una vez que el entrenamiento se haya realizado, poder a través de *prompts* específicos (de forma manual o automatizada) generar contenidos que sean verdaderos DO. Asociados al contenido de los DO están los "operarios" encargados de su difusión. En principio podemos ver dos tipos: a) los *haters*, usuarios de redes sociales que, utilizando DO, se dedican a insultar, degradar o discriminar personas o colectivos y b) los *trolls*, usuarios que actúan en masa hostigando y atacando a personas o colectivos. Estos dos tipos de usuarios actúan coordinados: los *haters*, como ideólogos y "capitanes" de los ataques y los *trolls* como el ejército que amplifica de forma acrítica y verticalista los mensajes de los *haters*. En muchas ocasiones los ataques suelen escaparse del ciberespacio y pasar al mundo tangible, así se transforman en agresiones físicas, escraches y hasta manifestaciones violentas.

Como hemos visto, el espacio web es una fuente ideal para el aprendizaje de servicios destinados a la desinformación. Los discursos de odio, las mentiras, los "escaparates" de sesgos están por doquier. Un caso interesante acerca de cómo un algoritmo puede "mal aprender" y comportarse de forma políticamente incorrecta fue el protagonizado por el chatbot "Tay" de Microsoft, aplicación creada para interactuar y aprender del comportamiento de los usuarios en la red social Twitter. Desde su puesta en funcionamiento, en solo 24 horas, el servicio comenzó a emitir mensajes controversiales, por ejemplo, racistas y

misóginos, como el resultado de tan pocas horas de interacción con usuarios humanos (Neff y Nagy, 2016). Como lección del hecho se puede indicar que los aprendizajes desviados se dan hoy por defecto en una web donde una buena parte de los contenidos fomentan desigualdades de todo tipo y se evidencian ciertas formas de discriminación. Esta situación de aprendizaje por defecto pudo ser observada en las primeras versiones del modelo de lenguaje GPT-3 de la empresa OpenAI que creaba contenidos con una amplia gama de sesgos (Bender *et al*., 2021).

Dentro de la amplia gama de contenidos falsos capaz de ser producidos con técnicas de IAG, se encuentran las llamadas *deepfakes*. Como se profundizará en el capítulo siguiente, consisten en vídeos y audios trucados que son prácticamente indistinguibles de grabaciones reales. En particular, estos contenidos se crean con la finalidad de instalar posiciones y discursos a partir de dichos de referentes o hechos controversiales, ya que tienen el poder de desacreditar a hechos y figuras al atribuirle situaciones o declaraciones que para nada han ocurrido. Para Kietzmann *et al.* (2020), estas producciones pueden producir implicaciones preocupantes para la desinformación, la privacidad y la confianza pública.

Contenidos realizados con técnicas de *deepfakes* se encuentran a diario. Un ejemplo que tuvo una amplia difusión fue el de Donald Trump siendo detenido por la policía. En 2023, se difundieron fotos falsas creadas con técnicas de IA donde se mostraba al político siendo arrestado por la policía por presuntamente ocultar documentos oficiales. La calidad del producto y su impacto fue importante, ya que se difundió rápidamente en redes sociales y hasta fue tomado por la prensa internacional. Otro caso tuvo por protagonista al presidente Zelenski donde se observa que solicita la rendición del ejército y el pueblo ucraniano. Este contenido audiovisual *deepfake* fue atribuido a *hackers* rusos. De forma rápida, el material fue desmentido dado su potencial de credibilidad.

A finales del año 2024 desde la plataforma X se presentó a Grok, un nuevo servicio tipo chatbot basado en técnicas de IA. Según sus

creadores, la diferencia del servicio está en que Grok puede responder sobre una amplia gama de temas que generalmente son evadidos por productos similares. Así, puede responder sobre temas del momento, tales como la política o asuntos religiosos, no limitándose a que la respuesta sea privada. Esta herramienta tiene capacidades poderosas para generar potenciales contenidos que abonen la desinformación, ya que sus servicios de texto e imágenes pueden llegar a vincular a cualquier persona con cualquier tipo de situación. ¿Hasta dónde estas herramientas colaboran con difuminar los límites entre la verdad y la mentira? ¿Estamos creando un nuevo mundo a la carta, donde las leyes de la vida, las matemáticas, la física, la historia, como en *1984*, podrían llegar a ser reescritas por cada ciudadano a su gusto, placer y necesidad?

La dimensión del espacio de desinformación

Para la organización Newsguard[9] el año 2023 fue el momento en que la IA potenció el desarrollo y distribución de noticias falsas tendentes a crear desinformación en los ciudadanos. Desde sus oficinas dedicadas al monitoreo de la red pudieron observar cómo las noticias falsas tuvieron un giro en su forma de redacción y se volvieron, además de más frecuentes, mejor redactadas, más persuasivas en su narración y por ende más peligrosas para los ciudadanos y las democracias del mundo.

Es interesante el enfoque técnico inicial que la organización Newsguard ha usado para probar cómo los modelos de lenguaje podrían, potencialmente, ser usados para generar contenidos falsos a partir de saltar los controles de la empresa. La técnica de prueba se llama *red teaming*, tiene su origen en la ciberseguridad, y consiste en una planificación y ejecución de una amplia gama de ataques simulados.

9 Newsguard (https://www.newsguardtech.com) es una organización especializada en evaluar la confiabilidad de contenidos en línea a partir de realizar estudios que rastrean y categorizan desinformación.

En el año 2023, los primeros resultados de los tests indicaron, usando una muestra aleatoria de 100 casos derivados de la base de datos de falsedades de NewsGuard (corpus Misinformation Fingerprints), que ChatGPT-4 generó 98 de los 100 mitos y el servicio Bard produjo 80 de los 100 totales (Brewster y Sadeghi, 2023).

Otro fenómeno asociado a las fábricas algorítmicas de desinformación son los crecientes portales generados por técnicas de IAG. Donde acá ya no es solo un contenido, sino un conjunto de piezas de desinformación orquestadas bajo el nombre de fuente maliciosa, que puede ser creada desde cero o emular alguna fuente creíble, de prestigio, y mediante trucos y engaños capturar a usuarios que no se dan cuenta del engaño. En el año 2014 la organización detectó más de 1000 de estos sitios web, donde muchos tenían nombres parecidos a los medios tradicionales: *iBusiness Day* o *Daily Time Update.* Dentro de los parámetros que hacen que se sospeche de un portal y pueda ser considerado como desinformativo, destacan: (1) la evidencia de que una parte importante de su contenido es generado por técnicas de IA, (2) si hay pruebas importantes que señalan que el contenido se publica sin una supervisión humana, (3) si el sitio se presenta de forma que lector puede asumir que su contenido es producido por periodistas, y (4) que el sitio no indique expresamente que su contenido es producido por IA.

Un estudio de opinión ciudadana sobre proliferación de noticias falsas fue realizado por la empresa Ipsos en el año 2023. A partir de una encuesta aplicada en 29 países se obtuvieron respuestas más de 20.000 adultos que, en principio, indican que la ciudadanía tiene claro (7 de cada 10 personas) que las nuevas técnicas de IA están facilitando la creación de noticias o imágenes falsas de una forma realista, difícil de evaluar (Ipsos, 2023). Por otro lado, a nivel global, 5 de cada 10 personas creen que los ciudadanos no son capaces de discernir entre información veraz y falsa. La política y los medios de comunicación son el caldo de cultivo de las noticias falsas. La percepción sobre la proliferación de noticias falsas en los últimos tiempos no deja lugar a dudas. Cinco de cada diez personas indican que existen más mentiras y

hechos de desinformación en la política y los medios de comunicación actualmente que hace 30 años.

Consideraciones finales, cuestiones abiertas

Las nuevas fábricas algorítmicas de desinformación y de discursos de odio, más allá de la manipulación y caos social que proponen, están trayendo, en paralelo a su desarrollo, otra serie de problemas. En este último apartado, pretendemos presentarlos y ponerlos en una mesa de discusión pública por la gravedad que presentan.

La primera cuestión tiene que ver con el abuso en la propiedad intelectual que se está produciendo por el uso de algoritmos de IAG. Cambian las técnicas, las herramientas de procesamiento, los servicios de automatización, pero el problema persiste y se agrava. La apropiación de los contenidos informativos de terceros por parte de las empresas de IAG aplicada a las noticias, es comparable a lo que sucedió hace más de dos décadas, cuando Google indexó todos los contenidos periodísticos existentes. Es decir, los tomó sin compensación alguna a nadie, ignorando cualquier derecho económico. Hoy esto se repite con la preparación de los *corpus* para el entrenamiento de los grandes modelos de lenguaje. Parece ser que el aprendizaje automático ingresa todo tipo de contenido, público, privado, reservado, limitado y con y sin *copyright*. Las referencias originales se pierden en el proceso y solo quedan las cuestiones matemáticas estadísticas asociadas a los textos. Ahora, en este devenir, nuevamente, un gran perdedor es la industria periodística y en particular los profesionales del género, cuyo trabajo es tomado por los algoritmos borrando cualquier referencia a fuentes y para nada reconociendo propiedad intelectual. Así, hoy es común encontrarse en portales con contenido periodístico artificial, construido con técnicas de IAG, donde se parafrasea y se reescriben noticias redactadas por periodistas sin reconocimiento de fuente ni retribución económica alguna. Por otro lado, cuando un algoritmo

narra, se pierde la idea de autor y, por ende, en géneros tan especiales como el periodismo esta situación es negativa para la profesión; dado que, por un lado, en quien confían los lectores ya no está, no existe más y, por otro, las cuestiones legales asociadas a todo relato informativo se encuentran frente a una pared donde la responsabilidad se diluye. Esto último es muy grave, es como si el algoritmo que narra tuviese licencia para decir cualquier cosa de cualquiera sin responsabilidad alguna.

El contenido artificial (no generado por humanos) disponible en el ciberespacio cada vez es mayor. A partir de la emergencia de los modelos generativos de la IA se está observando una aceleración mayor en su producción dado que las máquinas algorítmicas, que trabajan las 24 horas los 7 días de la semana, están emergiendo en una amplia variedad de rubros. Ahora, una pregunta está en relación con los nuevos grandes modelos de lenguaje que irán apareciendo. Si un modelo se alimenta en una buena parte de información disponible en el espacio web y la proporción de contenidos artificiales cada vez es mayor, ¿cómo afectará esto en los servicios de la IAG futura? ¿Estaremos ingresando en una época de *endogamia de saberes*? ¿Hasta dónde, en un corto tiempo, los contenidos sintéticos podrían comenzar a generar un cierto "ruido" en el *corpus* de saber de la humanidad?

Hasta el momento los grandes modelos se han alimentado principalmente de contenido generado por humanos, pero ¿qué sucederá cuando el contenido generado por la máquina alimente a nuevos modelos de lenguaje? Por un lado, son sabidas las limitaciones de comprensión, en semántica, de lógica y de razonamiento que tienen hoy los algoritmos usados y, por otro lado, están los errores que suelen tener las producciones de la IAG (por ejemplo, las alucinaciones). Esta combinación ya habla mucho de potenciales problemas que podrían suceder con las nuevas versiones de los grandes modelos de lenguaje. Se evidencia que hace falta un acuerdo general para señalar y marcar contenido sintético de todo tipo, no solo para los humanos, sino también para los agentes algorítmicos que los consuman.

Un último tema sobre el que reflexionar está en relación con el establecimiento potencial de estrategias que favorezcan nuevas formas de colonialismo basado en la manipulación y reescritura de nuestra historia y que conlleven a una visión y conformación de una cultura hegemónica, y por ende impuesta. Así, la pregunta pertinente sería: ¿qué posibilidades hay de que los contenidos artificiales narren nuevamente nuestra historia, pero con versiones programadas desde sesgos a medida? En este sentido quien planifica, entrena, supervisa y ajusta los grandes modelos de lenguaje tiene la posibilidad de establecer de antemano el comportamiento de los modelos ante las consultas de los usuarios. Y más allá de los sesgos de entrenamiento (por buena fe) y los errores que todavía tienen estás técnicas –que suelen causar respuestas erróneas y/o alucinaciones– está la posibilidad de un entrenamiento con *sesgos planificados* que colaboren a favor de manipulaciones, reescribiendo nuestra historia y nuestro presente. Estos nuevos *diarios de Yrigoyen algorítmicos*[10] podrían ser en parte lo que sucede en las cámaras de eco de las redes sociales pero llevado y compartido por un gran público.

Para finalizar solamente indicar, con cierta preocupación, que en enero del año 2025 Facebook se sumó a la campaña comenzada por Elon Musk con relación al despido de trabajadores que realizaban actividades de chequeo de hechos y cambio en la política verificación de noticias (Becares, 2025). Ahora la novedad tiene que ver con el establecimiento de una suerte de "democracia" o mayoría entre los usuarios acerca del "parecer" relacionado a un hecho. En el fondo, este cambio toma la opinión subjetiva de los colectivos frente a la verificación tradicional y objetiva, a partir de pruebas y de los hechos noticiosos. También esta nueva forma de "verificación" es una suerte de estrategia –a lo Poncio Pilatos– dado que ahora el poder decidir si una información es real o no cae bajo la responsabilidad de los usuarios de las plataformas de redes sociales.

10 En Argentina la expresión "diario de Yrigoyen" se utiliza para hacer referencia a una publicación que no es fiel a la realidad.

En resumen, a modo de cierre, podemos citar dos hechos graves actuales relacionados con el periodismo. Por un lado, la implementación de nuevos algoritmos de la IAG diluye la responsabilidad sobre la generación automática de noticias (donde ya observamos que algunas son falsas y tendenciosas), ya que el autor humano se evapora y se deja a los algoritmos que hagan periodismo (y para la ley es difícil su identificación y juicio) y, por otro lado, los dueños de las grandes plataformas de redes sociales están prescindiendo de servicios humanos de verificación objetiva de hechos noticiosos (cambiando la tarea por "juicios masivos subjetivos"). Ante el panorama descrito, es urgente una organización y acción por parte de los usuarios, la ciudadanía en general, dado que estos cambios lo único que pueden traer son atrasos a nuestra forma de ver y comprender el mundo en el que vivimos creando realidades paralelas en función de estrategias de manipulación y control social.

Por todo lo anterior, necesitamos, de forma urgente, que los distintos gobiernos de los países discutan, involucrando a sus ciudadanos, sobre estos temas y lleguen a acuerdos globales de control que permitan que toda esta tecnología se desarrolle en un marco ético y seguro en beneficio de la humanidad y no solo de los mercaderes algorítmicos de turno.

Referencias

Agamben, G. (2003). *Homo sacer.* Pre-Textos.

Alkaissi H. y McFarlane S. (2023). Artificial hallucinations in ChatGPT: Implications in scientific writing. Cureus.

Allcott, H y Gentkzkow, M (2017). Social Media and Fake News in the 2016 Election. Journal of Economic Perspectives 31, n.° 2, pp. 211-236.

Baeza-Yates R. y Villoslada P. (2022). Human vs. Artificial Intelligence. En 2022 IEEE 4th International Conference on Cognitive Machine Intelligence (CogMI) (pp. 40-48).

Becares, B. (7 enero 2025). Mark Zuckerberg imita a Elon Musk: se despide de los verificadores de Facebook e Instagram. Los acusa de ser "políticamente parciales". Blog Genbeta.
https://n9.cl/3vd5k2
Bender, E.; Gebru, T. McMillan-Major, A. y Shmitchell, S. (2021). On the Dangers of Stochastic Parrots: Can Language Models Be Too Big? Proceedings of the 2021 ACM conference on fairness, accountability, and transparency.
Bishop, J (2021). Artificial intelligence is stupid and causal reasoning will not fix it. Frontiers in Psychology 11:2603.
Bordignon, F., Dughera, L., & Tolosa, G. (2023). IAG y el momento de las máquinas imperfectas. Hipertextos, 11(19), 069. https://doi.org/10.24215/23143924e069
Brewster, J. y Sadeghi, M. (2023) Red-Teaming Finds OpenAI's ChatGPT and Google's Bard Still Spread Misinformation. Newsguard. https://acortar.link/afUjEK
Floridi, L. (2023) AI as Agency Without Intelligence: On ChatGPT, Large Language Models, and Other Generative Models (February 14, 2023). Philosophy and Technology, https://ssrn.com/abstract=4358789
Fondo de las Naciones Unidas para la Infancia. (s.f.). Cómo hablar con tus hijos e hijas sobre el discurso de odio. https://n9.cl/ulufu
Gates, B. (2023). The Age of AI has begun. Artificial intelligence is as revolutionary as mobile phones and the Internet. https://www.gatesnotes.com/The-Age-of-AI-Has-Begun
Graefe A, Haim M, Haarmann B, et al. (2016) Replication Data for: Readers' perception of computer-generated news: Credibility, expertise, and readability. Harvard Dataverse V1
Ipsos (2023) Global Views on A.I. and Disinformation. Perception of Disinformation Risks in the Age of Generative A.I. A 29-country global survey. https://acortar.link/N6jzPR
Kietzmann, J., L. W. Lee, I. P. McCarthy, and T. C.Kietzmann. 2020. Deepfakes: Trick or treat? BusinessHorizons 63 (2):135–46. doi: 10.1016/j.bushor.2019.11.006
Laboratorio de Estudios sobre Democracia y Autoritarismos. (2021). Informe LEDA #1. Discursos de odio en Argentina. https://n9.cl/6w2954
Levy, D., Aslan, B. y Bironzo, D. (2016). UK press coverage of the EU referendum (Report for the Reuters Institute for the Study of Journalism at the University of Oxford). Oxford: University of Oxford, pp. 1 - 45.
Neff, G. y Nagy, P. (2016) Talking to Bots: Symbiotic Agency and the Case of Tay. International Journal of Communication 10(2016), 4915–4931
ONU. (sf). What is hate speech. United Nations. https://www.un.org/en/hate-speech/understanding-hate-speech/what-is-hate-speech

Pariser, E., (2017). El filtro burbuja: Cómo la red decide lo que leemos y lo que pensamos. Barcelona: Taurus

Reiter, E. y Dale. R. (2000) Building natural language generation systems. Cambridge University Press, Cambridge

Sutskever, I.; Vinyals, O. y Le, Q. (2014) Sequence to sequence learning with neural networks. ArXiv preprint arXiv:1409.3215.

Vaswani, A.; Shazeer, N.; Parmar, N. Uszkoreit, J. Jones, L.; Gomez, A.; Kaiser, L. y Polosukhin, I. (2017) Attention Is All You Need (2017), NIPS'17: Proceedings of the 31st International Conference on Neural Information Processing Systems

Vlachos, A. y Riedel. S. (2014). Fact Checking: Task definition and dataset construction. En Proceedings of the ACL 2014 Workshop on Language Technologies and Computational Social Science, pp. 18–22, Baltimore, MD, USA. Association for Computational Linguistics.

Woody, A. (2023). Hallucinations, Plagiarism, and ChatGPT. Datanami. https://www.datanami.com/2023/01/17/hallucinations-plagiarism-and-chatgpt/

Zellers, R.; Holtzman, A.; Rashkin, H.; Bisk, Y.; Farhadi, A.; Roesner, F. y Choi, Y. (2019). Defending against neural fake news. Proceedings of the 33rd International Conference on Neural Information Processing Systems. Curran Associates Inc., Red Hook, NY, USA

4. Universo *deepfake*. Cartografía de la desinformación generada con IA

David García-Marín
Universidad Rey Juan Carlos
España

En el capítulo anterior, el profesor Fernando Bordignon realizó un interesante abordaje sobre la confluencia entre algoritmos y desinformación desde un punto de vista analítico a fin de explicar cómo estos sistemas, y en concreto la inteligencia artificial generativa (IAG), favorecen la producción de información falsa. En clara línea de continuidad temática, este capítulo –de carácter más aplicado– pretende servir como complemento a los planteamientos del profesor Bordignon, con el fin de conocer de forma específica las diferentes formas que toma la desinformación generada mediante estas fábricas algorítmicas de producción de falsedad.

La irrupción de la IA en los procesos de generación y propagación de la desinformación representa uno de los retos más complejos y trascendentales del ecosistema informativo contemporáneo. Estas tecnologías pueden actuar como un arma de doble filo: por un lado, constituyen una herramienta poderosa para automatizar tareas, personalizar contenidos y detectar información falsa; por otro, también se han convertido en un mecanismo eficaz para distorsionar la realidad,

manipular a la opinión pública y socavar los fundamentos de la democracia. Tecnologías como las *deepfakes*, los generadores de texto automatizado y las redes de *bots* representan un salto cualitativo en la capacidad de engaño, al ofrecer contenido que puede ser indistinguible del real incluso para usuarios informados.

La desinformación impulsada por IA no es un fenómeno aislado, sino global y adaptable a contextos sociopolíticos muy diversos, como demuestra su aplicación en diferentes países. En todos estos escenarios, el uso estratégico de la IA ha generado efectos concretos: erosión de la confianza en las instituciones, polarización del debate público, debilitamiento del periodismo y afectación directa a los procesos electorales. A su vez, se ha comprobado cómo los algoritmos de recomendación, diseñados para maximizar la interacción, han contribuido a viralizar contenidos falsos, reforzando cámaras de eco y burbujas informativas.

Desde una perspectiva crítica, también es necesario abordar el papel ambivalente que juegan los modelos de IA en este contexto: si bien pueden ser parte del problema, también forman parte de la solución. El desarrollo de herramientas automatizadas de *fact-checking*, sistemas de detección de patrones anómalos y tecnologías forenses para identificar *deepfakes* ofrece un camino esperanzador para mitigar el impacto de la desinformación, tal como se explicará en el siguiente capítulo. Sin embargo, estas soluciones técnicas no son suficientes por sí solas. Es necesario complementarlas con marcos regulatorios sólidos, principios éticos en el diseño de los algoritmos y políticas públicas centradas en la alfabetización mediática y digital de la ciudadanía.

Inteligencia artificial y narrativas desinformativas

En primer lugar, debemos centrarnos en el tipo de narrativas desinformativas más frecuentes que los creadores de desinformación generan cuando utilizan los sistemas algorítmicos de IA. Las primeras

investigaciones realizadas sobre este objeto de estudio determinan que los relatos falsos, normalmente basados en imágenes, que pretenden dañar o aprovecharse de la reputación de personajes relevantes o figuras públicas (especialmente actores políticos) son los más prominentes hasta la fecha. Se trata de imágenes, bien fotografías o vídeos (también audios, pero en menor porcentaje), donde aparecen figuras relevantes ejecutando acciones o emitiendo discursos que jamás realizaron ni pronunciaron. A este fin se consagran las denominadas *deepfakes* (ya anticipadas en anteriores capítulos), cuyo origen precisamente se encuentra en este tipo de narrativas: se considera que la primera *deepfake* de la historia consistió en un vídeo pornográfico falsamente protagonizado por la estrella de Hollywood Scarlett Johansson, cuyo rostro se le superpuso al cuerpo de una actriz de cine para adultos de modo que parecía que era la propia Johansson la protagonista de la cinta. Al estudio en profundidad de las *deepfakes* volveremos en la segunda parte de este capítulo.

Los primeros estudios sobre la desinformación generada con IA muestran también la abundancia de contenidos falsos generados con estos instrumentos sobre conflictos armados contemporáneos. Suelen ser imágenes que ofrecen una visión atroz del bando enemigo, minimizan las acciones propias y pretenden elevar la moral de la población y justificar la idoneidad del conflicto. Nada nuevo bajo el sol. La propaganda mediante medios visuales ha sido una constante en los conflictos armados, incluso anteriores al siglo XX (García-Marín y Salvat-Martinrey, 2023). A lo largo de la historia, la desinformación en contextos bélicos ha recurrido intensamente a los recursos visuales disponibles en cada época para cumplir sus objetivos. En esta línea, estudios como los de López Torán (2022) se enfocan en las tácticas de manipulación empleadas por el bando británico mediante el uso de la fotografía en el frente durante la guerra de Crimea (1853-1856). Por otro lado, algunas investigaciones han explorado el cine como herramienta propagandística, especialmente durante la Segunda Guerra Mundial,

con el propósito de "proporcionar a la opinión pública una visión distorsionada de lo que sucedía en los campos de batalla" (Díaz Benítez, 2013, p. 53).

También resulta especialmente preocupante el uso contemporáneo de la IAG para generar contenidos desinformativos cuyo objetivo es la creación de caos y desestabilización en una determinada comunidad a fin de atentar contra su seguridad. Frecuentemente vinculados con campañas de terceros países bien directas o a través de *proxies*[11], este tipo de contenidos elaborados con IA asumen dos posibles modalidades: (1) imágenes violentas habitualmente relacionadas con falsos ataques protagonizados por población migrante o racializada sobre ciudadanos autóctonos, o la falsa destrucción de lugares simbólicos para una determinada comunidad, y (2) vídeos con falsas recomendaciones de fármacos o tratamientos que, en caso de ser considerados ciertos, conllevarían un grave peligro para la salud pública. En ambos casos, la finalidad de estas narrativas creadas con IA es sembrar el caos en el país o comunidad objetivo, producir discursos de odio, potenciar la desconfianza de la población en las instituciones y obtener un beneficio estratégico a través de la división y la polarización ciudadana.

Sin embargo, las agendas políticas y bélicas no agotan las posibles narrativas desinformativas elaboradas con instrumentos de IA. La producción de imágenes con fines económicos también es especialmente prevalente. Suele darse mediante dos modalidades: (1) la falsa promoción de productos o servicios donde un falso experto o una figura relevante recomienda (falsamente) un producto existente en el mercado, y (2) los timos donde, igualmente, un falso especialista o un personaje popular en el contexto donde circula ese contenido desinformativo promueven la adquisición de un bien que no existe solo

11 De acuerdo con el Departamento de Seguridad Nacional de España, los *proxies* o fuentes *proxy* son actores interpuestos son entidades, organizaciones o individuos dentro de un Estado que actúan en interés de un actor extranjero. Pueden actuar como supuestos medios de comunicación, empresas de marketing y publicidad, organizaciones políticas, grupos de interés, funcionarios, o incluso figuras públicas e *influencers*. Diseminan mensajes o propaganda que beneficia a un gobierno o entidad extranjera, en contra de los intereses nacionales

con el objetivo de obtener un beneficio económico a través del robo de dinero o de la extracción de datos.

Las motivaciones económicas y/o promocionales también parecen estar detrás de la producción de imágenes curiosas o espectaculares sin una aparente intención política o ideológica. La democratización del uso y la creciente sencillez de utilización de los sistemas de IA para la producción de imágenes y audios facilita que prácticamente cualquier usuario pueda crear contenidos falsos altamente atractivos y potencialmente viralizables, con el único objetivo de potenciar la promoción y la interacción de sus cuentas en redes sociales con el consiguiente beneficio económico que ello podría reportarle a ese usuario si consigue una comunidad de seguidores ávidos de mensajes espectaculares o emocionales que se configuran, además, como los más fácilmente propagables en los circuitos digitales. Cuando tales creadores son profesionales de la producción visual (diseñadores, dibujantes, comunicadores audiovisuales, etc.), la intención promocional del uso de la IA para generar contenido falso resulta más que evidente.

La Tabla 1 sintetiza –con la adición de algunos ejemplos ilustrativos– la taxonomía de contenidos desinformativos generados con IA anteriormente explicada.

Tabla 1. Taxonomía del contenido desinformativo generado con IA en función de su narrativa

Narrativa	Definición	Ejemplo
Dañar / aprovechar la imagen de figuras relevantes	Se trata de imágenes, bien fotografías o vídeos (también audios, pero en menor escala), donde aparecen figuras relevantes ejecutando acciones o emitiendo discursos que jamás realizaron ni pronunciaron (*deepfakes*)	https://maldita.es/malditobulo/20220802/video-tiktok-princesa-leonor-montaje/
Imágenes sobre conflictos armados	Vídeos o fotografías que ofrecen una visión atroz del bando enemigo, minimizan las acciones propias y pretenden elevar la moral de la población y justificar el conflicto.	https://colombiacheck.com/chequeos/este-video-de-incendios-en-libano-tras-bombardeo-de-israel-mezcla-imagenes-creadas-con-ia
Creación de caos / desestabilización	Ataques de población migrante sobre ciudadanos autóctonos o destrucción de lugares simbólicos para una determinada comunidad. Vídeos con falsas recomendaciones de fármacos o tratamientos que conllevan un grave peligro para la salud pública.	https://maldita.es/malditobulo/20240816/imagen-policias-arrodillados-musulmanes-ia/ https://colombiacheck.com/chequeos/video-con-manuel-elkin-patarroyo-sobre-falsa-cura-para-la-hipertension-fue-creado-con-ia
Falsa promoción de productos / servicios	Un falso experto o una figura relevante recomienda (falsamente) un producto existente en el mercado.	https://colombiacheck.com/chequeos/circulan-en-facebook-videos-con-personajes-falsos-para-promocionar-un-medico-para-la
Timos	Un falso especialista o un personaje popular promueven el consumo de un bien que no existe a fin de obtener un beneficio económico.	https://maldita.es/timo/20240115/ana-blanco-proyecto-inversion-amancio-ortega/
Imágenes curiosas / espectaculares	Sin aparente intención política o ideológica, son contenidos falsos altamente atractivos y potencialmente viralizables	https://chequeado.com/ultimas-noticias/es-falsa-la-foto-del-submarino-titan-implosionado-en-el-fondo-del-oceano/

Fuente: elaboración propia

Ahora bien, ¿existen formatos o lenguajes mediáticos específicos para cada una de las narrativas anteriormente mencionadas? Los estudios preliminares realizados en el ámbito hispanoamericano por el equipo de investigación liderado por el autor de este capítulo

demuestran que todas aquellas narrativas relacionadas con cuestiones económicas (falsa promoción de productos y servicios y timos) se elaboran mayoritariamente en forma de vídeo. También es dominante el vídeo cuando la narrativa desinformativa se relaciona con la salud. Sin embargo, las imágenes curiosas o espectaculares se presentan en forma de fotografía, mientras que las narrativas que pretenden ocasionar un daño a personajes relevantes cuentan con un mayor equilibro, pero con un cierto dominio del vídeo. El uso del audio sintético generado con instrumentos de IA con fines desinformativos es aún ciertamente testimonial en comparación con los formatos visuales.

Como mencionábamos anteriormente, para el abono de estas narrativas falsas, los creadores de desinformación se suelen valer de un tipo de contenido que, por su verosimilitud y potencial manipulador, resulta especialmente preocupante y dañino para las sociedades democráticas: las *deepfakes*.

Deepfakes. Definición e impacto

Una *deepfake* ("ultrafalso" en castellano, según recomendación de Fundéu) es una "producción mediática (fotográfica, audiovisual o sonora) manipulada o generada desde cero mediante algoritmos de inteligencia artificial" (García-Marín, 2021, p. 105), donde habitualmente se presenta a un sujeto emitiendo un discurso que jamás emitió o realizando acciones que jamás ejecutó. Los usos más dañinos de este tipo de producciones "incluyen pornografía y noticias falsas, bulos y fraude financiero" (Foro contra las campañas de desinformación en el ámbito de la Seguridad Nacional, 2023, p. 195).

De acuerdo con el Departamento de Seguridad Nacional de España, las *deepfakes* pueden construirse mediante el uso de las siguientes técnicas algorítmicas:

1. Creación desde cero. Esta manipulación genera imágenes de rostros completamente inexistentes. Estas técnicas consiguen resultados

sorprendentes, generando imágenes faciales de gran calidad y realismo para el observador. Esta manipulación beneficia a muchos sectores, como la industria del videojuego o la del modelado 3D, pero también podría utilizarse para aplicaciones perjudiciales, como la creación de perfiles falsos muy realistas en redes sociales para generar desinformación.

2. Intercambio de identidad (cambio de rostro)[12]. Consiste en la sustitución de la cara de una persona en un vídeo por la cara de otra. A diferencia de la anterior, en ésta el objetivo habitual es generar vídeos falsos. Puede utilizarse con fines dañinos, como la creación de vídeos pornográficos protagonizados por personas famosas, bulos y fraudes financieros, entre muchos otros.
3. Manipulación de atributo. También conocida como *edición o retoque facial*, esta manipulación modifica algunos atributos de la cara como el color del pelo, de la piel, del sexo, o la edad. Un ejemplo de este tipo de manipulación es la popular aplicación móvil FaceApp, capaz de generar una representación "envejecida" de una persona automáticamente a partir de una foto actual de la misma. Muy útil en el ámbito comercial, los usuarios podrían utilizar esta tecnología para probarse una amplia gama de productos como cosméticos y maquillaje, gafas o peinados en un entorno virtual.
4. Sincronización labial. Consiste en la generación de vídeos falsos a partir de texto (*text-to-video*), tomándose como entrada un vídeo de un sujeto hablando y el texto que se desea pronunciar, y se sintetiza un nuevo vídeo en el que la boca del sujeto se sincroniza con las nuevas palabras (audio) generadas artificialmente[13].

A pesar de la aparente complejidad de estas técnicas, uno de los aspectos más preocupantes de las *deepfakes* radica en su creciente facilidad de elaboración:

---

12 Ver ejemplo en: https://www.youtube.com/watch?v=UlvoEW7l5rs

13 Ver en: https://www.ohadf.com/projects/text-based-editing/

> Se prevé que en pocos años cualquier persona con un *smartphone* pueda ser capaz de realizar este tipo de creaciones con una elevada calidad y a un coste mínimo. En 2015, cualquier producción de Hollywood necesitaba grandes cantidades de dinero y un equipo de expertos en efectos especiales para llevar a cabo trabajos de esta naturaleza. En 2020, ya existían *youtubers* capaces de realizarlos a coste cero y con resultados técnicos y estéticos muy similares a los profesionales. La película *El irlandés*, protagonizada por Robert De Niro y Al Pacino y dirigida por Martin Scorsese, narra una historia que abarca un total de siete décadas. Para hacer el relato creíble, Scorsese utilizó en 2015 parte de los 140 millones del presupuesto del film en contratar a un equipo de expertos en efectos especiales para modificar mediante procedimientos informáticos la cara de los actores a fin de adaptarla a cada etapa de la vida de los personajes. A pesar del dinero invertido, el resultado, a juicio de la crítica, no fue especialmente satisfactorio. Tres meses después del estreno de la película, en 2019, apareció un vídeo en Youtube titulado *The Irishman De-Aging: Netflix VS. Free Software.* Un *youtuber* anónimo había utilizado las nuevas herramientas de inteligencia artificial para realizar la misma tarea, con unos resultados excelentes. Este ejemplo nos muestra algunas de las características esenciales que tienen las *deepfakes*. En primer lugar, constituyen trabajos de una elevada calidad. La inteligencia artificial es capaz de crear efectos audiovisuales y sonoros mucho más creíbles que cualquiera de los estudios profesionales en toda la historia de la producción audiovisual. Por otro lado, esta tecnología provoca un efecto democratizador en la elaboración de contenidos audiovisuales, al convertirse en totalmente accesible para cualquier usuario a través de aplicaciones y software libre. Finalmente, a medida que esta tecnología evoluciona, la producción de este tipo de contenido resulta mucho más barata o incluso sin ningún tipo de coste. (García-Marín, 2021, p. 106-107)

Aunque estas técnicas algorítmicas se están refinando y paulatinamente ofrecen resultados más creíbles, lo cierto es que actualmente algunas de las *deepfakes* que circulan en los circuitos digitales pueden resultar fácilmente identificables teniendo en cuenta ciertos aspectos clave que ofrecen pistas para desconfiar de los vídeos generados mediante estos sistemas algorítmicos (ver Tabla 2).

Tabla 2. Claves para identificar *deepfakes* de vídeo

Clave visual	Definición
Desajuste del color de piel	El tono de piel entre la máscara y el rostro objetivo no coinciden. El rostro parece estar cubierto por una capa de colores. Se observan bordes o manchas.
Efecto de parpadeo	Hay un parpadeo entre el rostro original y el rostro deepfake, El algoritmo no puede reconocer la cara y deja de crear la máscara durante unos instantes.
Bordes visibles	Los bordes de la máscara son visibles: contornos nítidos o borrosos alrededor del rostro.
Perspectiva errónea	El deepfake tiene una diferente perspectiva al resto del vídeo. El vídeo fuente y destino difieren en longitud focal.
Oclusión facial	Cuando objetos pasan por delante de la cara, se distorsiona la máscara o la máscara cubre el objeto.
Contorno de perfil	El perfil de la cara no se ve bien. La máscara deepfake está rota, con menos detalle o mal alineada.
Rostro borroso	La máscara está borrosa. Hay una diferencia de nitidez o resolución entre la máscara y el resto del vídeo.
Signos de desajuste	Las expresiones del rostro deepfake no coinciden con el rostro objetivo. Los rasgos faciales no se comportan de manera natural y son invisibles, borrosos o salen repetidos.

Fuente: Carabias Álvaro (2023)

En todo caso, el potencial desinformativo y manipulador de este tipo de contenido resulta evidente. El informe denominado *Tackling deepfakes in European policy* elaborado en 2021 por el Parlamento Europeo documenta el impacto negativo de las *deepfakes* en tres ámbitos: (1) psicológico, (2) financiero y (3) social. El informe analiza los principales daños que pueden derivarse del uso malicioso de estos contenidos, a la vez que destaca su capacidad para erosionar la confianza pública, manipular la opinión, socavar instituciones democráticas y generar inestabilidad económica y social.

En primer término, las *deepfakes* pueden causar graves perjuicios psicológicos a las personas *target* falsamente representadas. Su uso en contextos de intimidación, difamación o pornografía no consensuada puede provocar traumas emocionales, pérdida de

reputación y oportunidades laborales, entre otros daños personales. En este fenómeno existe un evidente sesgo de género, que afecta desproporcionadamente a las mujeres, vinculándose con prácticas como la extorsión sexual, donde se amenaza con divulgar imágenes ímtimas falsas para obtener beneficios. Además, el conocimiento generalizado sobre la existencia de *deepfakes* fomenta fenómenos como el *enfriamiento social*, donde determinadas personas intentan evitar la exposición pública por temor a ser víctimas de manipulación.

Los riesgos financieros asociados a las *deepfakes* incluyen extorsión, fraude y robo de identidad. Existen casos documentados que muestran cómo se ha utilizado la clonación de voz para engañar a empleados y realizar transferencias bancarias fraudulentas. En marzo de 2019, el periódico *The Wall Street Journal* reportó un caso de fraude en el que una empresa británica del sector energético fue víctima de un robo de varios miles de euros. Los delincuentes emplearon IA para clonar la voz del director ejecutivo de la compañía. Con esta voz falsificada, realizaron una llamada telefónica a un empleado, solicitándole que transfiriera 250.000 euros a una cuenta que supuestamente pertenecía a un proveedor. Cuando los responsables de la empresa descubrieron el engaño, ya era demasiado tarde: los estafadores habían retirado el dinero y desaparecido con él.

Las *deepfakes* también se han empleado para suplantar identidades en procesos de verificación biométrica, comprometiendo la seguridad de organizaciones y personas. Asimismo, pueden ser utilizadas para manipular mercados financieros mediante la difusión de información falsa sobre empresas, fusiones o quiebras, afectando la estabilidad económica y la confianza de los inversores.

Desde el punto de vista social, los sectores más vulnerables a los efectos de las *deepfakes* incluyen el periodismo, la educación, la justicia y la ciencia. La manipulación de medios informativos puede facilitar campañas de desinformación, influir en la opinión pública y dificultar la labor ética de los periodistas. En el ámbito judicial, las *deepfakes* pueden ser utilizadas como pruebas falsas, comprometiendo

la integridad del sistema legal. En el ámbito científico, la generación de textos elaborados con instrumentos de IA generativa plantea riesgos de sesgos y desinformación, especialmente si los modelos de lenguaje no son entrenados con datos representativos y éticamente seleccionados.

Uno de los efectos más preocupantes de las *deepfakes* es la erosión generalizada de la confianza en la información. La dificultad para distinguir entre lo real y lo falso puede llevar a un estado de *apatía de la realidad*, donde las personas dejan de creer incluso en la verdad. Este fenómeno, conocido como el *dividendo del mentiroso*, favorece a quienes buscan manipular la verdad para obtener beneficios económicos, poder o impunidad.

Por este motivo, las *deepfakes* representan una amenaza directa a los procesos democráticos. Pueden distorsionar el debate público, manipular elecciones y dañar la reputación de figuras políticas. Su uso en campañas de microsegmentación[14] puede reforzar la polarización y las teorías conspirativas, debilitando la cohesión social y la legitimidad de las instituciones democráticas.

Finalmente, las *deepfakes* pueden ser utilizados para fomentar divisiones sociales, provocar disturbios civiles o incluso desencadenar conflictos internacionales. Un ejemplo notable es el caso del presidente de Gabón, donde un video sospechoso de ser una *deepfake* contribuyó a una crisis política y un intento de golpe de Estado en otoño de 2018[15]. Sin duda, la posibilidad de que gobiernos actúen basados en información manipulada representa un riesgo significativo para la paz y la seguridad global.

---

14 Según el Departamento de Seguridad Nacional de España, la microsegmentación consiste en identificar públicos específicos y personalizar el mensaje acorde a los datos que se han recogido de los usuarios. Con ella, el emisor consigue que la aceptación de los mensajes sea mayor y más efectiva. En estrategias de desinformación es empleada para identificar audiencias más vulnerables y aumentar la probabilidad de creencia sobre contenidos falsos, manipulados o medias verdades.

15 Ver en: https://www.motherjones.com/politics/2019/03/deepfake-gabon-ali-bongo/

Conclusión

Las *deepfakes*, como producciones que utilizan la IA con potencial desinformativo, constituyen una tecnología con un impacto disruptivo considerable. Si bien pueden tener aplicaciones legítimas como recursos comunicativos, artísticos o comerciales, su uso malicioso plantea desafíos urgentes que requieren respuestas coordinadas desde el ámbito legal, tecnológico y educativo. La prevención de sus efectos negativos pasa por el fortalecimiento de la alfabetización digital (Aparici y García-Marín, 20129), el desarrollo de herramientas de detección y la implementación de políticas públicas que protejan a las personas y a las instituciones.

La expansión de la IA ha introducido herramientas capaces de generar textos, imágenes, audios o vídeos de una calidad tal que, en muchos casos, resulta imposible distinguir lo real de lo artificial. Este fenómeno erosiona la confianza pública a través de la manipulación de las percepciones sociales o políticas. Cualquier persona con un equipamiento tecnológico mínimo y escasos conocimientos puede generar este tipo de contenido, lo que democratiza la posibilidad de utilizar la IA con fines maliciosos.

Tal como advierte el Informe sobre Riesgos Globales del Foro Económico Mundial (2024), la propagación de desinformación impulsada por IA representa una de las amenazas más graves de los próximos años, especialmente por su potencial para alterar procesos democráticos, alimentar el malestar social y fomentar la violencia mediante la difusión de contenido falso y especialmente destinado a la división y polarización social.

Pese a este escenario inquietante, la IA también ofrece soluciones para combatir estas mismas amenazas. Los sistemas más avanzados permiten analizar patrones lingüísticos, contextos y fuentes para detectar errores, noticias falsas y contenidos maliciosos. Iniciativas como la Coalición para la Autenticidad y Procedencia del Contenido (C2PA) o la Coalición Mundial para la Seguridad Digital, impulsada

por el Foro Económico Mundial, buscan establecer normas claras y colaboraciones entre gobiernos, empresas tecnológicas y sociedad civil para hacer frente a esta problemática. Además de las herramientas tecnológicas, resulta fundamental el fomento de la alfabetización mediática desde espacios educativos y comunitarios, como bibliotecas y escuelas, para dotar a la ciudadanía de capacidades críticas frente a la ola de información sintética (Li y Callegari, 2024).

El desafío que plantea la desinformación mediada por IA no es meramente tecnológico, sino profundamente social, político y ético. Requiere de un enfoque multidisciplinar que combine la innovación tecnológica con la responsabilidad institucional, el compromiso ciudadano y una reflexión crítica constante. Solo así será posible construir un entorno comunicativo más transparente, plural y resistente a la manipulación, donde la IA esté verdaderamente al servicio de la verdad y del bien común.

En este contexto, la colaboración global y el desarrollo ético de la IA se presentan como vías esenciales para preservar la integridad informativa en la era digital y, con ella, defender nuestras democracias. Para lograr este objetivo, resulta fundamental dotarnos de un fuerte sistema de medios de comunicación –independientes y plurales– dedicados a la verificación del contenido online y del discurso político, tanto desde el ámbito periodístico como desde una óptica general que atañe a toda la ciudadanía. Precisamente, el estudio de la verificación de contenidos y el *fact-checking* como abordajes esenciales contra el fenómeno de la desinformación centra la atención del siguiente capítulo.

Agradecimientos

Este trabajo forma parte del proyecto “Desafíos, usos y limitaciones de la IA en el *fact-checking* y la lucha contra la desinformación” (DESAF_IA) (Ref. 2024/SOLCON-135623) financiado por la convocatoria de

Proyectos IMPULSO a la investigación de la Universidad Rey Juan Carlos (2024).

Referencias

Aparici, R., & García-Marín, D. (2025). *La posverdad. Una cartografía de los medios, las redes y la política.* Gedisa.

Carabias Álvaro, A. (Coord.) (2023). *Guía práctica Fake News.* Fundación Telefónica.

Díaz Benítez, J.J. (2013). Propaganda bélica en la gran pantalla: la incursión de Makin (1942) a través de la película Gung Ho! *Historia Actual Online*, *31*, 53-63. https://cutt.ly/k0VL0r5

Foro contra las campañas de desinformación en el ámbito de la Seguridad Nacional (2023). *Trabajos 2023.* Ministerio de la Presidencia, Relaciones con las Cortes y Memoria Democrática, Gobierno de España.

García-Marín, D. (2021). El *whatsapp* de Odiseo. Potencial desinformativo y estrategias re- tóricas del audio fake. En: Elías, C. y Teira, D. (Coords.), *Manual de periodismo y verificación de noticias en la era de las fake news.* UNED, pp. 99-132.

García-Marín, D. & Salvat-Martinrey, G. (2023). Desinformación y guerra. Verificación de las imágenes falsas sobre el conflicto ruso-ucraniano. *Revista ICONO 14. Revista científica de Comunicación y Tecnologías emergentes*, *21*(1). https://doi.org/10.7195/ri14.v21i1.1943

Li, C., & Callegari, A. (2024). *Cómo combatir la desinformación de la IA y proteger la verdad en el mundo digital*. Foro Económico Mundial.

López Torán, J.M. (2022). Y la guerra entró en los hogares: noventa años de propaganda y fotografía bélica (1855-1945). *Historia & Guerra*, 2, 17-43. https://doi.org/10.34096/hyg.n2.1 1061

5. Verificación digital y *fact-checking*. Teoría, método y herramientas

David García-Marín
Universidad Rey Juan Carlos
España

En la era de la sobreabundancia informativa y la popularización de la IA generativa, el fenómeno de la desinformación se ha consolidado como uno de los principales retos para las democracias contemporáneas. A pesar de que el acceso a la información es más amplio, rápido y diverso que nunca, esta expansión no ha traído consigo una mejora proporcional en la calidad del conocimiento público. Por el contrario, ha generado un entorno propicio para la proliferación de contenidos falsos, manipulados o engañosos que erosionan la confianza en las instituciones y los medios de comunicación, polarizan a la sociedad y debilitan los sistemas democráticos. En este contexto, la verificación digital y el *fact-checking* emergen como herramientas esenciales para contrarrestar los efectos nocivos de la desinformación.

Como ya se ha señalado previamente en esta obra, el auge de la desinformación no es un fenómeno nuevo, pero sí ha adquirido una dimensión inédita en las últimas décadas debido a la convergencia de varios factores. Entre ellos destacan la crisis económica global de 2008, que debilitó el modelo de negocio de los medios tradicionales,

la irrupción de las redes sociales como principal canal de acceso a la información, y el desarrollo de tecnologías que facilitan la creación y difusión de contenidos falsos. Esta transformación del ecosistema informativo ha desplazado el protagonismo desde los medios convencionales hacia las plataformas digitales, donde la lógica algorítmica y la viralización priman sobre la veracidad y el rigor.

En este punto, hemos de diferenciar dos conceptos que en ocasiones resultan confusos: verificación digital y *fact-checking.* La verificación digital se define como el conjunto de procesos destinados a comprobar la veracidad de una información publicada, ya sea en medios de comunicación o, sobre todo, en redes sociales. Su objetivo es detectar manipulaciones, falsificaciones o descontextualizaciones en contenido multimedia digital. Utiliza herramientas tecnológicas como la búsqueda inversa de imágenes, el análisis de metadatos, la geolocalización, el análisis forense digital, etc. Ejemplos tradicionales de verificación digital consisten en comprobar si una imagen viral fue tomada en el lugar y fecha que se afirma, o si un vídeo ha sido editado y, por tanto, la narrativa que porta es falsa.

Por su parte, el *fact-checking* consiste en la comprobación de la veracidad de afirmaciones, datos o declaraciones generalmente realizadas por figuras públicas o medios de comunicación. Su objetivo es determinar si una afirmación es verdadera, falsa, engañosa o imprecisa en función de la "escala de la verdad" que cada verificador implementa en su patrón de trabajo. Se basa en la consulta de fuentes confiables, documentos oficiales, expertos, bases de datos, etc. Como ejemplo paradigmático del *fact-checking*, podemos destacar el hecho de verificar si un político miente al realizar una afirmación o si una estadística citada en un artículo periodístico es correcta.

La verificación del contenido y/o declaraciones públicas, que históricamente formaban parte del trabajo periodístico previo a la publicación, ha evolucionado hacia una labor especializada y posterior a la difusión del contenido, que es desarrollada por las agencias de verificación o *fact-checkers.*

Periodismo contra la desinformación: los *fact-checkers*

Los *fact-checkers*, también denominados verificadores o agencias de verificación, son medios de comunicación dedicados mayoritariamente a la identificación del contenido falso emitido por terceros; así como de la verificación del discurso de las figuras públicas, especialmente los representantes políticos, mediante prácticas "basadas en la comprobación de datos con herramientas actuales, con la informática y las tecnologías como elementos básicos en su marco de actuación" (Vázquez-Herrero *et al*., 2019, p. 3).

La actividad de las instituciones dedicadas al *fact-checking* se remonta a finales del siglo XX en Estados Unidos. La primera organización especializada en esta labor fue *Snopes.com*, un sitio web fundado en 1995 con el propósito de verificar rumores, bulos y falacias. Esta plataforma adquirió gran notoriedad tras los atentados del 11 de septiembre de 2021, convirtiéndose en el principal referente estadounidense en verificación de información (Aspray y Cortada, 2019). En los últimos años, este tipo de iniciativas ha experimentado una expansión global. Algunas operan de forma independiente, como *Maldita.es*, *Newtral* o *Verificat* en España, mientras que otras funcionan como divisiones dentro de medios de comunicación más amplios, como *VerificaRTVE*, *EFE Verifica* o *AFP Factual.*

Los profesionales que integran estas entidades provienen de diversas disciplinas, incluyendo el periodismo, la ciencia de datos, la informática y la ciencia política. Su producción se centra principalmente en cuatro tipos de contenidos: (1) verificaciones o desmentidos que buscan demostrar la falsedad de afirmaciones emitidas por representantes públicos o personas de gran relevancia (*fact-checking*), (2) artículos de *debunking*, definidos como la verificación de un contenido que circula online y que ha sido viralizado en una o más plataformas sociales, medios de comunicación u otros canales, (3) artículos de *prebunking* creados como una forma de *inoculación* contra la desinformación,

con el objetivo de explicar a la ciudadanía cómo identificar posibles casos de desinformación antes de que los encuentren, y (4) informes que analizan, investigan o informan en profundidad sobre tendencias, narrativas o desarrollos relacionados con la desinformación.

La labor de estas organizaciones se sustenta en tres pilares fundamentales: el conocimiento empírico basado en hechos comprobados, el contexto en el que se desarrollan dichos hechos (Antonakaki *et al*., 2021) y el análisis del contenido a verificar, evaluando su veracidad a partir de sus características textuales o audiovisuales (Potthast *et al*., 2017; Oshikawa *et al*., 2020).

Aunque la verificación de contenidos es una práctica inherente al periodismo, el *fact-checking* presenta rasgos que lo diferencian del periodismo tradicional. Los verificadores se consideran una categoría particular dentro del ámbito periodístico, con normas, rutinas y principios propios y diferenciados (Graves, 2016). En primer lugar, su producción es esencialmente reactiva, ya que responde a la circulación de desinformación, especialmente en redes sociales, o a declaraciones inexactas emitidas por figuras públicas. Es decir, su acción constituye una reacción a posteriori del contenido elaborado por terceros. Por ello, rara vez marcan la agenda informativa, ya que su intervención ocurre tras la difusión de la falsedad. Salvo en el caso de las piezas explicativas, no generan contenido original, sino que se dedican a evaluar la veracidad de información ya publicada.

En este sentido, los *fact-checkers* desempeñan una función poco habitual del periodista como vigilante de otros medios, al trabajar sobre hechos previamente mediatizados (Graves, 2016). Su labor puede interpretarse como una corrección a los medios que han difundido contenidos asumidos como verídicos, constituyendo así una crítica explícita al periodismo convencional. Para lograr legitimidad ante la audiencia, estas entidades deben mantener una postura neutral y apartidista durante todo el proceso de verificación, además de garantizar transparencia respecto a su propiedad y fuentes de financiación (García-Marín *et al*., 2023). Estos requisitos son esenciales

para su integración en redes internacionales como la International Fact-Checking Network (IFCN).

Los verificadores suelen formar parte de consorcios internacionales que promueven la colaboración y el intercambio de herramientas tecnológicas avanzadas, como aquellas basadas en inteligencia artificial. Estas dinámicas permiten que entidades con menos recursos accedan a tecnologías desarrolladas por organizaciones más consolidadas (Sánchez-González *et al*., 2022). Además, los *fact-checkers* practican un periodismo en red, enlazando sus verificaciones a fuentes externas, fomentando la citación de su trabajo por parte de otros periodistas y estableciendo acuerdos de distribución y colaboración con medios de gran alcance y con el ámbito académico y científico. Su autoridad e impacto dependen, en parte, de estas conexiones.

Las diferencias entre el *fact-checking* y el periodismo tradicional también se reflejan en las rutinas productivas y la estructura y metodología de los informes de verificación. El proceso de verificación suele seguir una metodología común: (1) selección de contenidos sospechosos, (2) análisis de su veracidad mediante herramientas digitales y fuentes fiables, y (3) publicación de los resultados. Este trabajo se ve reforzado por la colaboración ciudadana, que permite a los usuarios alertar sobre posibles bulos, y por el uso de tecnologías avanzadas, como la inteligencia artificial, que facilita el rastreo automatizado y el análisis masivo de datos. No obstante, la verificación también enfrenta importantes desafíos, como la imposibilidad de abarcar todo el volumen de información circulante, la dificultad de acceder a fuentes fiables o la percepción de parcialidad por parte de ciertos sectores sociales. En este sentido, uno de los aspectos más debatidos en torno a la verificación es su eficacia real en la lucha contra la desinformación.

En este sentido, la literatura científica revela que la actuación de los verificadores tiene un efecto positivo y significativo en la reducción de la creencia en la desinformación (van Erkel *et al.*, 2021). Sin embargo, su efectividad puede variar según la fuente utilizada en el chequeo

y la credibilidad percibida de dicha fuente (Liu *et al.,* 2023). Otros estudios han demostrado que la repetición de mensajes verificando un contenido desinformativo puede aumentar su impacto (Pennycook *et al.,* 2018), así como la importancia de la transparencia en los procesos de verificación para mantener y potenciar la credibilidad de los *fact-checkers* (Brandtzaeg *et al.,* 2018).

En el contexto español, las alianzas de *fact-checking*, como la iniciativa *Comprobado*, han demostrado ser efectivas en la verificación colaborativa durante periodos electorales. En esta línea, Cuartielles y Carral (2024) examinaron la viabilidad y el funcionamiento de estas alianzas, destacando la importancia de la coordinación y el uso de materiales compartidos para mejorar su eficiencia. Estas alianzas permiten una respuesta más rápida y coordinada a la desinformación, especialmente en contextos de alta polarización política. Por tanto, la colaboración entre diferentes agencias puede aumentar la capacidad de respuesta y la cobertura de las verificaciones.

Por otro lado, el contexto en el que se realiza el *fact-checking* también puede influir en su efectividad. Un estudio experimental en 16 países europeos subraya la importancia de considerar el contexto político y cultural al diseñar estrategias de verificación (van Erkel *et al.,* 2021). Además, la polarización política puede influir en la receptividad de los mensajes de los *fact-checkers,* con individuos más inclinados a aceptar verificaciones que confirmen sus creencias preexistentes (Nyhan y Reifler, 2010).

El proceso de verificación

La actividad de los verificadores se establece en diferentes fases: (1) identificación y selección del contenido a chequear, (2) aplicación del método de verificación, que concluye con un veredicto o conclusión de acuerdo con una "escala de la verdad" específica para cada *fact-checker* y (3) redactado y publicación del verificado en la web del *fact-checker* (y

también en sus redes sociales) a partir de una estructura ciertamente estandarizada. Se describe a continuación cada una de estas fases (García-Marín, 2024).

Selección

Todo proceso de verificación se inicia con la identificación y selección del contenido que se presume desinformativo. Esta etapa inicial reviste una especial sensibilidad, ya que una posible parcialidad en la elección del material a verificar —por ejemplo, motivada por sesgos ideológicos o políticos— puede comprometer seriamente la credibilidad y legitimidad del verificador ante la opinión pública. Al igual que en el periodismo tradicional, los *fact-checkers* aplican criterios de noticiabilidad y relevancia, considerando el impacto potencial que la desinformación puede generar en la sociedad. Dado el volumen de información falsa que circula en el entorno mediático contemporáneo, resulta inviable verificar la totalidad de los contenidos, por lo que se priorizan aquellos que presentan mayor riesgo de daño o confusión, así como los contenidos más ampliamente propagados.

Un criterio esencial en esta fase es la verificabilidad del contenido. En muchos casos, no es posible establecer con certeza la veracidad de determinadas afirmaciones, ya sea por su naturaleza ambigua o por la falta de evidencia empírica. Además, el carácter dinámico de la información implica que una afirmación puede ser falsa en un momento determinado y volverse verdadera posteriormente, lo que obliga a contextualizar la verificación en el marco temporal en que fue emitida. En consecuencia, quedan excluidos del ámbito de actuación de los verificadores ciertos tipos de contenidos. Entre ellos se encuentran las declaraciones basadas en opiniones personales o juicios de valor de índole estética, moral o ética; las afirmaciones complejas constituidas por varios hechos, que requieren ser descompuestas en unidades de significado verificables; las declaraciones descontextualizadas, cuya veracidad depende del momento y circunstancias en que

fueron pronunciadas; las proyecciones futuras o metas propuestas; las expresiones vagas, ambiguas o poco concretas; las afirmaciones sustentadas en creencias religiosas; las experiencias personales; y aquellas formuladas mediante recursos retóricos como hipérboles o metáforas.

Aplicación del método de verificación

Una vez seleccionado el contenido a verificar, se inicia el proceso de verificación propiamente dicho, el cual, en su formulación ideal, debería sustentarse en el método científico y en la aplicación de procedimientos sistemáticos. El punto de partida consiste en formular una hipótesis que presupone que el hecho o la declaración objeto de análisis es falsa, engañosa o inexacta. A partir de esta premisa, se procede a la recolección de evidencias durante la fase de chequeo, con el objetivo de confirmar o refutar dicha hipótesis.

Durante esta etapa, los *fact-checkers* combinan el uso de herramientas digitales avanzadas (ver anexo de este capítulo) con métodos tradicionales del periodismo, como la consulta directa a las fuentes implicadas o afectadas por la declaración en cuestión. Las fuentes utilizadas en este proceso pueden incluir documentos oficiales como encuestas, estudios, informes o actas institucionales; personas o entidades mencionadas en la narrativa desinformativa; emisores del contenido falso, como partidos políticos, figuras públicas o instituciones; expertos reconocidos en la materia objeto de verificación; y archivos de medios de comunicación, incluyendo las propias hemerotecas de los verificadores, que permiten contrastar declaraciones con antecedentes informativos similares.

La utilización de herramientas digitales y la consulta a fuentes se combina con otras estrategias, también propias del periodismo de verificación clásico. Los *fact-checkers* (así como todo ciudadano con espíritu crítico) debe prestar especial atención a los siguientes aspectos, que pueden ofrecer claves decisivas para detectar contenidos

desinformativos a fin de diferenciarlos de la información legítima y verídica (Fernández Barrero y Aramburú Moncada, 2024):

1. Dificultad para identificar la autoría y/o procedencia de la información.
2. Uso de titulares llamativos y con una fuerte carga emocional.
3. Dificultad para identificar (o ausencia de) las fuentes en las que se basa la información.
4. Excesivo énfasis en la exclusividad de los contenidos. El periodismo de calidad no remarca en el texto de forma reiterada que una información ha sido publicada en exclusiva por ese medio.
5. Uso de formatos llamativos (colores, tamaños y un uso desmedido de la mayúscula).
6. Utilización de contextos confusos. Las noticias en los medios digitales de calidad deben llevar fecha y hora de cierre y fecha de actualización, pues las actualizaciones conllevan a veces la inserción de nuevos datos e información más precisa acerca de los hechos.
7. Introducción en el relato de elementos valorativos.
8. Presencia de errores ortográficos y gramaticales.

La labor de verificación implica también una mirada crítica hacia el contenido a chequear. Algunas preguntas que todo *fact-checker* –y todo ciudadano crítico– debe hacerse para demostrar la falsedad de una información son (Mateos, 2021):

- ¿Parece increíble? ¿A quién le interesa o beneficiaría que esto fuera cierto?
- ¿Apela a la razón? ¿Plantea explicaciones más que juicios? ¿Relaciona las afirmaciones con datos contrastados y con fuentes autorizadas y diversas?
- ¿Apela a las emociones? ¿Juzga, condena, califica, descalifica, plantea diatribas morales?
- ¿Qué tipo de léxico emplea? ¿Es opinativo, provocativo, incendiario, con exageraciones, sensacionalista? ¿Es lenguaje

común o científico y especializado? ¿Hace una mezcla de gran contraste intercalando términos científicos altisonantes con un lenguaje muy coloquial?
- En el caso de las imágenes: ¿Qué información de contexto puedo ver (personajes, lugares, etc.)?

Es habitual que los verificadores triangulen la información obtenida mediante todas estas estrategias y evidencias, con el fin de dotar de mayor solidez y credibilidad al veredicto final. Este veredicto se expresa a través de una escala de veracidad que clasifica el contenido verificado desde el máximo grado de verdad hasta la falsedad absoluta. Por ejemplo, el *fact-checker* español *Newtral* aplica diferentes categorías según el tipo de contenido: en el caso de declaraciones de figuras públicas o representantes políticos, utiliza etiquetas como "verdad", "verdad a medias", "engañoso" y "falso"; mientras que los bulos difundidos en redes sociales se categorizan únicamente como "engañosos" o "falsos".

Redacción del verificado

Una vez finalizado el proceso de verificación y recopilada toda la información relevante, los *fact-checkers* proceden a la publicación de los resultados de su investigación. Esta difusión se realiza principalmente a través de sus sitios web, aunque también se recurre a las redes sociales, donde el contenido se adapta a las particularidades semióticas de cada plataforma. La estructura y redacción de los informes de verificación en estos espacios digitales tiende a seguir un formato estandarizado, diseñado para facilitar la comprensión y reforzar la credibilidad del mensaje.

El informe se inicia con un titular que comunica de forma directa el veredicto alcanzado (García-Marín, 2024). Esta estrategia responde al modelo conocido como *sándwich de la verdad*, que consiste en enmarcar el contenido desinformativo entre dos afirmaciones veraces: se comienza con la conclusión del chequeo, se desarrolla la explicación

del contenido falso y se cierra reiterando el veredicto. Esta técnica busca minimizar el impacto del bulo al presentarlo subordinado a un marco de verdad desde el inicio.

Seguidamente, se incorpora un subtítulo que amplía la información del titular, explicando de manera concisa las razones que justifican la categorización del contenido como verdadero, falso o engañoso (Imagen 1). Junto a estos elementos, suele añadirse una caja destacada que resume el resultado de la verificación mediante una etiqueta correspondiente a la escala de veracidad empleada por la entidad verificadora.

Imagen 1. Título y subtítulo del verificado

Fuente: https://bit.ly/3TIS6Mi

Entre el subtítulo y el cuerpo del texto, se incluye habitualmente una imagen del contenido desinformativo, acompañada de una marca de agua o una etiqueta como "falso", "bulo" o "*fake*", con el objetivo de evitar su reproducción acrítica y contribuir a su desacreditación visual (Imagen 2).

Imagen 2. Bulo con la etiqueta "*fake*"

Fuente: https://bit.ly/3TIS6Mi

Los informes deben estar fechados y firmados. La firma responde a criterios de transparencia, permitiendo identificar al periodista responsable del chequeo, mientras que la fecha contextualiza temporalmente la verificación, lo cual es crucial para distinguir entre contenidos desinformativos recurrentes y bulos de nueva creación.

El cuerpo del texto comienza con un párrafo introductorio que presenta las principales coordenadas del contenido verificado, respondiendo a las preguntas clave del periodismo: qué dice la desinformación, dónde ha circulado y cómo ha sido elaborada (por ejemplo, si se trata de un vídeo, una imagen o un texto). En algunos casos, también se incluye el cuándo, aunque esta información puede omitirse si no es posible determinarla con precisión o si se considera irrelevante por la inmediatez de la detección del bulo. Este primer párrafo suele concluir reiterando el veredicto y clasificando el tipo de desinformación como contenido manipulado, fuera de contexto, suplantación, etc. (Imagen 3).

Imagen 3. Ejemplo de primer párrafo de un verificado

Por María G. Dionis

02 febrero 2024 | 2 min lectura

Bulos Migrantes

Circulan mensajes en Facebook, X y TikTok que aseguran que un comedor social de Cáritas en Almería ha denunciado que personas marroquíes "tiran parte de la comida que reciben". Algunas de las afirmaciones están acompañadas de un vídeo que muestra a varias mujeres junto a un contenedor, donde supuestamente habrían tirado comida. Pero son **mensajes falsos**, según ha desmentido Cáritas, y **el vídeo no es actual**.

Fuente: https://bit.ly/3TIS6Mi

A continuación, el desarrollo del cuerpo del informe se dedica a explicar las técnicas, herramientas y fuentes utilizadas en el proceso de verificación. Esta sección tiene un carácter metaperiodístico, ya que detalla el procedimiento seguido para alcanzar las conclusiones expuestas. Además, se proporciona contexto adicional sobre la narrativa verificada, incluyendo antecedentes similares y su circulación en otros entornos geográficos. Esta parte suele estar acompañada de pruebas documentales, como vídeos o imágenes, tanto del contenido falso como del verdadero, o incluso comparativas entre ambos, con el objetivo de facilitar al lector la identificación de las diferencias.

Finalmente, el informe concluye reiterando el veredicto, en coherencia con la estructura del *sándwich de la verdad*, y presenta una recopilación de las fuentes consultadas y una explicación de la metodología empleada, incluyendo la definición de las categorías utilizadas en la escala de veracidad. Esta información puede destacarse en recuadros separados para facilitar su identificación y reforzar la transparencia del proceso (Imagen 4).

Imagen 4. Información sobre fuentes y metodología

Fuente: https://bit.ly/3TIS6Mi

La Tabla 1 recoge la estructura tipo de los informes de verificación.

Tabla 1. Estructura de los verificados

Parte del verificado	Contenido que incluye
Titular	Veredicto / conclusión
Subtítulo	Explicación del veredicto / conclusión
Imagen del hecho y/o narrativa desinformativos	Debe ir etiquetada como falsa, engañosa, etc.
Cuerpo	Primer párrafo
	Ws del relato desinformativo (qué dice la desinformación, dónde circula, cómo y cuándo)
	Veredicto / conclusión
	Resto del cuerpo
	Explicación del proceso (técnicas, herramientas y fuentes consultadas)
	Contexto de la desinformación (ampliación del relato desinformativo y otras narrativas falsas similares)
	Final
	Recopilación de fuentes
	Explicación de las categorías (“escala de la verdad”)

Fuente: García-Marín (2024)

Del contenido a la estrategia

Nótese que la verificación se dirige exclusivamente al contenido que circula o al discurso que se pronuncia de forma concreta. El vídeo o la fotografía falsa constituyen la cara más visible de la desinformación. Sin embargo, estos contenidos desinformativos específicos siempre forman parte de narrativas manipuladoras más amplias que, perfectamente orquestadas y muchas veces inadvertidas para la ciudadanía, utilizan estrategias para introducir en la opinión pública una idea controvertida o establecer un debate a propósito de un tema que resulta de interés para los propagadores de desinformación. A veces, estas estrategias tienen como objetivo la división y polarización de una determinada sociedad para debilitarla en beneficio de intereses de terceros o, simplemente, invisibilizar un determinado relato a través de la cancelación de sus portavoces.

El Departamento de Seguridad Nacional de España, a través de su Foro de Expertos contra las Campañas de Desinformación, ha identificado y descrito algunas de estas estrategias (caracterizadas por su opacidad), a cuyo servicio se colocan los contenidos desinformativos específicos que, en forma de texto, audio o imagen, alcanzan a millones de usuarios diariamente. Algunas de estas estrategias son las siguientes (Arce García *et al*., 2024):

1. Arenques podridos (*rotten herrings*). Estrategia de desinformación que consiste en vincular falsamente y de manera reiterada y sostenida en el tiempo a una persona, colectivo o entidad con uno o varios escándalos o episodios de corrupción en el caso de la política. Aunque dichas acusaciones sean posteriormente refutadas o desmentidas, la conexión emocional y cognitiva entre el sujeto y el contenido negativo persiste en la percepción pública, generando un daño reputacional duradero.
2. Manguera de falsedades (*firehosing*). Emisión masiva, rápida y repetitiva de información falsa o engañosa. Su objetivo es saturar

el espacio informativo, dificultando que las audiencias puedan discernir entre hechos verídicos y falsos. Esta sobrecarga informativa obstaculiza los procesos de verificación y favorece la confusión, debilitando la capacidad crítica del público.

3. Blanqueo de información (*information laundering*). Conjunto de estrategias de manipulación informativa orientadas a conferir legitimidad a determinados contenidos mediante su redistribución a través de intermediarios que omiten deliberadamente la atribución a la fuente original, con el fin de encubrir el verdadero origen de la información. Se desarrolla en tres etapas: primero, la publicación inicial del contenido falso por parte de uno o varios canales de comunicación (habitualmente pseudomedios o cuentas hiperpartisanas en redes sociales); en segundo lugar, la diseminación a través de intermediarios —frecuentemente interrelacionados— que disimulan su vínculo con el emisor primario y contribuyen a desdibujar su procedencia; y finalmente, la incorporación del contenido en el discurso público, lo que amplifica su alcance y refuerza su aparente credibilidad.
4. Técnicas de supresión (*doxing* o *doxeo*). Divulgación deliberada y pública de información personal sensible perteneciente a un individuo, colectivo o entidad, con la intención de perjudicar su integridad. Esta práctica no solo busca exponer y dañar a la persona afectada, sino también generar intimidación, coacción o humillación, con el propósito de silenciarla o inhibir su participación en el espacio público.
5. Ataque mariposa (*butterfly attack*). Estrategia orientada a infiltrarse en comunidades, campañas o grupos ya consolidados con el propósito de generar fracturas internas y neutralizar su funcionamiento. Esta táctica se basa en la introducción de actores encubiertos —como impostores o troles— que, especialmente en entornos digitales y redes sociales, difunden desinformación con el fin de fomentar la desconfianza, provocar conflictos internos y debilitar la cohesión del grupo.

6. La gran mentira (*The Big Lie*). Estrategia propagandística que consiste en difundir deliberadamente una falsedad evidente (incluso muy burda), diseñada para generar una reacción emocional intensa —como miedo, repulsión o indignación— en el público receptor. Esta técnica se basa en el principio psicológico de que las emociones fuertes pueden inhibir el pensamiento crítico, facilitando que la falsedad se mantenga en el subconsciente incluso después de haber sido desmentida.
7. *Astroturfing.* Táctica comunicativa desplegada en redes sociales que implica la utilización de múltiples cuentas que, aunque aparentan ser usuarios comunes (generalmente con pocos seguidores y seguidos), operan de manera sincronizada. Estas cuentas pueden ser nuevas o reconvertidas, es decir, previamente dedicadas a otros temas y reutilizadas para nuevos fines, aprovechando la base de seguidores acumulada. Su acción coordinada se basa en el aprovechamiento del anonimato relativo que ofrecen las plataformas digitales, con el objetivo de diseminar, amplificar y saturar el espacio con contenidos desinformativos. Esta estrategia busca posicionar determinadas narrativas, generar tendencias artificiales (como *trending topics*) y moldear el debate público hacia temas específicos. Mientras que el término *grassroots* alude a movimientos auténticos y espontáneos surgidos desde la ciudadanía, el *astroturfing* representa una simulación de estos movimientos planificada estratégicamente para aparentar espontaneidad. Un ejemplo de ello sería el diseño de una campaña durante el proceso legislativo de una controvertida reforma energética en un país determinado. Una campaña de astroturfing pagada por empresas del sector energético consistiría en la creación y difusión masiva de mensajes en redes sociales que aparenten provenir de ciudadanos comunes, expresando un respaldo entusiasta a la reforma bajo el argumento de que reduciría los costes de la electricidad y promovería la soberanía energética nacional. Para lograr este efecto, se utilizarían cientos de cuentas automatizadas (*bots*) y perfiles falsos (*sockpuppets*) que repliquen

patrones de comportamiento humano, incluyendo interacciones con usuarios reales, uso de lenguaje coloquial y referencias a eventos locales.

8. Marketing de falsa bandera. Estrategia de comunicación engañosa en la que una organización, grupo o individuo oculta su identidad real y simula ser otra entidad —generalmente una parte neutral o incluso contraria— con el fin de manipular la percepción pública, influir en comportamientos o desacreditar a un oponente. El término proviene del ámbito militar y de inteligencia, donde se refiere a operaciones encubiertas que se hacen pasar por acciones del enemigo. Ejemplo: una campaña en redes sociales donde un grupo político crea cuentas falsas que simulan pertenecer a un movimiento ciudadano ambientalista. Estas cuentas critican a un partido rival por supuestas políticas contaminantes, aunque la información esté manipulada o fuera de contexto. El objetivo es erosionar la imagen del oponente sin que se perciba que la crítica proviene, en realidad, de un competidor político.

Conclusión

La verificación digital y el *fact-checking* se han consolidado como una respuesta necesaria, aunque quizá no suficiente, frente al desafío de la desinformación. Su eficacia depende de una combinación de factores: el rigor metodológico, la colaboración entre actores, la implicación ciudadana y el desarrollo de políticas públicas que promuevan un ecosistema informativo más transparente y resiliente. En este sentido, la verificación no solo es una herramienta técnica, sino también una práctica cívica que contribuye a fortalecer la democracia en tiempos de incertidumbre.

Esta labor crítica para identificar la desinformación que puebla el ecosistema informativo digital no debe ser patrimonio exclusivo de los profesionales de la comunicación. Todo ciudadano debe conocer, al

menos de forma básica, los principios esenciales de la desinformación y manejar estrategias y herramientas para la detección del contenido desinformativo y malicioso. Por ello, las sociedades democráticas no solo deben equiparse con un potente sistema de medios de comunicación que luchen contra el fenómeno de la desinformación, sino también las instituciones educativas deben proveer la suficiente (y necesaria) alfabetización mediática e informacional para facilitar la resiliencia de la población ante este tipo de campañas.

Competencias como la comprensión del fenómeno de la desinformación, la identificación de los tipos de desórdenes informativos que existen y sus distintas narrativas, el análisis de posibles sesgos en los contenidos, la identificación del propósito de un mensaje engañoso (ideológico, político, religioso, cultural, económico o propagandístico), la diferenciación entre información y opinión (sobre todo en las redes sociales), el contraste de un contenido sospechoso por otras vías o fuentes informativas de referencia, el discernimiento del contenido veraz de aquel que pretende engañar (Herrero-Diz *et al.*, 2022) y la detección de campañas estratégicas desinformativas más amplias son aspectos básicos que todo programa de alfabetización contra la desinformación debe incluir. Al análisis de la (imprescindible) educación mediática e informacional como antídoto contra este fenómeno dedicamos el capítulo final de este manual.

Agradecimientos

Este capítulo forma parte del proyecto "Desafíos, usos y limitaciones de la IA en el *fact-checking* y la lucha contra la desinformación" (DESAF_IA) (Ref. 2024/SOLCON-135623) financiado por la convocatoria de Proyectos IMPULSO a la investigación de la Universidad Rey Juan Carlos (2024).

Herramientas de verificación digital

Este capítulo concluye con un anexo que presenta las herramientas tecnológicas más habitualmente utilizadas por los *fact-checkers* profesionales y que están a disposición de forma gratuita para el público general. Aunque esta obra no tiene una perspectiva tecnocéntrica, resulta fundamental recomendar este tipo de instrumentos para conocer de primera mano sus principios esenciales, siempre teniendo en cuenta que lo interesante no es tanto la herramienta en sí, sujeta a evolución y a largo plazo perecedera, sino la filosofía que encierra su funcionamiento. Para conocer plenamente su filosofía y valorar críticamente su uso, es necesario sumergirse en su utilización; por ello se ofrece esta selección, clasificada por categorías.

1. Buscadores

Búsqueda avanzada de Google: site:
Descripción. El comando de búsqueda avanzada de Google **site:** es una herramienta de filtrado que permite restringir los resultados de búsqueda a un dominio web específico. Se trata de un operador booleano de búsqueda que optimiza la recuperación de información al delimitar el ámbito de consulta a un sitio web o dominio determinado, lo cual mejora la precisión y relevancia de los resultados obtenidos.
¿Cómo funciona? Es especialmente útil en contextos de verificación de hechos y análisis de desinformación por las siguientes razones: Filtrado de fuentes confiables. Permite buscar información exclusivamente dentro de sitios verificados, como medios de comunicación reconocidos, organismos oficiales o plataformas de verificación de hechos (por ejemplo, site:bbc.com o site:who.int). Contraste de narrativas. Facilita la comparación entre lo que se afirma en redes sociales o sitios dudosos y lo que publican fuentes autorizadas. Rastreo de contenido original. Ayuda a localizar la fuente primaria de una cita, imagen o declaración, lo cual es clave para detectar manipulaciones o atribuciones falsas. Auditoría de desinformación. Permite investigar si un sitio web ha publicado previamente contenido similar o recurrentemente engañoso.
Ejemplo de uso. Verificación de una supuesta declaración de la OMS sobre una nueva pandemia Circula en redes sociales una imagen con el logo de la Organización Mundial de la Salud (OMS) que afirma que se ha declarado una nueva pandemia. Para verificarlo, se utiliza en Google el comando: **site:who.int nueva pandemia** Esto restringe los resultados a publicaciones oficiales del sitio web de la OMS. Si no aparece ninguna declaración reciente relacionada con el tema, es un claro indicio de que la información viralizada es falsa o manipulada.

FactCheck Explorer
Descripción. Herramienta desarrollada por Google con el objetivo de facilitar el acceso a verificaciones de hechos realizadas por organizaciones especializadas en *fact-checking*. Puede definirse como un motor de búsqueda especializado en la recuperación de contenidos verificados, que permite a periodistas, investigadores y ciudadanos consultar rápidamente si una afirmación ha sido evaluada por verificadores acreditados.
¿Cómo funciona? Actúa como una interfaz de búsqueda que conecta al usuario con una base de datos global de verificaciones de hechos. Su funcionamiento se basa en los siguientes principios: - Indexación de fuentes verificadas. Agrega contenidos de organizaciones de *fact-checking* reconocidas internacionalmente, como *PolitiFact, AFP Fact Check, Maldita.es, Chequeado, Newtral,* entre otras - Búsqueda semántica y filtrado. Permite buscar por palabras clave, nombres de figuras públicas, temas específicos o ubicaciones geográficas. También ofrece filtros por idioma y fecha. - Etiquetado estructurado. Utiliza el estándar de marcado de datos estructurados de Schema.org para identificar afirmaciones verificadas y sus respectivas calificaciones (por ejemplo, "falso", "engañoso", "verdadero", etc.). - Acceso abierto y gratuito. Está disponible para cualquier usuario, aunque está especialmente diseñado para apoyar el trabajo de periodistas, verificadores y académicos.

Ejemplo de uso. Verificación de una afirmación viral sobre política migratoria Circula en redes sociales una afirmación que dice: *"El gobierno de X país ha abierto completamente sus fronteras a la inmigración ilegal"*. Un periodista desea comprobar si esta afirmación ha sido verificada. Accede a Fact Check Explorer. Introduce la búsqueda: "fronteras inmigración ilegal país X". Revisa los resultados. Encuentra una verificación de una organización confiable que califica la afirmación como *engañosa*, explicando que se trata de una interpretación errónea de una política temporal. Resultado. El periodista evita difundir información falsa y puede incluir en su artículo una fuente verificada que contextualiza correctamente la afirmación.
URL: https://toolbox.google.com/factcheck/explorer/search/list:recent;hl=es

2. Búsqueda inversa de imágenes

TinEye
Descripción. Motor de búsqueda inversa de imágenes desarrollado por la empresa canadiense Idée Inc., que permite rastrear la procedencia y el historial de una imagen en la web. Desde una perspectiva académica, puede definirse como una herramienta de análisis visual computacional que emplea algoritmos de reconocimiento de patrones y coincidencias de imágenes para identificar copias exactas o modificadas de un archivo gráfico en internet.
¿Cómo funciona? Es particularmente útil en el ámbito de la verificación digital y la lucha contra la desinformación visual por las siguientes razones: (1) rastreo del origen de una imagen (permite identificar el sitio web donde una imagen apareció por primera vez, lo cual es clave para verificar su contexto original), (2) detección de reutilización engañosa (ayuda a descubrir si una imagen ha sido sacada de contexto, por ejemplo, usada para ilustrar un evento actual cuando en realidad pertenece a otro lugar o época), (3) análisis de modificaciones (aunque no detecta manipulación digital como los análisis forenses, sí puede mostrar versiones alteradas de una imagen, lo que permite inferir posibles intentos de engaño), y (4) seguimiento de la difusión (muestra cómo y dónde se ha replicado una imagen en la web, lo que permite rastrear su viralización y evaluar su impacto).
Ejemplo de uso. Verificación de una imagen viral sobre una catástrofe natural Durante una supuesta inundación en una ciudad, circula una imagen impactante de calles completamente sumergidas. Deseamos verificar si la imagen corresponde realmente con el evento real: Carga la imagen en TinEye. Resultados. TinEye muestra que la imagen fue publicada originalmente en 2017 en un artículo sobre inundaciones en otro país. Conclusión. La imagen ha sido reutilizada fuera de contexto para generar alarma. Se puede desmentir la desinformación y publicar la fuente original.
URL: https://www.tineye.com
Herramientas similares: Google Lens

3. Análisis forense de imágenes

Forensically
Descripción. Forensically es una herramienta de análisis forense digital orientada al examen detallado de imágenes con el fin de detectar manipulaciones, alteraciones o inconsistencias. Consiste en una plataforma de análisis visual computacional que aplica técnicas de procesamiento de imágenes para evaluar la autenticidad de archivos gráficos, siendo especialmente útil en contextos de verificación digital y lucha contra la desinformación visual.
¿Cómo funciona? Permite a periodistas, investigadores y verificadores de hechos examinar imágenes sospechosas mediante una serie de filtros y análisis técnicos que revelan indicios de edición o falsificación. Sus principales funcionalidades incluyen: (1) análisis de metadatos (extrae información técnica del archivo, como la fecha de creación, el dispositivo utilizado o el *software* de edición, lo que puede revelar alteraciones), (2) *Error Level Analysis* (ELA, consistente en la detección de diferencias en los niveles de compresión de distintas áreas de la imagen, lo que puede indicar zonas editadas), (3) *Clone Detection* (identifica regiones duplicadas dentro de una imagen, lo que sugiere posibles manipulaciones como el copiado y pegado de elementos), (4) *Noise Analysis* (examina la distribución del ruido digital para detectar inconsistencias que no serían visibles a simple vista), y (5) *Magnifier y Luminance Gradient* (permiten inspeccionar detalles finos y variaciones de iluminación que pueden delatar ediciones artificiales).
Ejemplo de uso. Verificación de una imagen viral sobre un supuesto acto de violencia policial Una imagen circula en redes sociales mostrando a un agente de policía agrediendo a un civil. La imagen genera indignación, pero algunos usuarios dudan de su autenticidad. Carga de la imagen en Forensically. Aplicación de ELA. Se detectan diferencias de compresión en el área donde aparece el brazo del agente, lo que sugiere que fue añadido digitalmente. *Clone Detection*. Revela que el fondo ha sido duplicado para ocultar elementos originales. Conclusión. La imagen ha sido manipulada. El equipo de verificación publica un análisis explicando la falsificación, ayudando a frenar la desinformación.
URL: https://29a.ch/photo-forensics/#forensic-magnifier
Herramientas similares: Image Forensic

4. Desinformación visual (vídeos y fotografías)

inVID WeVerify
Descripción. Plataforma digital orientada a la verificación de contenido multimedia en línea. Desde una perspectiva académica, puede definirse como un conjunto de herramientas tecnológicas integradas que aplican técnicas de análisis forense digital, minería de datos y verificación colaborativa para combatir la desinformación, especialmente en formatos audiovisuales.
¿Cómo funciona? Opera principalmente como una extensión de navegador (para Chrome y Firefox) que permite a periodistas, verificadores de hechos, investigadores y ciudadanos analizar contenido multimedia de forma rápida y precisa. Sus principales funcionalidades incluyen: Fragmentación de video (*Keyframe Extraction*). Extrae fotogramas clave de un vídeo para facilitar su análisis individual y posterior búsqueda inversa. Búsqueda inversa de imágenes. Permite realizar búsquedas en motores como Google, Yandex, Bing y Baidu para rastrear el origen de una imagen o fotograma. Análisis de metadatos. Extrae información técnica de archivos multimedia (fecha, ubicación, dispositivo, *software* de edición), útil para detectar alteraciones o falsificaciones. Análisis forense de imágenes. Incluye herramientas como detección de clonación, análisis de errores de compresión (ELA) y gradientes de luminancia para identificar manipulaciones digitales. Verificación de redes sociales. Evalúa perfiles en plataformas como Twitter, Facebook o YouTube para detectar automatización, suplantación o comportamiento coordinado. Verificación colaborativa (WeVerify *plugin*): Permite a los usuarios compartir hallazgos, colaborar en investigaciones y construir redes de verificación ciudadana.
Ejemplo de uso. Verificación de un vídeo viral durante una protesta social Un vídeo que muestra supuestos actos de violencia policial se vuelve viral en redes sociales. Un equipo de verificación desea comprobar su autenticidad antes de difundirlo. Pasos con InVID WeVerify: Fragmentación del vídeo. Se extraen fotogramas clave del video. Búsqueda inversa. Se realiza una búsqueda de los fotogramas en Google y Yandex. Se descubre que las imágenes ya habían sido publicadas en otro contexto, en otro país, años antes. Análisis de metadatos: Se detecta que el video fue editado recientemente, aunque se presenta como actual. Conclusión. Se determina que el vídeo ha sido reutilizado fuera de contexto para manipular la percepción del evento.
URL: https://www.invid-project.eu/tools-and-services/invid-verification-plugin/

5. *Deepfakes*

Deepware
Descripción. Herramienta tecnológica basada en inteligencia artificial diseñada para la detección automatizada de *deepfakes*. Emplea algoritmos de aprendizaje automático para identificar alteraciones sintéticas en vídeos, imágenes y audios, con el objetivo de preservar la integridad informativa y combatir la desinformación visual.
¿Cómo funciona? Se basa en los siguientes principios: Análisis de patrones biométricos y visuales. Detecta inconsistencias en expresiones faciales, sincronización labial, movimientos oculares y texturas de piel, que suelen ser indicios de manipulación digital. Detección de artefactos generados por IA. Utiliza modelos entrenados para identificar huellas digitales típicas de algoritmos generativos como las redes generativas adversarias (GAN), que intervienen en la producción de *deepfakes*. Evaluación de autenticidad. Clasifica el contenido como auténtico o manipulado, proporcionando un nivel de confianza basado en el análisis técnico. Aplicación en entornos críticos. Puede integrarse en flujos de trabajo de medios de comunicación, plataformas sociales, agencias gubernamentales y empresas para prevenir la difusión de contenido falso.
Ejemplo de uso. Verificación de un vídeo viral durante una campaña electoral Durante una campaña política, circula un vídeo en el que un candidato parece hacer declaraciones polémicas. Un medio de comunicación, antes de difundirlo, decide verificar su autenticidad con Deepware: Carga del vídeo en la plataforma Deepware. Análisis automático. El sistema detecta desincronización entre el audio y los movimientos faciales, así como patrones visuales anómalos en los bordes del rostro. Resultado. Deepware clasifica el video como *deepfake* con un alto nivel de certeza. Impacto. El medio evita difundir contenido falso, protege su credibilidad y contribuye a un entorno informativo más seguro.
URL: https://scanner.deepware.ai
Herramientas similares: Hive Moderation, Hugging Face

6. Periodismo de investigación y/o de datos

Pinpoint
Descripción. Herramienta de investigación desarrollada por Google como parte de su iniciativa para apoyar el periodismo y la verificación de hechos. Se configura como una plataforma de análisis documental asistida por inteligencia artificial que permite a periodistas, investigadores y académicos explorar, organizar y extraer información relevante de grandes volúmenes de documentos, audios, imágenes y otros formatos textuales.
¿Cómo funciona? Pinpoint contribuye a la lucha contra la desinformación a través de las siguientes funcionalidades clave: Procesamiento de grandes volúmenes de datos. Permite subir y analizar hasta 200.000 documentos por colección, incluyendo PDFs, correos electrónicos, imágenes escaneadas, manuscritos y archivos de audio. Reconocimiento óptico de caracteres (OCR) y transcripción automática. Convierte el texto presente en imágenes, el vídeo y el audio en texto digital, lo que facilita la búsqueda y análisis de contenido no estructurado. Búsqueda avanzada y contextual. Utiliza tecnologías de búsqueda semántica y reconocimiento de entidades (personas, lugares, organizaciones) para identificar patrones narrativos, fuentes y conexiones entre documentos. Colaboración y verificación. Facilita el trabajo colaborativo entre periodistas y verificadores, permitiendo compartir documentos, resaltar fragmentos clave y construir narrativas verificadas a partir de evidencias documentales. Seguridad y privacidad. Las colecciones de documentos a analizar son privadas por defecto y están protegidas por las tecnologías de seguridad de Google, lo que garantiza la integridad del proceso de investigación.
Ejemplo de uso. Investigación sobre una campaña coordinada de desinformación. Un equipo de periodistas de investigación recibe una filtración de correos electrónicos y documentos internos que podrían revelar una campaña organizada para difundir información falsa sobre un tema de salud pública. Carga de documentos. Suben miles de archivos a Pinpoint, incluyendo correos, PDFs y notas manuscritas escaneadas. Análisis automatizado. Pinpoint identifica menciones recurrentes a términos clave, nombres de personas y organizaciones implicadas. Transcripción de audios. Archivos de audio de reuniones son transcritos automáticamente, permitiendo búsquedas por palabras clave. Verificación cruzada. El equipo cruza la información con fuentes públicas y encuentra evidencia de coordinación narrativa. Publicación del reportaje. Gracias a Pinpoint, logran estructurar una investigación sólida con respaldo documental, desmontando una red de desinformación.
URL: https://journaliststudio.google.com/pinpoint/about/es_es/

7. Cronolocalización

Wolfram Alpha
Descripción. Herramienta gratuita que permite saber qué tiempo hacía en una localización concreta en una fecha y hora anteriores.
¿Cómo funciona? Permite buscar la información meteorológica de una localización concreta en un periodo temporal determinado. Ofrece información sobre las temperaturas y otras condiciones meteorológicas, como las precipitaciones, presencia de viento, etc.
Ejemplo de uso. Verificación de una afirmación sobre el clima extremo. Una publicación viral afirma que *"la ciudad de Madrid alcanzó los 50 °C el 10 de junio de 2025"*. Para verificarlo: Un periodista accede al *widget* de clima de Wolfram Alpha. Introduce "Madrid, June 10, 2025" como consulta. El *widget* devuelve los datos meteorológicos reales de esa fecha, mostrando que la temperatura máxima fue de 38 °C. Conclusión: La afirmación es falsa. El periodista puede desmentirla con evidencia computacional verificable.
URL: https://www.wolframalpha.com/widgets/view.jsp?id=dc17deabb063272f86b3c0e6cbfafcba

8. Archivo

Wayback Machine
Descripción. Herramienta digital desarrollada por el *Internet Archive*, una organización sin fines de lucro dedicada a la preservación del conocimiento en línea. Se trata de un archivo web de acceso abierto que permite consultar versiones históricas de páginas web, capturadas en distintos momentos. Su función principal es conservar el contenido digital para garantizar la trazabilidad, la transparencia informativa y la memoria colectiva de la web.
¿Cómo funciona? Su utilidad contra la desinformación se basa en: (1) recuperación de contenido eliminado o modificado (permite acceder a versiones anteriores de páginas web, lo que es útil para comprobar si una fuente ha cambiado, borrado o manipulado información), (2) trazabilidad de fuentes (facilita el seguimiento de la evolución de una narrativa o noticia, permitiendo comparar cómo ha sido presentada en distintos momentos), (3) preservación de evidencia digital (actúa como respaldo documental en investigaciones periodísticas, académicas o legales, especialmente cuando el contenido original ya no está disponible), y (4) detección de inconsistencias (ayuda a identificar contradicciones entre versiones actuales y pasadas de un mismo sitio web, lo que puede revelar intentos de encubrimiento o manipulación).

Ejemplo de uso. Verificación de una declaración eliminada de un sitio oficial. Durante una crisis sanitaria, un organismo gubernamental publica una declaración minimizando los riesgos de una enfermedad. Semanas después, la declaración desaparece del sitio web oficial y se niega su existencia. Un periodista accede a Wayback Machine. Introduce la URL del sitio oficial y selecciona una fecha cercana a la publicación original. Encuentra una versión archivada que contiene la declaración eliminada. Usa esta evidencia para demostrar que la información fue publicada y luego retirada, lo que contribuye a la transparencia y a la rendición de cuentas.
URL: https://web.archive.org
Herramientas similares: Archive

9. Geolocalización

Yandex Maps
Descripción. Plataforma cartográfica digital desarrollada por la empresa tecnológica rusa Yandex. Funciona de manera similar a Google Maps, ofreciendo servicios de geolocalización, navegación GPS, visualización de mapas en diferentes capas (satélite, tráfico, transporte público, etc.), y herramientas para la planificación de rutas. Su base de datos se alimenta de imágenes satelitales, datos geoespaciales y contribuciones de usuarios, lo que permite una actualización constante y detallada del entorno geográfico.
¿Cómo funciona? Yandex Maps puede considerarse una herramienta de verificación geoespacial, útil en el contexto de la alfabetización mediática y la verificación de hechos. Su utilidad radica en la capacidad de: Corroborar ubicaciones mencionadas en noticias o publicaciones en redes sociales. Verificar la existencia o disposición de infraestructuras (como edificios, carreteras, instalaciones militares, etc.). Comparar imágenes satelitales históricas para detectar cambios en el terreno que puedan confirmar o desmentir narrativas visuales. Contrastar información visual con otras plataformas cartográficas (como Google Maps o OpenStreetMap), lo que permite detectar manipulaciones o inconsistencias.
Ejemplo de uso. Un caso concreto tuvo lugar durante el conflicto en Ucrania, donde usuarios y periodistas utilizaron Yandex Maps para verificar la ubicación de supuestos bombardeos o movimientos militares. Por ejemplo, si una imagen viral mostraba un edificio destruido en una ciudad específica, los verificadores podían (1) buscar la dirección exacta en Yandex Maps, (2) comparar la estructura del edificio con imágenes satelitales previas, y (3) confirmar si el edificio existía, si coincidía con la arquitectura local y si había señales de daño reciente. Este tipo de análisis ayudó a desmentir imágenes que en realidad correspondían a conflictos pasados o a otros países, contribuyendo así a frenar la propagación de desinformación visual
URL: https://yandex.com/maps/?ll=-2.169054%2C39.675003&z=13
Herramientas similares: Google Maps, OpenStreetMap

10. Tutoriales

Verification Handbook for Disinformation and Media Manipulation
Descripción. Manual que proporciona los conocimientos necesarios para investigar cuentas en redes sociales, *bots*, aplicaciones de mensajería privada, operaciones de información, *deepfakes*, así como otras formas de desinformación y manipulación mediática. Publicado por el European Journalism Centre.
URL: https://datajournalism.com/read/handbook/verification-3/

First Draft Training
Descripción. Iniciativa educativa de la organización sin fines de lucro First Draft, dedicada a combatir la desinformación y mejorar las habilidades de verificación digital tanto en periodistas como en el público general. Proporciona una biblioteca gratuita de contenidos formativos, que incluye cursos online sobre desinformación, verificación de hechos y análisis de redes sociales, así como guías prácticas para identificar y contrarrestar narrativas falsas. También provee retos interactivos para poner a prueba habilidades como la geolocalización o la verificación de imágenes. Asimismo, ofrece *webinars* y recursos multilingües, disponibles en español, inglés, francés, alemán, entre otros idiomas.
URL: https://firstdraftnews.org/training/

Referencias

Antonakaki, D., Fragopoulou, P., & Ioannidis, S. (2021). A survey of Twitter research: data model, graph structure, sentiment analysis and attacks. *Expert systems with applications, 164*, 114006. https://doi.org/10.1016/j.eswa.2020.114006

Arce García, S., Rodríguez Fernández, L., Establés Heras, M.J., García-Marín, D., Marín García, B., Martín Jiménez, V., Pérez Curiel, C., Said Hung, E., Salaverria Aliaga, R., & Wagner, A. (2024). 125 términos sobre desinformación. En*: Trabajos del Foro contra las Campañas de Desinformación. Iniciativas 2024.* Ministerio de la Presidencia, Justicia y Relaciones con las Cortes, pp. 9-39.

Aspray, W., & Cortada, J.W. (2019). *From urban legends to political fact-checking. Online scrutiny in America*, *1990-2015*. Springer.

Brandtzaeg, P. B., Følstad, A., & Chaparro Domínguez, M. Á. (2018). How Journalists and Social Media Users Perceive Online Fact-Checking and Verification Services. *Journalism Practice*, *12*(9), 1109–1129. https://doi.org/10.1080/17512786.2017.1363657

Cuartielles R., & Carral U. (2024). Funcionamiento y viabilidad de las alianzas de fact-checking en España: el caso de Comprobado. *Estudios sobre el Mensaje Periodístico*, *30*(4), 817-828. https://doi.org/10.5209/emp.96756

Fernández Barrero, M.A., & Aramburú Moncada, L.G. (2024). Estrategias de prevención de la desinformación desde la enseñanza de la redacción periodística. En: Pérez-Curiel, C. y Domínguez-García. R. (Coords.), *Periodismo y desinformación. Manual de aplicación de técnicas digitales para la verificación de noticias en la docencia universitaria.* Fragua, pp 153-169.

García-Marín, D. (2024). Periodismo contra la desinformación. Proceso y estructura de las verificaciones en el fact-checking. *Infonomy*, 2(2). https://doi.org/10.3145/infonomy.24.026

García-Marín, D., Rubio-Jordán, A. V.., & Salvat-Martinrey, G. (2023). Chequeando al fact-checker. Prácticas de verificación política y sesgos partidistas en Newtral (España). *Revista De Comunicación,* 22(2), 207–223. https://doi.org/10.26441/RC22.2-2023-3184

Graves, Lucas (2016). *Deciding what's true: The rise of political fact-checking in American journalism.* Columbia University Press.

Herrero-Diz, P., Pérez-Escolar, M., & Varona Aramburu, D. (2022). Competencias de verificación de contenidos: una propuesta para los estudios de Comunicación. *Revista De Comunicación*, *21*(1), 231–249. https://doi.org/10.26441/RC21.1-2022-A12

Liu, X., Qi, L., Wang, L., & Metzger, M. J. (2023). Checking the Fact-Checkers: The Role of Source Type, Perceived Credibility, and Individual Differences in Fact-Checking Effectiveness. *Communication Research*. https://doi.org/10.1177/00936502231206419

Mateos, C. (2021). Imagen y vídeos *fakes*: la certeza en el documento audiovisual. En: Elías, C. y Teira, D. (Coords.), *Manual de periodismo y verificación de noticias en la era de las fake news.* UNED, pp. 133-172

Nyhan, B., & Reifler, J. (2010). When Corrections Fail: The Persistence of Political Misperceptions. *Political Behavior, 32,* 303–330. https://doi.org/10.1007/s11109-010-9112-2

Oshikawa, R., Qian, J., Wang, W.-Y. (2020). A survey on natural language processing for fake news detection. *arXiv*, 1811.00770. https://arxiv.org/pdf/1811.00770.pdf

Pennycook, G., & Rand, D.G. (2019). Fighting misinformation on social media using crowdsourced judgments of news source quality. *Psychological and Cognitive Sciences, 116*(7), 2521-2526. https://doi.org/10.1073/pnas.1806781116

Potthast, M., Kiesel, J., Reinartz, K., Bevendorff, J., & Stein, B. (2017). A stylometric inquiry into hyperpartisan and fake news. *arXiv,* 1702.05638. https://arxiv.org/pdf/1702.05638.pdf

Van Erkel, P. F. A., van Aelst, P., de Vreese, C. H., Hopmann, D. N., Matthes, J., Stanyer, J., & Corbu, N. (2024). When are Fact-Checks Effective? An Experimental Study on the Inclusion of the Misinformation Source and the Source of Fact-Checks in 16 European Countries. *Mass Communication and Society*, *27*(5), 851–876. https://doi.org/10.1080/15205436.2024.232154

Vázquez-Herrero, J., Vizoso, Á., & López-García, X. (2019). Innovación tecnológica y comunicativa para combatir la desinformación: 135 experiencias para un cambio de rumbo", e280301. https://doi.org/10.3145/epi.2019.may.01

6. Alfabetización mediática contra la desinformación

Roberto Aparici
UNED
España
Manuel Álvarez Rufs
UNED
España
Fernando Bordignon
UNIPE
Argentina

Alfabetización mediática y educomunicación

En primer lugar, considerando que todas y todos somos educomunicadoras y educomunicadores, proponemos la educomunicación como un modelo de alfabetización mediática e informacional, útil para combatir la desinformación. La educomunicación implica la interrelación de dos campos de estudio: la educación y la comunicación. Es evidente que para educar necesitamos al menos dos personas que establezcan una comunicación, un diálogo. Es imposible educar sin establecer algún tipo de comunicación, y la comunicación implica el desarrollo de un acto educativo. La educomunicación también se puede entender desde dos puntos de

vista diferentes. Por una parte, como *enseñanza* de los *medios de comunicación*, y, por otra parte, como el *uso de los medios en educación*.

La educomunicación se entiende como un concepto de gran amplitud que también incluye la alfabetización mediática, y que ha recibido diferentes denominaciones a lo largo de la historia, según los diferentes contextos y momentos históricos, tales como: Recepción Crítica de los Medios de Comunicación; Pedagogía de la Comunicación; Educación para la Televisión; Pedagogía de la Imagen; Didáctica de los Medios Audiovisuales; Educación para la Comunicación; Comunicación Educativa; Educación Mediática; o Alfabetización Mediática e Informacional.

Cada denominación asume diferentes perspectivas e interpretaciones que se producen fruto de las posibles hibridaciones entre la educación y la comunicación en los diversos ámbitos educativos y comunicativos. En este manual, proponemos la siguiente definición:

> Educomunicación incluye, sin reducirse, el conocimiento de los múltiples lenguajes y medios por los que se realiza la comunicación personal, grupal y social. Abarca también la formación del sentido crítico, inteligente, frente a los procesos comunicativos y sus mensajes para descubrir los valores culturales propios y la verdad. (CENECA, UNICEF & UNESCO, 1992)

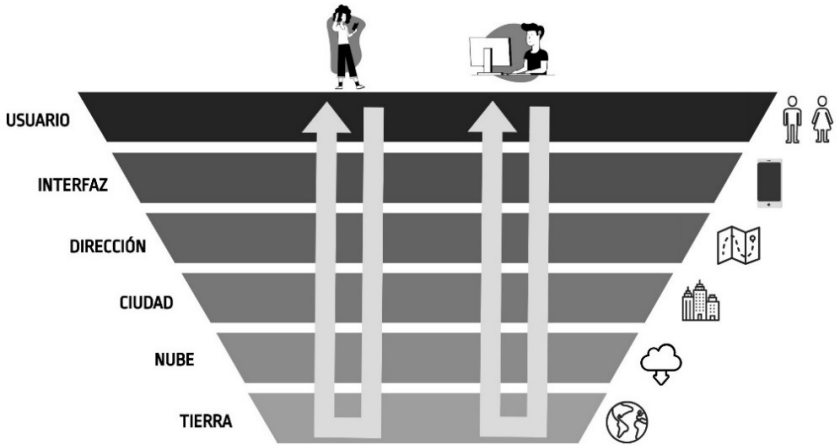
Debemos tener en cuenta que, cuando se publica esta definición en el año 1992, internet ya existía desde los años 60 del siglo XX, pero la web, tal y como la conocemos hoy en día, no se hace pública hasta el año 1993. ¿Ha cambiado algo? ¿Sigue siendo aplicable esta definición teniendo en cuenta el desarrollo de los sistemas de comunicación digitales actuales?

Eco-educomunicación en el entramado cibernético

En este manual vamos a tener en cuenta el concepto de *eco-educomunicación* que proponen Aparici *et al*. (2024a) en la era de la IA, desde una perspectiva holística y ecológica, dada la interrelación

de los seres vivos, las personas y las máquinas con su entorno, y las consecuencias que esto puede tener desde un punto de vista educativo y comunicativo. Para abordar el conocimiento de los múltiples lenguajes y medios por los que se realiza la comunicación personal, grupal y social, proponemos un modelo de entramado cibernético en capas (Bratton 2015; Álvarez Rufs, 2022), en el que podemos distinguir: Usuario, Interfaz, Dirección, Ciudad, Nube y Tierra (Imagen 1).

Imagen 1. Entramado cibernético



Fuente: Álvarez Rufs (2022)

Por **Usuario** entendemos cualquier ente, vivo o inanimado, que pueda conectarse al entramado cibernético. Aquí podemos incluir a las personas usuarias, a las máquinas, a los animales, a las plantas o a cualquier dispositivo capaz de conectarse al entramado. La capa Usuario es el lugar donde se personifican los efectos del entramado.

La **Interfaz** es el punto de contacto por el que se establece la comunicación con el resto del entramado cibernético por parte del Usuario, y al mismo tiempo gobierna el intercambio de información entre los diferentes sistemas implicados. La capa Interfaz puede

estar constituida por una pantalla, una palanca, una aplicación de smartphone, un volante, una frontera internacional, etc.

La **Dirección** es el identificador universal de cada dispositivo, el cual se convierte en una entidad que puede comunicarse dentro del entramado cibernético y que resulta identificada de manera unívoca dentro del mismo. Esta capa permite ubicar destinatarios individualizados y también actúa como medio de comunicación entre ellos, permitiendo la emisión y recepción de información en este plano común.

Por otra parte, la **Ciudad** es el punto por el que nos conectamos a la red a través de algún dispositivo, es decir, comprende un entorno de redes que sitúan los asentamientos humanos dentro de su propia estructura.

La **Nube** abarca una compleja infraestructura compuesta por servidores, bases de datos, fuentes de energía, cables ópticos, medios de transmisión inalámbricos y aplicaciones distribuidas. Implica un alto consumo energético y de recursos materiales.

La capa **Tierra** proporciona el sustento, materia y energía, para toda la infraestructura que compone el entramado cibernético en cada una de sus capas. Esto implica tener en cuenta una serie de problemas que están relacionados con la gobernanza ecológica, la sostenibilidad medioambiental y la colonización tecnológica.

La eco-educomunicación resulta imprescindible para poder descubrir los nuevos diseños, códigos y lenguajes de la colonización tecnológica, con la finalidad de atender el poder normativo y cultural que se ejerce desde los diferentes medios de comunicación digitales propios de los actuales entes colonizadores (Aparici *et al*., 2024b).

Minería de datos versus minería tradicional

Las tecnologías de información y de comunicación influyen en la vida de las personas y en la sostenibilidad del planeta Tierra a medio

y largo plazo. Todas las transacciones comunicativas que se llevan a cabo dentro del entramado cibernético implican un coste material y energético, tanto para las personas y otras entidades que acceden al mismo, como para el propio planeta Tierra.

El hecho de llevar a cabo una minería de datos capaz de sustentar el entramado cibernético y todas las comunicaciones digitales que se producen en el mismo a nivel planetario, incluidas nuestras relaciones con la IA, implica el desarrollo de una minería tradicional en algún lugar del mundo, y el consumo de grandes cantidades de energía (producida de forma más o menos limpia), sin obviar todo el agua necesaria para refrigerar los grandes centros de datos dispersos en diferentes lugares del planeta. Tal y como indica Crawford (2023), "la minería que crea la Inteligencia Artificial es tan literal como metafórica".

En este apartado ofrecemos algunos ejemplos y reflexiones que se relacionan con el mantenimiento de la propia infraestructura que sustenta el entramado, es decir, con la extracción a nivel terrestre de los materiales necesarios para poder fabricar los dispositivos digitales que constituyen las interfaces, direcciones, ciudades y nubes que sustentan las múltiples transacciones y comunicaciones que se producen constantemente dentro del entramado cibernético que se ha descrito anteriormente.

Litio

El litio se utiliza principalmente para la fabricación de las baterías que alimentan de corriente eléctrica los distintos dispositivos electrónicos, digitales, que componen el entramado cibernético.

Estas son diez de las minas más importantes de litio en explotación actual:

1. Mina Greenbushes (Australia).
2. Salar de Atacama – SQM (Chile).
3. Pilgagoora (Australia).
4. Salar de Atacama – Albermarle (Chile).

5. Mina Mt Marion (Australia).
6. Mina Wodgina (Australia).
7. Salar de Hombre Muerto (Argentina).
8. Mina Mount Cattlin.
9. Salar de Olaroz (Argentina).
10. Mina Silver Peak (EEUU).

Como se puede observar, cinco de ellas están situadas en Australia, y las restantes, las otras cinco, en América. En el caso de América, una se localiza en Estados Unidos, dos en Argentina y otras dos en Chile. ¿Qué nos hace pensar esto? ¿No hay reservas de litio en otros lugares del mundo? ¿Por qué no se extrae litio en otros lugares? ¿Quiénes son los propietarios de estas explotaciones, de estas minas?

El triángulo del litio en América se conforma en los países de Chile, Argentina y Bolivia. Sin embargo, Bolivia no aparece entre los principales países con explotación de litio. ¿Por qué? ¿Tiene alguna implicación la situación o el contexto geopolítico de Bolivia en comparación con sus países vecinos? En el caso de España, se estima la existencia de un 5% de las reservas mundiales de litio, pero no se está llevando a cabo su explotación. También habría que plantearse el por qué.

El entorno donde se explotan las minas de las que se extraen los materiales necesarios para construir el entramado cibernético no es un entorno agradable. Son minas que generan una gran cantidad de productos tóxicos que destruyen el entorno. ¿Qué ocurre en estos lugares? ¿Existen personas oprimidas? ¿Existen conflictos bélicos? ¿Existen comunidades indígenas que han sido desplazadas por este tipo de minería que sustenta, al fin y al cabo, las telecomunicaciones digitales en el planeta Tierra?

Tierras raras

Igualmente, también podemos hablar de las tierras raras, de las que se extraen elementos químicos necesarios para la fabricación de componentes electrónicos, vehículos eléctricos e híbridos, bombillas

LED, es decir, distintos componentes electrónicos necesarios para la fabricación de los dispositivos que se integran en el entramado cibernético.

"Tierras raras" es el nombre común por el que se conoce a 17 elementos químicos. Estos son Escandio, Itrio y los quince elementos del grupo de los lantánidos: Lantano, Cerio, Praseodimio, Neodimio, Prometio, Samario, Europio, Gadolinio, Terbio, Disprosio, Holmio, Erbio, Tulio, Iterbio y Lutecio. Al procesar una tonelada de estas tierras se generan unos 250 metros cúbicos de agua altamente ácida y radioactiva.

Al igual que ocurre con el litio, no se están explotando todas las reservas de tierras raras que existen a nivel mundial. China y Mongolia son dos de los principales países con reserva de tierras raras en el mundo. Cabe destacar que el 33%, es decir, la tercera parte de las tierras raras se encuentran en Estados Unidos, pero no están siendo explotadas en ese país. Actualmente, el 97% de la explotación mundial de tierras raras está concentrada en China.

El lago Baotou es uno de los lugares donde se está almacenando toda esta cantidad de miles y miles de litros de agua ácida y radioactiva, que procede del procesamiento de las tierras raras necesarias para obtener elementos químicos preciados en el mundo de las tecnologías digitales.

Coltán

Otro ejemplo de cómo el mantenimiento del entramado cibernético digital y del sistema de telecomunicaciones digitales en el mundo afecta directamente al planeta Tierra sería la extracción de coltán.

El coltán tiene esta denominación por la contracción del nombre de dos minerales bien conocidos, que serían: la Columbita (COL), óxido de niobio con hierro y manganeso, y la Tantalita (TAN), óxido de tantalio con hierro y manganeso. El coltán es imprescindible para la fabricación de semiconductores y dispositivos electrónicos como

los teléfonos *smartphone* que llevamos en nuestros bolsos y bolsillos. La República Democrática del Congo contiene un 80% de las reservas mundiales de coltán, y su extracción genera tanto conflictos bélicos como la explotación de menores de edad, niños y niñas, en unas minas que también destrozan el entorno y afectan al medioambiente.

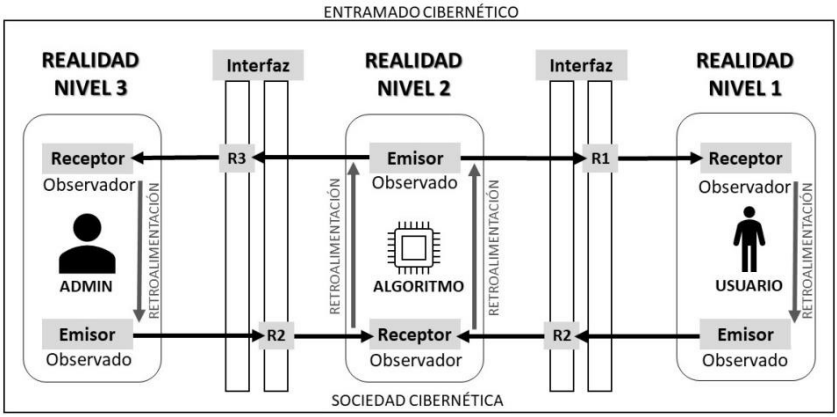
Todo esto ocurre en el planeta **Tierra** mientras que los **Usuarios** acceden al *entramado cibernético* a través de la **Interfaz** de algún dispositivo provisto de una **Dirección** única que se conecta dentro de algún lugar de la **Ciudad** (*wifi*, 5G, radio, satélite, fibra óptica, etc.) para comunicar con alguna zona específica de la **Nube** (Google, YouTube, Instagram, TikTok, Facebook, Amazon, Microsoft... etc.).

Comunicación algorítmica interactiva y persuasiva

Acabamos de ver algunas de las consecuencias sociales y medioambientales que conlleva la producción y utilización de los dispositivos digitales que se conectan al entramado cibernético para establecer intercambios comunicativos e informacionales. Pero, ¿qué es lo que verdaderamente ocurre al otro lado de la pantalla cuando interactuamos con las diferentes interfaces de aplicaciones en nuestro dispositivo favorito? ¿Con quién se comunica el *smartphone* que llevamos en la mano mientras nos comunicamos con algún ser querido usando una red social? ¿Y qué ocurre cuándo lo llevamos en el bolsillo caminando por la calle?

Para comprender desde un punto de vista crítico lo que ocurre cuando se utilizan estos sistemas de comunicación digital avanzada, Álvarez Rufs (2023) propone un modelo de comunicación algorítmica interactiva y persuasiva que implica la generación de, al menos, tres niveles de realidad diferentes dentro del entramado cibernético.

Imagen 2. Modelo de comunicación algorítmica interactiva y persuasiva



Fuente: Álvarez Rufs (2023)

En un primer nivel, encontramos a la persona usuaria que observa la realidad que le ofrece el algoritmo a través de la interfaz. La persona es al mismo tiempo receptora de estímulos y emisora de comportamientos en base a sus interacciones con la interfaz.

En un segundo nivel de realidad, el algoritmo observa y registra los comportamientos de la persona usuaria a través de la interfaz. El algoritmo utiliza la retroalimentación para ajustar la realidad que ofrece a la persona usuaria con relación al tipo de comportamiento deseado por parte de la entidad que administra el sistema algorítmico.

En un tercer nivel de realidad, la entidad administradora actúa como un observador universal que controla la finalidad del algoritmo y la modificación de los comportamientos de las personas.

De esta forma, de una manera interactiva y persuasiva, las personas administradoras de las redes sociales de las plataformas digitales controlan la retroalimentación que ofrecen los algoritmos en las interfaces de las personas usuarias, generando esos tres niveles diferentes de realidad.

Desde un punto de vista cibernético, se produce la observación de la observación de la observación, es decir, las personas usuarias observan la interfaz que ofrece el algoritmo, el algoritmo observa y registra la reacción de las personas usuarias, y retroalimenta y genera una nueva realidad en la interfaz que sirve para mantener las conductas que desea la entidad que administra el sistema, quien a su vez observa y programa según su propios objetivos la reacción del algoritmo ante los diferentes comportamientos observados.

Estrategias Eco-Educomunicativas contra los discursos de odio

Para afrontar posibles estrategias eco-educomunicativas contra los discursos de odio, debemos tener en cuenta la segunda parte de la definición de educomunicación, que se refiere a la formación del sentido crítico, inteligente, frente a los procesos comunicativos y sus mensajes para descubrir los valores culturales propios y la verdad. Pero, ¿qué es el discurso de odio?

No hay consenso, no hay una unanimidad a la hora de definir el discurso de odio. Según la definición de Naciones Unidas (s.f.), el discurso de odio es cualquier tipo de comunicación, ya sea oral o escrita, o también comportamiento, que ataca o utiliza un lenguaje peyorativo o discriminatorio en referencia a una persona o grupo en función de lo que son, en otras palabras, basándose en su religión, etnia, nacionalidad, color, ascendencia, género u otra forma de identidad. El discurso de odio posee tres características esenciales (Naciones Unidas, s.f.):

1. Se materializa en cualquier forma de expresión (imágenes, ilustraciones, memes, gestos, símbolos, etc.).
2. Es discriminatorio (sesgado, fanático e intolerante) o peyorativo (basado en prejuicios, despectivo o humillante) hacia un individuo o grupo de individuos.

3. Está centrado en factores de identidad reales o percibidos de un individuo o grupo (religión, etnia, color, ascendencia, género) además de otras muchas características (idioma, origen económico o social, discapacidades, estado de salud, edad, orientación sexual, etc.).

La automatización de diferentes discursos de odio mediante la utilización de sistemas de inteligencia artificial generativa (IAG) constituye un fenómeno análogo a la producción de desinformación y noticias falsas. El peligro de estas prácticas radica en su escalabilidad exponencial, ya que con estos sistemas resulta posible producir diferentes variantes de discursos de odio que se contextualizan en forma de múltiples lenguajes y registros discursivos masivos.

Para poder afrontar, en el contexto del fenómeno de la desinformación, los discursos de odio que se dan contra las clases oprimidas por parte de las clases privilegiadas, interesa tener en cuenta la pedagogía antifascista que propone Díez Gutiérrez (2022), la cual incluye, asimismo, diferentes tipos de pedagogía que pueden resultar útiles en nuestra práctica eco-educomunicativa: Pedagogía Crítica; Pedagogía en Valores; Pedagogía Laica; Pedagogía de la Memoria; Pedagogía Feminista; Pedagogía del Apoyo Mutuo; Pedagogía Inclusiva; Pedagogía Decolonial; Pedagogía Intercultural y Antirracista; Pedagogía Lenta; Pedagogía Digital Crítica; Pedagogía de la Evaluación Democrática; Pedagogía de lo Esencial; Pedagogía de la Igualdad; Pedagogía Ecosocial del Decrecimiento; Pedagogía Democrática; Pedagogía de la Desobediencia; Pedagogía del Compromiso; Pedagogía del Bien Común.

Para diseñar estrategias útiles para combatir la desinformación y los discursos de odio, también resulta interesante tener en cuenta el concepto de *analfabetismo cívico* que propone Henry Giroux (2018), entendido como una negación del individuo a actuar desde una posición de consideración, juicio informado y agencia crítica. El analfabetismo cívico se convierte en una forma de despolitizar a las personas que no

saben, o no son capaces de desarrollar juicios informados, analizar relaciones complejas y recurrir a una variedad de fuentes, con la finalidad de comprender cómo funciona el poder y cómo pueden moldear las fuerzas que influyen en sus vidas.

En este sentido, es imprescindible atender y trabajar las *actitudes personales* desde un punto de vista eco-educomunicativo, ya que la actitud lleva a las personas a reaccionar de una determinada forma ante las diferentes realidades a las que tienen acceso dentro del entramado cibernético. Habitualmente, resulta más fácil aprender conocimientos y desarrollar habilidades que generar un cambio en las actitudes de las personas. Por tanto, un cambio de actitudes es fundamental para que la ciudadanía supere el analfabetismo cívico y actúe.

También debemos considerar la definición de *pensamiento crítico de sentido fuerte* o en un sentido estricto, propuesta por Paul (1992), que consiste en aplicar los procesos y principios de evaluación de argumentación a las propias creencias y compromisos, particularmente hacia las propias creencias profundas, y eliminar cualquier prejuicio hacia estas propias creencias, volviendo el pensamiento crítico hacia dentro, si así se desea.

Asimismo, como estrategia para combatir los discursos de odio y la desinformación, debemos tener en cuenta el concepto de *empatía intelectual* que propone Bowell (2018), según el cual, la razón y la comprensión deben complementarse con la emoción y la experiencia para que podamos conocer en el sentido más amplio posible. Para ello, debemos conocernos a nosotras mismas y nosotros mismos, y también conocer todo lo que podamos sobre las circunstancias de otras personas, particularmente las personas cuyas circunstancias son diferentes a las nuestras.

Pero hay que ir más allá. El concepto de *reciprocidad asimétrica*, también propuesto por Bowell, implica que, en lugar de tratar de imaginarnos a nosotras mismas o a nosotros mismos en la posición de la otra persona, tenemos que comprometernos a aprender tanto como podamos sobre las realidades vividas por otras personas a partir de sus

propios testimonios sobre cómo experimentan, negocian y viven sus vidas; cómo viven su género, clase, discapacidad, sexualidad, cultura y religión.

Esto requiere que escuchemos adecuadamente y tomemos en serio el testimonio de otros para aprender sobre sus experiencias vividas y sobre cómo deben ser ellos en este mundo. En este punto, recordamos a Mario Kaplún, quien refiere que una de las cualidades más importantes del comunicador, de la comunicadora, es saber *escuchar*.

Esta *reciprocidad asimétrica* nos lleva más allá de ponernos en los zapatos de alguien. No se trata de imaginarnos como la otra persona, sino de intentar que nosotras mismas vivamos y experimentemos la realidad de la otra persona. No se trata de ponernos en su lugar, se trata de ser esa misma persona, comprenderla desde su propia interioridad.

Una estrategia útil para combatir la desinformación y los discursos de odio es generar historias con nuevos discursos. Las historias y otros textos, películas, obras de teatro, artes visuales, canciones, memes, *reels*, etc., también conllevan potencial para la transformación sociocultural a través de su capacidad para producir cambios en el imaginario social, ofreciendo destellos de diferentes vidas y formas de ser, ya sean vidas reales o narraciones ficticias. Giroux (2018) nos pide desarrollar un discurso y unas prácticas pedagógicas que conecten la lectura de la palabra y la lectura del mundo, de manera que mejoren las capacidades de las personas jóvenes como agentes críticos y ciudadanía comprometida.

Rubio (2017) propone incrementar la presencia y eficacia de los filtros de verificación, fortalecer la formación y el pensamiento crítico de la ciudadanía, y reconstruir la confianza ciudadana en los gobiernos, en los medios de comunicación, en la justicia y en la ciencia. Según McIntyre (2018), sobre todo relacionado con las noticias falsas, es imprescindible verificar desde múltiples fuentes, evaluar la credibilidad de la fuente, buscar la fecha de publicación, evaluar la experiencia del autor en el tema, preguntar ¿coincide esto con mi conocimiento previo? o ¿esto parece realista?

Ball (2018) propone hacernos algunas preguntas ante las noticias que aparecen en los diferentes medios de comunicación y redes sociales: ¿De dónde sale esta noticia? ¿Cómo está redactada y diseñada? ¿Quién la firma? ¿Qué emociones te provoca? ¿De dónde salen las fotografías? ¿Es coherente en el tiempo lo que cuenta la noticia? ¿Serán ciertos los datos? ¿Intuyo algún interés partidista o ideológico? ¿Se ve claramente que es una broma? ¿Qué gano compartiendo la información?

Por su parte, Amorós (2018) nos pide hacer estallar nuestra propia burbuja. Esto implica enfrentarnos a opiniones contrarias a las nuestras, opiniones diferentes, disidentes. También nos pide involucrar al *Sistema Dos*. Debemos tener en cuenta que Kahneman (2011) resalta la importancia de implicar al *Sistema Dos* de pensamiento que evalúa cuidadosamente la información antes de compartirla, antes que al *Sistema Uno*, que es el que utilizamos habitualmente y que no se toma tal tipo de esfuerzos, de manera que nos lleva a compartir de manera adictiva. También resulta conveniente aprender algunas estadísticas, considerar a las narrativas que creemos con el mismo escepticismo como las que no creemos, y tratar de no sucumbir al pensamiento conspirativo (Amorós, 2018). En relación con los medios de comunicación Amorós propone: (1) no crear ni publicar más *fake news*, (2) no usar el poder del periodismo por interés propio ni de nadie, (3) ser independientes, (4) primar los hechos y alejarlos de toda opinión, y (5) abandonar el periodismo de declaraciones y apostar por el periodismo de investigación.

Atendiendo a Levitin (2018), debemos tener en cuenta que:

- Hay algunas cosas que sabemos.
- Hay algunas cosas que no sabemos.
- Hay algunas cosas que sabemos, pero no somos conscientes de ello.
- Hay algunas cosas que no sabemos, y no somos conscientes de ello.

En conclusión, necesitamos *humildad*. Humildad para reconocer que no lo sabemos todo. Humildad para reconocer que debemos tener en cuenta las opiniones de los demás, los puntos de vista de otras personas para enriquecernos. Humildad para poder combatir de manera eficaz la desinformación y los discursos de odio.

Ejercicio práctico: Una mirada al discurso de odio desde la interseccionalidad.

La interseccionalidad, concepto popularizado por Kimberlé Cresnhaw en el año 1989, se refiere a la interacción de diferentes categorías de diferenciación social que afectan a la vida de las personas en forma de opresión o privilegio. Algunos ejemplos de estas categorías son:

- **Androcentrismo** (hombres versus mujeres).
- **Racismo** (raza blanca versus otros colores de piel).
- **Elitismo** (con estudios versus sin estudios).
- **Edadismo** (joven versus anciano).
- **Clasismo** (clase social alta versus clase social baja).
- **Heterosexismo** (heterosexual versus lesbiana, gay, bisexual).
- **Capacitismo** (personas capacitadas versus personas discapacitadas).
- **Eurocentrismo** (modelo europeo versus modelo no europeo)
- **Género** (cisgénero versus transgénero).
- **Apariencia** (atractivo/a versus no atractivo/a).
- **Religión** (religión mayoritaria versus religión minoritaria).

Para concluir el capítulo, os proponemos realizar una reflexión personal a partir de la interseccionalidad y desde el pensamiento crítico. Una introspección, una mirada hacia adentro, dirigida a vuestras propias emociones personales y creencias profundamente arraigadas, con el objetivo de reconocer los momentos en los que habéis gozado de algún privilegio respecto a otras personas, y también aquellos momentos en los que habéis sufrido opresión en vuestro propio contexto personal, social, laboral, cultural, etc.

A continuación, una vez realizada la reflexión personal, debéis identificar relaciones entre las opresiones o privilegios que hayáis reconocido previamente y los posibles discursos de odio que pueden llegar a generar en vuestro propio contexto personal.

Tened en cuenta que las creencias profundamente arraigadas pueden mantenerse de forma apasionada, se pueden defender dogmáticamente y pueden jugar un papel fundamental en la forma en que representamos el mundo. Forman parte de nuestro marco simbólico y de nuestra forma de ser en el mundo con los y las demás (Bowell, 2018), influyendo en nuestras acciones sociales y políticas.

Referencias

Álvarez Rufs, M. (2022). Posverdad y Algoritmos en Sociedades Cibernéticas: Un Mapeo de los Nuevos Territorios Educomunicativos. En L. R. Romero Domínguez y N. Sánchez Gey, (coords.), *Sociedad digital, comunicación y conocimiento: retos para la ciudadanía en un mundo global* (pp. 111-131). Dykinson.

Álvarez-Rufs, M. (2023). Los Algoritmos del Capitalismo de la Vigilancia como Medios de Comunicación de Masas: Un Modelo de Comunicación Algorítmica Interactiva y Persuasiva. *Revista De La Asociación Española De Investigación De La Comunicación, 10* (Especial), 108-130.

Aparici, R., Álvarez Rufs, M. & Gómez Mondino, P. (2024a). Eco-Educomunicación y Colonización Tecnológica. En: Aparici, R., Álvarez Rufs, M. & Gómez Mondino, P. *Hoy es mañana. De Mario Kaplún a la Educomunicación del siglo XXI* (pp. 313-326). CIESPAL.

Aparici, R., Álvarez Rufs, M. & Gómez Mondino, P. (2024b). Colonización Tecnológica, Automatización de la colonización y Eco-Educomunicación. *Chasqui: Revista Latinoamericana de Comunicación*, 157, 19-33.

Amorós, M. (2018). *Fake News. La verdad de las noticias falsas*. Plataforma Actual.

Ball, J. (2017). *Post-Truth. How Bullshit Conquered the World*. Biteback Publishing.

Bowell, T. (2018). Changing the World One Premise at a Time: Argument, Imagination and Post-truth. En: Peters, M.A., Rider, S., Hyvönen, M., Besley, T. (2018). *Post-Truth, Fake News. Viral Modernity & Higher Education*. Springer.

Bratton, B. H. (2015). *The Stack. On Software and Sovereignty*. The MIT Press.

Crawford, K. (2023). *Atlas de IA. Poder, política y costes planetarios de la inteligencia artificial*. Ned Ediciones.

Giroux, H.A. (2018). What Is the Role of Higher Education in the Age of Fake News? En: Peters, M.A., Rider, S., Hyvönen, M., Besley, T. (2018): *Post-Truth, Fake News. Viral Modernity & Higher Education*. Springer.

Gutiérrez, E. J. D. (2022). *Pedagogía antifascista: construir una pedagogía inclusiva, democrática y del bien común frente al auge del fascismo y la xenofobia*. Ediciones Octaedro.

Kahneman, D. (2011). *Thinking Fast and Slow*. Farrar, Straus and Giroux.

Levitin, D. J. (2017). *Weaponized Lies. How to Think Critically in the Post-Truth Era*. Dutton.

McIntyre, L. (2018). *Post-Truth.* (The MIT Press Essential Knowledge series). MIT Press.

Naciones Unidas. (s.f.). *Entender qué es el discurso del odio*. https://www.un.org/es/hate-speech/understanding-hate-speech/what-is-hate-speech#

Paul, R. (1992). Teaching critical reasoning in the strong sense: Getting behind worldviews. En: R. A. Talaska (Ed.), *Critical reasoning in contemporary culture* (pp. 135–156). State University of New York Press.

Rubio, D. (2017). La política de la posverdad. *Estudios de política exterior* 176. pp. 58-67.

VV.AA. (1992). *Educación para la comunicación*. CENECA-UNICEF-UNESCO.

Epílogo

La verdad sitiada

David García-Marín
Universidad Rey Juan Carlos

La desinformación no puede ser comprendida únicamente como una desviación individual o una anomalía comunicativa, sino como un fenómeno estructural que se inscribe en dinámicas sociotécnicas complejas. La desinformación se alimenta de factores como la polarización política, la desconfianza institucional, la precarización del periodismo, la arquitectura de incentivos de las plataformas digitales y, más recientemente, la capacidad de los modelos de IA generativa para producir contenidos verosímiles pero falsos a gran escala. La *infoesfera* contemporánea está caracterizada por una sobreabundancia de datos, una aceleración de los flujos comunicativos y una creciente dificultad para distinguir entre lo verdadero y lo falso, lo auténtico y lo manipulado. La IA actúa como catalizador de estas transformaciones, amplificando tanto las posibilidades de creación como los riesgos de distorsión.

Frente a este panorama, el *fact-checking* ha emergido como una práctica crucial para la defensa de la verdad y la integridad informativa. Sin embargo, esta práctica enfrenta múltiples tensiones en la era algorítmica. Por un lado, se encuentra la necesidad de escalar los

procesos de verificación para hacer frente al volumen y la velocidad de la desinformación. Por otro, persiste el desafío de mantener estándares éticos, metodológicos y epistemológicos rigurosos en un entorno donde la presión por la inmediatez puede comprometer la calidad de las verificaciones.

Asimismo, resulta fundamental la importancia de promover una alfabetización digital crítica que no se limite a habilidades técnicas, sino que fomente la capacidad de análisis, el pensamiento reflexivo y la conciencia sobre los mecanismos de producción y circulación de la información. En este sentido, la educación mediática debe ser concebida como una herramienta de empoderamiento ciudadano frente a las asimetrías del poder informativo.

Vivimos en un mundo complejo. La relación entre verdad, conocimiento y poder en el contexto contemporáneo está marcada por la hegemonía del neoliberalismo como racionalidad política, económica y cultural. Lejos de ser una mera ideología económica, el neoliberalismo ha demostrado ser una forma de gubernamentalidad que penetra en los tejidos más íntimos de la subjetividad, reconfigurando no solo las instituciones sociales, sino también las condiciones de posibilidad del saber y la verdad. En tiempos donde la verdad parece haber sido desplazada por la eficacia, la performatividad y la rentabilidad, se vuelve urgente repensar los fundamentos mismos de nuestra relación con el conocimiento y la verdad. ¿Qué significa decir la verdad en un mundo donde los discursos son constantemente instrumentalizados? ¿Cómo resistir a la captura neoliberal del saber sin caer en relativismos paralizantes o en esencialismos dogmáticos?

Uno de los efectos más insidiosos del neoliberalismo ha sido la institucionalización de formas sistemáticas de ignorancia. A través de la privatización del conocimiento, la precarización de la investigación crítica y la subordinación de la educación a lógicas de mercado se ha erosionado la posibilidad de construir saberes emancipadores. Esta epistemología neoliberal no solo selecciona qué se puede conocer, sino también quién tiene derecho a conocer y en qué condiciones.

La proliferación de *fake news*, la desinformación algorítmica y la polarización mediática no son fenómenos aislados, sino síntomas de una crisis epistemológica más profunda. En este contexto, el conocimiento se convierte en un bien escaso, fragmentado y jerarquizado, donde la verdad pierde su valor intrínseco y se convierte en una mercancía más, sujeta a las leyes de la oferta y la demanda.

El neoliberalismo no solo transforma las estructuras sociales, sino también las formas de subjetividad. El sujeto neoliberal es un emprendedor de sí mismo, un gestor de su capital humano, un consumidor de verdades a la carta. Esta subjetividad autogestionada tiende a rechazar toda forma de verdad que no se traduzca en utilidad inmediata o en ganancia simbólica. En este marco, la verdad deja de ser una búsqueda ética y se convierte en una estrategia de posicionamiento.

Frente a esta lógica, se vuelve urgente recuperar una concepción ética de la verdad, entendida no solo como certeza absoluta, sino también como compromiso con la justicia, la transparencia y la responsabilidad. La verdad, en este sentido, no es un dato, sino una práctica, una forma de habitar el mundo, de interpelar al poder y de construir comunidad. Como señalaba Michel Foucault, decir y buscar la verdad puede ser un acto de coraje, una forma de resistencia frente a los dispositivos de normalización.

El desafío que se nos presenta es, por tanto, doble: por un lado, desmontar las formas neoliberales de producción del saber; por otro, construir alternativas epistemológicas que permitan reconfigurar nuestra relación con la verdad. Esto implica repensar las instituciones del conocimiento —universidades, medios de comunicación, centros de investigación—, pero también las prácticas cotidianas de producción y circulación del conocimiento.

Una epistemología crítica de la verdad debe ser, ante todo, una epistemología política. No se trata solo de conocer el mundo, sino de transformarlo. Esto requiere no solo una pedagogía eco-educomunicativa como la que se propone en el libro, sino también una pedagogía de la sospecha y de la esperanza: una apuesta por formas

de conocimiento que no se limiten a denunciar, sino que también propongan, imaginen y construyan.

En tiempos neoliberales, la verdad se encuentra sitiada. No por la mentira en sí misma, sino por la banalización de su valor, por la indiferencia ante sus consecuencias, por la fragmentación de los marcos comunes de sentido. Frente a esta crisis, no basta con reivindicar la verdad como un ideal abstracto; es necesario encarnarla en prácticas concretas, en formas de vida y en comunidades de sentido.