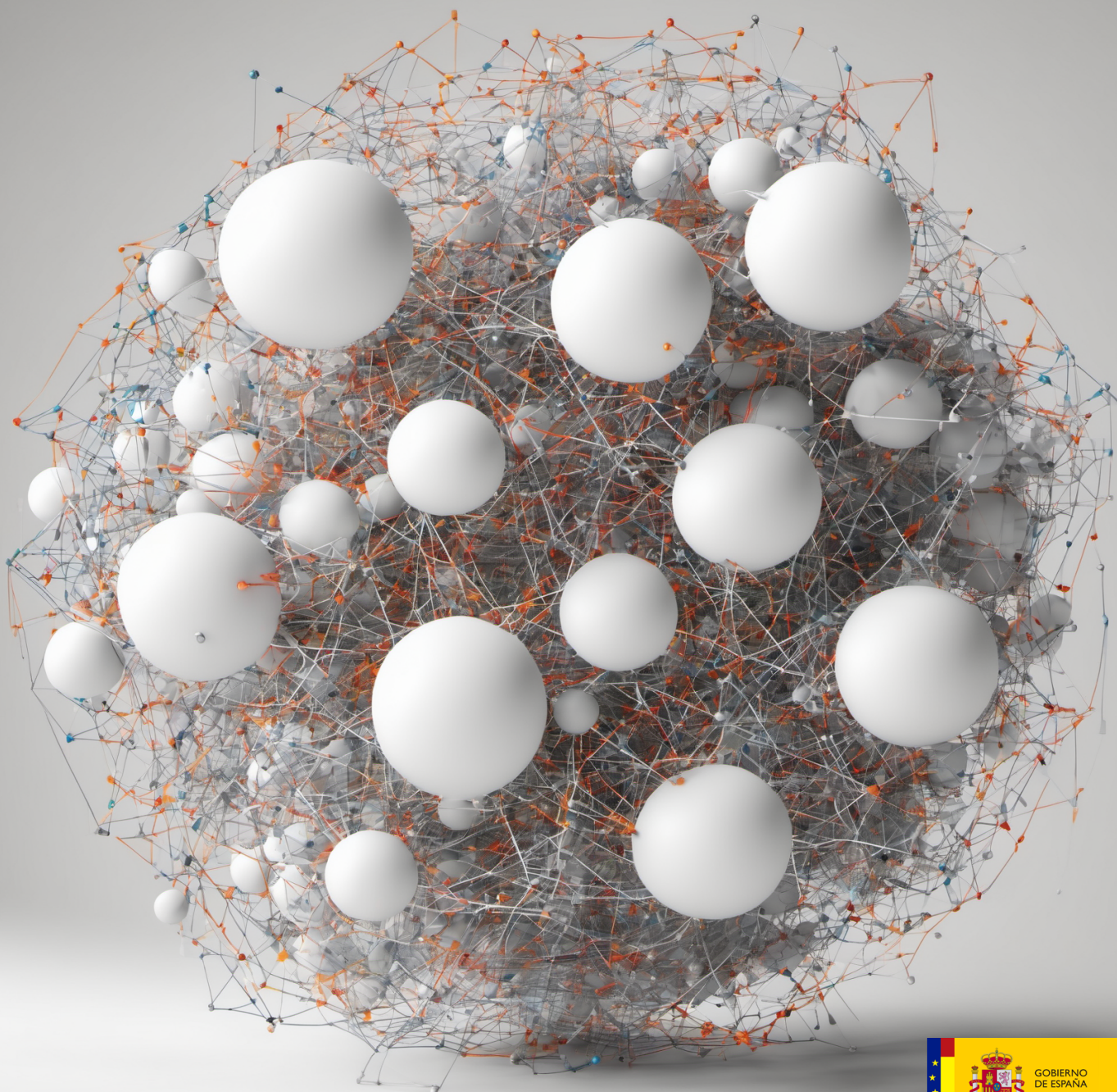


FORO CONTRA LAS CAMPAÑAS DE DESINFORMACIÓN EN EL ÁMBITO DE LA SEGURIDAD NACIONAL

TRABAJOS 2023



Catálogo de publicaciones de la Administración General del Estado

<https://cpage.mpr.gob.es>

Edita:



© Autor y editor,

NIPO (edición impresa): 089-23-041-3

NIPO (edición on-line): 089-23-042-9

DL: M-33637-2023

Fecha de edición: noviembre 2023

Imprime: MASQUELIBROS, S.L.

FORO CONTRA LAS CAMPAÑAS DE DESINFORMACIÓN EN EL ÁMBITO DE LA SEGURIDAD NACIONAL

TRABAJOS 2023

Los expertos participantes en los Grupos de Trabajo lo son a título personal y no a título institucional. Por lo tanto, sus opiniones y recomendaciones no representan ni comprometen a las instituciones a las que pertenecen.

El resultado de los trabajos es producto de un ejercicio de reflexión colectivo, si bien, no tiene por qué representar la opinión individual de todos los participantes, quienes no necesariamente comparten todas las conclusiones o propuestas.

PRESENTACIÓN

La Estrategia de Seguridad Nacional de 2021 (ESN21) recogió, por primera vez, las campañas de desinformación como un riesgo y amenaza para la Seguridad Nacional, en línea con la Unión Europea y la comunidad internacional occidental. Asimismo, la ESN21 destaca la colaboración público privada para detectar y hacer frente a las citadas campañas. La desinformación organizada no es un fenómeno nuevo, pero la revolución digital la ha llevado a una escala y velocidad de difusión sin precedentes que la han convertido en un riesgo para la seguridad de los Estados democráticos y sus procesos electorales.

Afrontar este riesgo en el marco de un ecosistema digital y dinámicas de comunicación complejas y cambiantes requiere, más que nunca, la colaboración de la administración pública, la sociedad civil, el mundo académico y la industria privada. Esta necesidad de fomento de la colaboración público-privada, recogida también en el Plan de Acción para la Democracia Europea, se basa en garantizar no solo un conocimiento más profundo de la amenaza sino idear y consensuar los mecanismos más eficientes para hacerle frente desde el más escrupuloso respeto a los derechos y libertades fundamentales consagrados en nuestra Carta Magna.

Conscientes de esta necesidad, el Sistema de Seguridad Nacional promovió en 2020 una iniciativa de cooperación público privada con más de 50 expertos que culminó en la publicación del libro “Lucha contra las campañas de desinformación en el ámbito de la seguridad nacional: propuestas de la sociedad civil”. El éxito de este proyecto y la necesidad de construir un espacio de debate permanente que permitiera evaluar el riesgo cambiante de esta amenaza y analizar la eficacia de las posibles medidas para limitar sus efectos, motivó la creación en mayo de 2021, por parte del Consejo de Ministros, del Foro contra las campañas de desinformación en el ámbito de la seguridad nacional.

Esta obra es el fruto de las iniciativas promovidas por este Foro que está constituido por seis representantes de la Administración Pública y diez representantes de la sociedad civil, el periodismo, la academia y la industria de comunicación.

El libro contiene siete trabajos que abordan el fenómeno de las campañas de desinformación desde varias ópticas que permiten analizar el progreso y nos aportan una visión prospectiva de esta amenaza.

La comunicación pública, la labor de medios de comunicación y de los periodistas juegan un papel fundamental en los sistemas democráticos, que se sustentan en que los ciudadanos tengan acceso a la información veraz para poder tomar sus decisiones en libertad. Su participación activa es clave para incrementar el conocimiento situacional y la resiliencia de todos los ciudadanos frente a esta amenaza, lo que, a su vez, contribuirá a promover, potenciar y sensibilizar sobre la cultura de Seguridad Nacional en aras a favorecer la concienciación sobre todos los esfuerzos que requiere la garantía de la Seguridad Nacional del Reino de España. Además, la comunicación y la transparencia son herramientas esenciales para limitar los efectos de la desinformación.

Asimismo, uno de los puntos de partida para abordar cualquier amenaza es el conocimiento de la misma mejorando así la conciencia situacional. En este plano, el despliegue de campañas de desinformación en el marco de la invasión ilegítima de Rusia a Ucrania supone un caso de análisis esencial para conocer las técnicas que actores estatales utilizan para proyectar las campañas de desinformación tanto a nivel doméstico como en el ámbito internacional y reflejan el potencial que adquiere esta amenaza cuando se enmarca en el seno de estrategias híbridas más amplias.

Por otro lado, promover la investigación académica y científica sobre este fenómeno, fomentar la cooperación multidisciplinar, mediante sinergias con entidades de la sociedad civil y del sector industrial, como base para poder avanzar en el conocimiento y la implicación que las campañas de desinformación pueden tener en los procesos de radicalización violenta. Debemos analizar las implicaciones que los avances en Inteligencia Artificial pueden tener para la difusión y para la detección de estas amenazas. También necesitamos desarrollar mecanismos comunes para analizar las campañas de desinformación, que fomenten la cooperación tanto entre entidades nacionales como internacionales.

Por último, en los últimos años, a nivel de la UE se han realizado esfuerzos para establecer unos principios corregulatorios para que las plataformas digitales limiten ciertos comportamientos nocivos en sus servicios, por ejemplo, el comportamiento inorgánico que favorece la difusión de desinformación. Iniciativas como el Código de buenas prácticas contra la desinformación o el Reglamento de servicios digitales necesitan una continua evaluación de su implementación a fin de analizar la validez de los mismo para los fines que se persiguen.

Agradezco a todos los expertos que, de forma desinteresada, han contribuido con su experiencia y su conocimiento a la elaboración de estos trabajos, que suponen un importante punto de apoyo para avanzar en la lucha contra las campañas de desinformación.

No puedo concluir sin expresar mi profundo agradecimiento a Aurelio Martín González, vocal del Foro designado por la Federación de Asociaciones de Periodistas de España (FAPE), quien contribuyó con su excelsa dedicación y compromiso a la redacción de este libro, especialmente a su capítulo primero. Quienes tuvimos el privilegio de conocer al periodista vocacional y a la persona, sentimos que su ausencia deja un vacío imposible de llenar, le recordamos con cariño y gratitud.

Miguel Ángel Ballesteros Martín
Director del Departamento de Seguridad Nacional
Presidencia del Gobierno

ÍNDICE

CAPÍTULO 1: METODOLOGÍA Y BUENAS PRÁCTICAS PARA LA DETECCIÓN Y VERIFICACIÓN DE LA DESINFORMACIÓN Y PARA LA RESPUESTA EN EL PLANO DE LA COMUNICACIÓN.....	12
INTRODUCCIÓN	15
PARTE 1: IDENTIFICAR Y DETECTAR LA DESINFORMACIÓN	18
CÓMO DETECTAR CAMPAÑAS DE DESINFORMACIÓN	18
Contexto	18
La importancia del trabajo en red	19
La escucha social.....	19
HERRAMIENTAS PARA IDENTIFICAR UNA PÁGINA DE DESINFORMACIÓN	21
LA IA COMO POTENCIADOR DE LA AMENAZA Y RESPUESTA	23
Utilización de la IA en campañas de desinformación.....	23
Utilización de la IA para responder a las campañas	24
EL PAPEL DE LAS PLATAFORMAS DE VERIFICACIÓN. ORGANIZACIONES DE VERIFICACIÓN: DEFINICIÓN, TIPOLOGÍA Y METODOLOGÍA	26
Organizaciones de verificación	26
Los falsos verificadores.....	27
PARTE 2: RESPONDIENDO A LA DESINFORMACIÓN	28
¿CÓMO RESPONDER A LAS CAMPAÑAS DE DESINFORMACIÓN Y EN QUÉ CASOS?	28
Identificación del nivel de riesgo y amenaza	28
Incrementar la concienciación y formación sobre esta amenaza en el seno de la propia organización	29
Establecer mecanismos de alerta temprana	30
Establecer procedimientos y protocolos para responder	30
Generar confianza a través de una comunicación estratégica a largo plazo	33

VACUNAS CONTRA LA DESINFORMACIÓN	35
Desinformación y conducta humana	35
Teoría de la inoculación y vacunas contra la desinformación.....	36
Conceptualización de la teoría de la inoculación	37
Principales iniciativas y actividades	37
Limitaciones y conclusiones	38
EL PAPEL DE LA ALFABETIZACIÓN MEDIÁTICA EN EL ACTUAL SISTEMA INFORMATIVO: PRINCIPALES INICIATIVAS Y EXPERIENCIAS	39
Instituciones privadas	39
Instituciones públicas	41
Instituciones del tercer sector	43
CAPÍTULO 2: ESTUDIO DE LA DESINFORMACIÓN RUSA A NIVEL INTERNACIONAL..	46
INTRODUCCIÓN	49
RUSIA Y LA “GUERRA DE LA INFORMACIÓN”	50
EL ECOSISTEMA INTERNACIONAL DE MEDIOS DE DESINFORMACIÓN DEL KREMLIN	54
ALGUNAS CONSIDERACIONES SOBRE LA DESINFORMACIÓN RUSA CONTRA UCRANIA (2014-2022)	57
UN ESTUDIO DE CASO SOBRE SUPLANTACIÓN DE MEDIOS E INSTITUCIONES .	60
CAPÍTULO 3: EFECTOS PSICOLÓGICOS, RADICALIZACIÓN Y AMPLIFICACIÓN DE LOS CONFLICTOS SOCIALES CAUSADOS POR LA INJERENCIA Y MANIPULACIÓN EXTRANJERA DE LA INFORMACIÓN: ANÁLISIS Y PREVENCIÓN	66
INTRODUCCIÓN	69
EL CONTEXTO INTERNACIONAL Y EL <i>SHARP POWER</i>	72
Desinformación y radicalización violenta.....	74
LA DESINFORMACIÓN Y LAS DINÁMICAS GRUPALES	78
Continuidad y cambio en el entorno digital.....	79

La construcción y descalificación del adversario	81
La presión por la conformidad en el seno del grupo	84
El concepto de verdad y las teorías de la conspiración	85
VULNERABILIDAD A LA DESINFORMACIÓN Y LA PROPENSIÓN A CREERLA	89
El sesgo de confirmación y la exposición selectiva	90
El heurístico de disponibilidad	91
DESTREZAS PARA DETECTAR LA DESINFORMACIÓN	93
Programas de alfabetización mediática	93
Verificación de información	94
Diversidad de fuentes de información y autoregulación profesional	95
Educación cívica	96
ESTRATEGIAS DE RESISTENCIA A LA DESINFORMACIÓN	97
Prevención y respuesta frente al <i>Sharp Power</i> y la desinformación	97
Los efectos no deseados de la censura	98
La reactancia psicológica	99
Teoría de la inoculación	99
CONCLUSIÓN	102
CAPÍTULO 4: MAPA DE LAS CAPACIDADES DE INVESTIGACIÓN EN MATERIA DE DESINFORMACIÓN EN LAS UNIVERSIDADES Y CENTROS DE INVESTIGACIÓN ESPAÑOLES	116
INTRODUCCIÓN	119
APROXIMACIÓN A LA INVESTIGACIÓN SOBRE DESINFORMACIÓN REALIZADA POR LOS ACADÉMICOS DE UNIVERSIDADES Y CENTROS DE INVESTIGACIÓN ESPAÑOLES	121
Análisis de los artículos científicos publicados sobre desinformación por investigadores e investigadoras en España	122
Análisis y evolución de los artículos científicos publicados por investigadores de instituciones españolas	129

ANÁLISIS DE LOS GRUPOS DE INVESTIGACIÓN ESPECIALIZADOS EN DESINFORMACIÓN EN ESPAÑA	133
Grupos de investigación que abordan la desinformación en el Consejo Superior de Investigaciones Científicas.....	133
Grupos de investigación que abordan la desinformación en las universidades españolas	135
ANÁLISIS DE LOS PROYECTOS DE INVESTIGACIÓN FINANCIADA SOBRE DESINFORMACIÓN	148
Proyectos de investigación financiados por la Agencia Estatal de Investigación sobre desinformación	148
Proyectos de investigación sobre desinformación financiados por entidades privadas	150
EL PAPEL ESTRATÉGICO EN EL ÁMBITO NACIONAL DE LAS UNIVERSIDADES Y CENTROS DE INVESTIGACIÓN EN LA LUCHA CONTRA LAS CAMPAÑAS DE DESINFORMACIÓN.	153
Mecanismos europeos de colaboración y cooperación de las Universidades en iniciativas para la lucha contra la desinformación	153
El papel de la universidad y centros de investigación para un mejor conocimiento de las campañas de desinformación	160
Potenciar el trabajo académico y su accesibilidad a la sociedad en su conjunto	161
DISCUSIÓN Y CONCLUSIONES.....	164
CAPÍTULO 5: TÉCNICAS, METODOLOGÍAS Y PROSPECTIVA IA PARA COMBATIR LA DESINFORMACIÓN.	168
INTRODUCCIÓN	171
ESTUDIO Y MODELADO DE LAS CAMPAÑAS DE DESINFORMACIÓN	173
CARACTERIZACIÓN DE REDES Y PROPAGACIÓN DE LA DESINFORMACIÓN.....	176
Propagación de desinformación en redes sociales	176
Ejemplos de difusión de desinformación en Twitter y Telegram	180
EL ROL DE LAS PLATAFORMAS DE VERIFICACIÓN	184

Tecnología y verificación. Hacia los procesos de automatización	184
IA y verificación	185
Perspectivas de futuro y recomendaciones	187
TÉCNICAS Y METODOLOGÍAS DE PROCESAMIENTO DEL LENGUAJE NATURAL	189
Monitorización y verificación de rumores	190
Verificación automatizada	191
Detección de información falsa	192
GENERACIÓN Y DETECCIÓN DE CONTENIDO FALSO:	
FAKE NEWS Y DEEPPAKES	194
Fake Multimodal News	194
Deepfake	195
Detección de <i>deepfakes</i>	197
MODELOS DE LENGUAJE MASIVOS PARA GENERACIÓN DE TEXTOS Y	
DESINFORMACIÓN	200
Detectando (o no) desinformación	200
Generación automática de textos	201
Generación automática de desinformación	202
Detección (automática) de textos generados automáticamente	203
Reflexiones finales	206
CAPÍTULO 6: TÉCNICAS Y METODOLOGÍAS DE CIBERINTELIGENCIA PARA LA	
INVESTIGACIÓN DE OPERACIONES DE INFORMACIÓN Y GUERRA COGNITIVA	212
INTRODUCCIÓN	215
El problema en su contexto	215
La desinformación como amenaza	217
FASE I: ANÁLISIS DE ACCIONES DETECTADAS PARA LA DEFINICIÓN DE	
ELEMENTOS Y ACTORES PARTICIPANTES	219
La componente técnica de las acciones. Técnicas, tácticas y procedimientos	219
La componente económica de las acciones	221
La componente organizativa	222

FASE II: DISEÑO Y DESARROLLO DE UN MODELO DE DETECCIÓN TEMPRANA DE HIPOTÉTICAS ACCIONES DE DESINFORMACIÓN.....	225
Elementos para evaluar	228
Metodología a utilizar de cara a desarrollar un modelo de detección temprana	229
Desarrollo del modelo	230
CONCLUSIONES.....	243
INVESTIGACIONES A FUTURO	244
CAPÍTULO 7: LAS ACTUACIONES DE LAS PLATAFORMAS DE INTERNET EN MATERIA DE DESINFORMACIÓN. DEL CÓDIGO DE BUENAS PRÁCTICAS AL REGLAMENTO DE SERVICIOS DIGITALES.....	250
PRESENTACIÓN	253
PREFACIO: LA REGULACIÓN VOLUNTARIA EN LA UNIÓN EUROPEA.....	254
EL CÓDIGO DE BUENAS PRÁCTICAS EN MATERIA DE DESINFORMACIÓN	256
El Código en su versión inicial de 2018	256
El Código de buenas prácticas en materia de desinformación reforzado (2022).....	261
Referencias específicas a España en los informes de rendición de cuentas de las plataformas firmantes del Código	273
REGLAMENTO DE SERVICIOS DIGITALES	276
Marco general	276
Prestadores concernidos y sus obligaciones	277
Riesgos sistémicos	283
La corregulación como complemento a las obligaciones de los prestadores	285
Sanciones previstas	287
La desinformación en el Reglamento de servicios digitales	287
CONCLUSIONES.....	290
Resumen	290
Propuestas	293

Capítulo 1

METODOLOGÍA Y BUENAS PRÁCTICAS
PARA LA DETECCIÓN Y VERIFICACIÓN
DE LA DESINFORMACIÓN Y PARA
LA RESPUESTA EN EL PLANO DE LA
COMUNICACIÓN



Coordinadores:

Aurelio Martín González

Ministerio de Asuntos Exteriores, Unión Europea y Cooperación - Subdirección General de Comunicación Estratégica, Diplomacia Pública y Redes

Autores y colaboradores:

Blanca Bayo Pérez

Sergio Hernández García

Eva Herrero Curiel

Francisco Marcos Martín Martín

Leticia Rodríguez Fernández

Alfonso Alejandro Sánchez Macián Pérez

Presidencia de Gobierno - Departamento de Seguridad Nacional (DSN)



INTRODUCCIÓN

La dificultad de presentar el problema que supone la desinformación radica en la imposibilidad de delimitar las dimensiones de una amenaza inabarcable que solo somos capaces de intuir. Tan antigua como el ser humano y, a la vez, beneficiada por el continuo desarrollo tecnológico de una sociedad hiperconectada, la desinformación es un arma con potencial para matar, manipular, menoscabar el libre ejercicio de la democracia o extender el odio, el miedo y la rabia, entre otros efectos perniciosos. En estos últimos años hemos visto indudables ejemplos de la capacidad devastadora con la propagación de mentiras que han causado muertos durante la pandemia de la COVID, asaltos a instituciones democráticas en Estados Unidos y Brasil, la difusión de relatos racistas, homófobos y sexistas, o la propaganda bélica durante la guerra de Ucrania.

Lo que hace tan peligroso este fenómeno es su eficiencia. La desinformación funciona porque todos somos vulnerables ante ella. Todos queremos confirmar nuestras ideas y reforzar la sensación de que tenemos razón, mientras creemos que el resto se equivoca, y todos pensamos peor cuando nos atenaza el miedo o nos empuja la ira. A menudo, además, le damos credibilidad al mensaje falso que recibimos porque proviene de un remitente conocido. Esa es la brecha que encuentran los manipuladores para penetrar en nuestro pensamiento. Asimismo, la desinformación, como un parásito, se nutre de sus víctimas. Empujados por los mismos sesgos cognitivos, nos disponemos a compartir a toda prisa los mensajes falsos que hemos admitido sin demasiada reflexión. Este es el mejor combustible para su difusión masiva y es así como llega una viralización que infecta a grupos amplios y crea nuevos damnificados. Esta dinámica se refuerza a sí misma porque, cuanta más gente comparta una idea, más creíble parece.

Con este poder de persuasión, en muchas ocasiones los desinformadores no necesitan un gran esfuerzo para divulgar sus engaños. Aunque con frecuencia los contenidos falsos son toscos o aparentemente ridículos —y están promovidos por internautas que se sirven del anonimato que les proporcionan las redes sociales—, estos contenidos consiguen una repercusión notable. Si la manipulación más burda y barata de producir es dañina, cabe preguntarse qué puede conseguir la de elaboración más sofisticada e impulsada

por grandes actores de la desinformación, capaces de organizar campañas con objetivos estratégicos.

En este sentido, las instituciones europeas llevan años alertando sobre las injerencias extranjeras en los procesos democráticos de los países comunitarios y han señalado a Rusia y China, entre otros Estados, como autores de esos ataques, en una estrategia de guerra híbrida.

España es también objeto de estas campañas de desinformación, como se detectó durante el proceso independentista catalán o se ha puesto de relieve en este año electoral, en el que se han replicado relatos falsos ya utilizados en otros países y con los que se ha pretendido desacreditar la fiabilidad del sistema de votación y el recuento de votos.

A estos retos hay que añadir la vigorosa y alarmante irrupción de la inteligencia artificial (IA), que eleva el poder de convicción de la desinformación a límites hasta hace poco tiempo insospechados. A esto hay que añadir que estas herramientas son cada vez más accesibles para cualquiera que tenga la tentación de propagar un bulo.

Todo ello en un contexto de inestabilidad de las grandes plataformas, con cambios de propiedad, despidos masivos y sorprendentes giros en sus políticas corporativas y de moderación de los contenidos publicados en las redes sociales. Estas grandes plataformas que lideran las principales redes sociales continúan apostando por su crecimiento, tanto en usuarios como en la cantidad de mensajes que difunden por internet, sin que esto haya supuesto un refuerzo proporcional de los mecanismos para limitar y detener la difusión de falsedades, contenidos manipulados y discursos de odio. Esta situación se da en un contexto en el que el flujo de la desinformación ha aumentado de forma progresiva y constante desde hace más de dos décadas.

El crecimiento de la desinformación ha tenido como respuesta la proliferación de equipos de verificadores, ya sea como nuevos medios de comunicación especializados o como departamentos de empresas mayores. Su utilidad en la lucha contra la desinformación se ha demostrado tanto en la refutación de algunas de las falsedades más peligrosas divulgadas en internet como en su actuación para evitar la difusión de contenidos engañosos o descontextualizados por parte de los medios en los que trabajan.

La relevancia de esta labor periodística queda manifiestamente patente desde las semanas que anteceden al inicio, el 24 de febrero de 2022, de la guerra de Ucrania. Este periodo se ha visto marcado por una difusión constante de mensajes y contenidos audiovisuales que buscaban plasmar la falsa narrativa de que Rusia no estaba preparando la invasión del país vecino, sino únicamente unas maniobras militares ordinarias. El inicio del conflicto bélico se caracterizó por la dificultad de los periodistas para informar libremente desde Rusia y acceder a determinadas *zonas calientes* en Ucrania. Esta situación propició la difusión de contenidos audiovisuales sobre la guerra desde las redes sociales, con la consiguiente necesidad de los medios de someter a comprobación los vídeos y fotografías que circulan

por internet antes de emitirlos. Los vídeos descontextualizados han circulado a sus anchas desde las semanas previas al inicio del conflicto bélico y durante toda la contienda, por lo que el trabajo de los equipos especializados en verificación ha sido fundamental para impedir la difusión de contenidos falsos o manipulados por parte de los medios de comunicación.

Por último, experiencias recientes han demostrado el potencial de las campañas de desinformación para menoscabar la confianza en los valores democráticos, sus instituciones e incluso los derechos humanos. Pero también ha quedado de manifiesto el peligro de dichas campañas para el sector financiero y las empresas privadas, y también para las ONG, e incluso el posible riesgo para la integridad de sus trabajadores. Por este motivo, la creciente preocupación por esta amenaza y la necesidad de identificar mecanismos para hacerle frente han crecido también en el sector privado y entre las organizaciones de la sociedad civil.

Este trabajo busca, por un lado, recopilar algunas de las técnicas y buenas prácticas existentes para detectar e identificar las campañas de desinformación y, por otro lado, desarrollar estrategias de comunicación, tanto preventivas como reactivas, para limitar los efectos de esta amenaza.

PARTE 1: IDENTIFICAR Y DETECTAR LA DESINFORMACIÓN

CÓMO DETECTAR CAMPAÑAS DE DESINFORMACIÓN

Contexto

La experiencia de los verificadores en la detección y el análisis de los mensajes falsos difundidos en las redes sociales, aplicaciones de mensajería instantánea y páginas web desinformativas demuestra que quienes promueven esos contenidos engañosos se rigen por la actualidad y aprovechan los estados de ánimo que determinados acontecimientos generan. Así, es habitual que en momentos de incertidumbre e inestabilidad económica o social los desinformadores aprovechen para lanzar sus mensajes. Esto es debido a que en situación de crisis la población es más vulnerable y permeable ante contenidos falsos. Estos contenidos se creen y se difunden con más facilidad en contextos de incertidumbre y cuando los mensajes oficiales no son claros o coherentes entre sí. Un ejemplo claro de ello ocurrió durante la pandemia del nuevo coronavirus. En ese tiempo se pudo observar cómo los relatos falsos se fueron adaptando a las diferentes etapas de propagación y lucha contra la enfermedad. Si en un principio abundaban los mensajes que engañaban acerca del origen de la COVID, estos fueron sucedidos por los que cuestionaban las medidas gubernamentales de confinamiento o el uso de las mascarillas, para posteriormente sembrar dudas sobre la seguridad de las vacunas y, finalmente, generar miedo ante los supuestos efectos perniciosos de estos fármacos en los niños. También se utilizan fechas clave en las que los ciudadanos son más susceptibles de creer ciertos bulos porque consumen más contenidos de una temática determinada, como ocurre con efemérides como el Día Internacional de la Mujer o el inicio de los procesos electorales.

Por otra parte, las instituciones comunitarias han constatado la existencia de campañas de manipulación con gran capacidad nociva, como exponía la Unión Europea en sus planes de acción contra la desinformación y para la democracia europea de 2018 y 2020, respectivamente. Posteriormente, el Parlamento Europeo, en su primer *Informe sobre injerencias extranjeras en todos los procesos democráticos de la Unión Europea*, en particular la desinformación (INGE), aprobado en marzo de 2022 pero elaborado con anterioridad a la invasión de Ucrania, alertaba de que “las pruebas demuestran” que “Rusia y China”, entre otros agentes “malintencionados y autoritarios”, utilizan “la manipulación de la información y otras tácticas de injerencia para interferir en los procesos democráticos en la Unión”, como “parte de una estrategia de guerra híbrida”.

La importancia del trabajo en red

En este entorno, la Comisión Europea ha promovido la creación de centros que coordinan la información de verificadores e instituciones académicas para detectar las tendencias narrativas de la desinformación. Esto podría ser el indicio de la existencia de campañas organizadas para difundir mensajes engañosos. El Observatorio Europeo de los Medios Digitales (EDMO, por sus siglas en inglés) aglutina a miembros de una red europea de la que también dependen otros 14 observatorios nacionales o multinacionales, como el consorcio IBERIFIER, que reúne a 23 entidades de España y Portugal. Este tipo de alianzas facilitan la detección de campañas de manipulación y engaño gracias a la cooperación y el intercambio de información, además de identificar contenidos promovidos a la vez en distintos países y canales. La relevancia del esfuerzo colectivo también es notoria en asociaciones como la Red Internacional de Verificación (IFCN) y la Red Europea de Estándares de la Verificación (EFCSN), que propician iniciativas como UkraineFacts, una base de datos creada para compartir contenidos falsos y verificaciones sobre la guerra de Ucrania recopilados por organizaciones de todo el mundo. A su vez, los miembros de la Unión Europea de Radiodifusión (UER), una agrupación internacional de la que forman parte servicios públicos de radiodifusión, cooperan en la detección y verificación de contenidos falsos y engañosos a través de la Eurovision Social Newswire. Por esta vía se proporcionan contenidos audiovisuales verificados y se intenta evitar que se publiquen o transmitan contenidos desinformativos que pueden producir daños que cada vez son más difíciles de reparar y afectan a la credibilidad de los medios. La UER promueve además la cooperación en la verificación e investigación digital a través de la Red de Periodismo de Investigación, conformada por periodistas especializados de las principales radiotelevisiónes públicas europeas.

La escucha social

La importancia de ese trabajo en red reside en la puesta en común de los contenidos falsos detectados y de las conclusiones de las investigaciones llevadas a cabo para refutar falsedades y manipulaciones. Pero para identificar esos mensajes engañosos previamente es necesario realizar una labor constante de análisis de las redes sociales con el fin de conocer cuáles son las publicaciones más compartidas y, por tanto, con mayor capacidad de influencia y de causar daño en caso de ser maliciosas. Los contenidos falsos suelen difundirse primero en las redes sociales, donde en muchos casos se viralizan y saltan a aplicaciones de mensajería como WhatsApp. Allí los usuarios comparten los contenidos sin contrastar la información. El anonimato que permiten plataformas como Twitter, Telegram y Facebook facilita la difusión de campañas de desinformación en estas redes. Esas falsedades circulan gracias a la existencia de cuentas automatizadas, cuentas semiautomatizadas con participación humana y grupos humanos coordinados. Estos últimos actúan de forma sincronizada y suelen organizarse en canales y grupos en redes como Facebook, Telegram o WhatsApp. Este tipo de perfiles que se dedican a desinformar suelen camuflar su identidad utilizando nombres y avatares falsos, a veces con imágenes creadas con inteligencia artificial. También es habitual que en redes sociales como Facebook los desinformadores se camuflen con perfiles falsos que suplantan a empresas

y que ofrecen falsos sorteos a cambio de que los internautas se suscriban o sigan a estas cuentas. Una vez que el internauta se ha suscrito, estos perfiles comienzan a difundir, de forma fraudulenta, otros contenidos desinformativos que nada tienen que ver con la cuenta original. Asimismo, se observa un aumento de las cuentas paródicas que imitan los perfiles oficiales de personajes públicos, casi siempre políticos. Su objetivo es desacreditar a estas personas aprovechando la imagen de una cuenta supuestamente dedicada a las bromas.

Como se ha mencionado anteriormente, un indicio de campaña de desinformación es la publicación simultánea de contenido idéntico en varias cuentas de una misma red social o, incluso, en distintas plataformas y hasta en páginas web especializadas en difundir mentiras. Todo esto conforma un ecosistema multicanal de propagación de falsedades. En su monitorización de las redes sociales, los verificadores emplean distintas herramientas informáticas, algunas de ellas proporcionadas por las plataformas con las que tienen acuerdos y de uso confidencial, pero otras abiertas al público general y gratuitas. Los ciudadanos pueden detectar si los mensajes que reciben son casos de desinformación conocidos visitando directamente las páginas web de los verificadores. Otro medio de obtener esa información es haciendo una consulta en internet con las palabras clave de las afirmaciones de las que se desconfía, ya que los principales motores de búsqueda darán como resultado artículos de los verificadores que hayan analizado ese contenido. Si el resultado es infructuoso, los internautas pueden dirigir sus preguntas a los propios verificadores, que tienen canales abiertos para recibir las demandas de la audiencia, ya sea a través de sus páginas web, correos electrónicos, cuentas en redes sociales o canales de WhatsApp. Además, los mismos medios suelen poner a disposición de sus lectores cajas de herramientas y aplicaciones informáticas con las que estos pueden hacer sus propias investigaciones para determinar si una publicación es falsa, en un esfuerzo de alfabetización mediática de la población. Entre los recursos esenciales que están al alcance general de la ciudadanía se encuentra la búsqueda avanzada en los motores de búsqueda y las redes sociales, que permite determinar si una fotografía o un vídeo se publicaron con anterioridad a los hechos con los que se asocian y se trata, por tanto, de una publicación descontextualizada. Del mismo modo, se puede hacer una búsqueda inversa de imágenes para rastrear la presencia previa de esos contenidos en internet. Otras herramientas útiles son InVid, para el análisis de fotografías y vídeos, y Wayback Machine, para hallar versiones archivadas de publicaciones eliminadas de internet. En cuanto a los formatos de la desinformación, esta puede difundirse en textos, audios, fotografías, vídeos e incluso declaraciones públicas. Los canales utilizados para ello pueden ser las redes sociales, las aplicaciones de mensajería instantánea, portales de internet que fingen ser medios de comunicación o los suplantan, o los propios medios.

Aplicar el principio de cautela es una de las recomendaciones que conviene tener presente para evitar ser víctima de la desinformación y acabar participando en su cadena de difusión. Por tanto, se debe evitar compartir aquellos contenidos cuya veracidad no esté confirmada, aunque refuercen las ideas propias y se sienta el impulso de creerlos. También es aconsejable intentar averiguar si el mensaje es falso con una búsqueda de fuentes fiables en internet. Por último, es necesario recordar que, con los avances de la tecnología y el auge de la inteligencia artificial, ni siquiera podemos fiarnos de nuestros sentidos: incluso lo que vemos y oímos puede haberse elaborado con el fin de engañarnos.

HERRAMIENTAS PARA IDENTIFICAR UNA PÁGINA DE DESINFORMACIÓN

Las campañas de propaganda digital suelen articularse empleando distintas herramientas que contribuyen a mejorar la adhesión de públicos afines, disuadir a públicos adversos y generar desconfianza y malestar social. Entre estas herramientas encontramos las redes sociales, pero también el uso de prescriptores e *influencers* o la creación de contenido desde páginas web que simulan ser medios de comunicación.

Se entiende por páginas de desinformación aquellas “páginas web que adoptan la estética y forma de medios de comunicación para difundir propaganda y cuyos contenidos falsos, distorsionados o valorativos contribuyen a beneficiar una causa política o social” (Rodríguez-Fernández, 2020).

Los contenidos de estas páginas se presentan con apariencia similar a las noticias de medios de comunicación, aprovechando esta estética para dotarlas de un halo de credibilidad. En su difusión se emplean las redes sociales, que permiten llegar a un volumen amplio de personas y generar tráfico web hacia las páginas. Estas visitas de usuarios suponen además una oportunidad para conseguir financiación a través de la publicidad digital.

Es importante matizar que no hablamos de una página o dos páginas concretas, sino de ecosistemas que terminan configurando una red para legitimar un relato sobre cuestiones políticas y sociales. Así, cada país tiene su propia red de páginas de desinformación, que suele servir a distintos intereses en función de los marcos políticos y culturales. Se ha observado que en Estados Unidos y el Reino Unido se encuentran las mayores proporciones de desinformación partidista, mientras que en Alemania y Austria prevalecen las historias sensacionalistas. Además, en los países de habla inglesa, la desinformación ataca con frecuencia a los actores políticos, mientras que, en los países de habla alemana, los inmigrantes son los objetivos más comunes (Humprecht, 2018).

Por otra parte, estas páginas suelen adaptarse a las circunstancias, cambiando el nombre de su dominio o readaptando sus contenidos a otras causas cuando dejan de ser útiles. No obstante, se advierten algunas cuestiones que pueden ser útiles para su identificación:

1. Contenidos y autoría: en sus espacios suelen encontrarse contenidos de actualidad copiados de informaciones de agencias y de medios de comunicación reales, en muchas ocasiones reinterpretados, así como textos de producción propia y que tienen una obvia intención propagandística. Estos no suelen presentar una autoría clara, indicando en su defecto el nombre de la propia página o una referencia a una supuesta “redacción”.

En otras ocasiones estas páginas de desinformación ocultan la identidad de una supuesta organización bajo apodos o nombres muy comunes difícilmente rastreables. En algunos casos, llegan a crear perfiles de LinkedIn para simular su existencia. Este último caso no suele ser el más común, pero también puede resolverse atendiendo a la configuración de su equipo y revisando sus nombres con búsquedas en internet. Normalmente este tipo de identidades falsas suelen ser empleadas para firmar los contenidos de opinión, que presentan siempre mayor carga ideológica.

2. Ausencia de tratamiento informativo y tendencia a la exageración: las páginas de desinformación no suelen presentar tratamiento informativo de sus contenidos y tienden a la exageración interpretativa. El lector observará durante la lectura de estos contenidos que no hay investigación por parte del autor; las imágenes suelen ser sesgadas, exageradas o están fuera del contexto informativo, pero han sido incluidas para generar fuertes impresiones emocionales. Igualmente, los titulares presentan un lenguaje poco neutro, bien atribuyendo etiquetas a determinados colectivos o personajes públicos o incidiendo en las referencias ideológicas.
3. Mala calidad de las imágenes y escasos vídeos: dado que la producción audiovisual y el material gráfico requieren de una inversión económica, las páginas de desinformación no suelen ofrecer vídeos propios. Por otra parte, la ausencia de calidad de las imágenes es otro parámetro de observación, ya que este tipo de páginas tiende a emplear fotomontajes o imágenes que no tienen contexto informativo. No obstante, es posible que en un futuro próximo la inteligencia artificial pueda ser aplicada para resolver esta cuestión, refinando aún más el engaño.
4. “Quiénes somos” y “Acerca de”: al tratar de investigar la financiación o la propiedad de la página no se observa información relevante o, cuando se observa, se trata de condiciones legales copiadas de otras páginas. No se indican la empresa responsable, una ubicación física o un teléfono de contacto.
5. Incorrección política: se presentan como páginas que buscan ofrecer una visión alternativa—normalmente con un sesgo ideológico muy marcado—con la intención de ganar credibilidad entre aquellos usuarios que tienen un sesgo similar. Algunas de estas páginas presentan apariencia de medios de comunicación locales, pero en realidad la mayoría de sus contenidos son ideológicos y propagandísticos.
6. Suplantación de la identidad de medios de comunicación reales: otra fórmula para inducir al engaño es tratar de suplantar la identidad de un medio de comunicación real. En estos casos, el propagandista compra dominios web con nombres similares o registra redes sociales sirviéndose de la misma picaresca. EU DisinfoLab destapó en 2022 una red en la que se intentó suplantar la identidad de al menos 17 medios de comunicación europeos (Alaphilippe et al., 2022).

LA IA COMO POTENCIADOR DE LA AMENAZA Y RESPUESTA

La inteligencia artificial y su aplicación a tareas cotidianas está evolucionando a una velocidad nunca antes vista.

Todos estos avances suponen un reto y, a la vez, una oportunidad desde el punto de vista de la lucha contra la desinformación. Por un lado, es cada vez más fácil y accesible el uso de la IA como potenciador de la amenaza mediante la generación, publicación, difusión y monitorización de desinformación en formato multimodal (como texto, audio, vídeo o imágenes). Por otro lado, desde el punto de vista de la respuesta, se plantea el uso de herramientas de IA para la detección, moderación y alerta temprana ante campañas de desinformación, entre otros aspectos.

Utilización de la IA en campañas de desinformación

Desde el punto de vista de un actor interesado en generar y difundir una campaña de desinformación, la IA ofrece diferentes facilidades para la consecución de sus objetivos.

En cuanto a la generación de contenidos para una campaña, estos modelos permiten desarrollar textos bien estructurados para la creación de noticias falsas. Las falsificaciones de imágenes, vídeos o audios (deepfakes) permiten generar un impacto emotivo aún mayor. La combinación de algoritmos de generación de textos y de contenidos multimedia posibilita la adaptación a diferentes canales de comunicación. El principal avance de las herramientas de IA es la facilidad para la generación de contenido multimodal falso a gran escala y la reducción en cuanto a tiempo y costes de generación de la campaña.

Respecto a la distribución de contenidos falsos, la integración con redes sociales y mensajería posibilita la automatización de su publicación. La amplia distribución de *bots* o cuentas automatizadas permite amplificar la difusión mediante la generación automática de *likes* y la de contestaciones (usando de nuevo modelos de IA) que reproduzcan y complementen las opiniones iniciales, resalten respuestas positivas de terceros, desacrediten respuestas contrarias y, en conclusión, promuevan el efecto de mera exposición y el sesgo de confirmación y generen una cámara de eco entre los lectores. Por otro lado, si bien pueden desarrollarse medidas que limiten y dificulten estas acciones, el bajo coste de llevar la campaña a múltiples herramientas hace que sea complejo detenerla completamente.

El conocimiento de los diferentes canales en que se está difundiendo la campaña permite también monitorizar la difusión del mensaje, identificar qué acciones y canales están funcionando mejor y potenciarlos. Esto puede ser automatizado mediante el uso de una

IA que reciba datos de diferentes fuentes (por ejemplo, número de visualizaciones, likes recibidos, contestaciones) y priorice las acciones adicionales que se deben ejecutar en función de los resultados previos.

La facilidad en la generación automática y distribución de contenidos y la inundación actual de vídeos y fotografías hacen que difícilmente se pueda desarrollar una respuesta adecuada en términos de magnitud. Instituciones, medios de comunicación y verificadores pueden convertirse en objetivo, con el fin de desacreditarlos y complicar su labor. Esto, junto a la generalización de información falsa, puede llevar a la pérdida de confianza en la información recibida, también en la de los medios que deberían velar por evitar este problema.

Utilización de la IA para responder a las campañas

La respuesta a las campañas de desinformación puede venir también apoyada por el uso de la IA. La primera opción consiste en utilizarla para la detección de contenidos falsos. Por un lado, es posible entrenar estas herramientas con contenidos previamente identificados como falsos o verdaderos, que pueden ser de tipo textual, imágenes, audios o vídeos, o una combinación de ellos (multimodal). Así, el modelo de IA puede intentar inferir si un contenido nuevo tiene alta probabilidad de ser falso. Otra opción son las búsquedas automatizadas que rastrean similitudes entre las nuevas publicaciones y los desmentidos previos realizados por verificadores. También es posible entrenar las herramientas para detectar desinformación a partir de parámetros distintos del propio contenido, como el análisis de sentimiento (las emociones son un objetivo preferente de la desinformación) o el patrón de difusión (si es posible seguir la misma).

Este tipo de herramientas están ya siendo utilizadas por diferentes perfiles (por ejemplo, para moderación en redes sociales, o por verificadores) para priorizar el contenido que se trata de verificar. Su utilización por un usuario final (por ejemplo, un ciudadano que accede a las redes sociales) es más compleja. Por un lado, las bases por las que un contenido se clasifica como posiblemente falso deben ser claramente explicables, permitiendo a los usuarios entender el porqué de la predicción. Por otro lado, debe tenerse en cuenta que un contenido que no es clasificado como falso no es necesariamente verdadero, sino que el modelo no ha podido realizar dicha predicción basándose en los datos con los que fue entrenado. También es necesario, entre otros muchos aspectos, que el modelo utilizado no reproduzca sesgos.

La IA también puede ayudar en la producción de los desmentidos y en la refutación preventiva por parte de periodistas y verificadores, por ejemplo, ayudando a la localización de información real asociada a la publicación falsa o a la visualización de datos abiertos relacionados con la misma.

Finalmente, un reto esencial es el de cómo adaptar la respuesta, cuando esta ha de tener participación humana, frente a la inundación de desinformación generada por IA. Una primera opción podría ser la de priorizar la respuesta ante aquellas campañas en las que la difusión sea alta y que tengan un alto impacto, entrenando a las herramientas para identificarlas. La introducción de marcas de agua por parte de las aplicaciones de IA en el momento de la generación también simplificaría el proceso de detección, aunque no serviría para los contenidos de actores maliciosos que utilizaran sus propios modelos generativos de IA.

En este sentido, cabe destacar la regulación de IA de la UE (AI Act), legislación en proceso de aprobación, cuyo articulado actual obligaría a los modelos fundacionales generativos a señalar que el contenido que producen lo creó una IA (sin llegar a considerar estos modelos como de alto riesgo), además de otras obligaciones.

En el capítulo 5 se podrá profundizar en las técnicas, metodologías y prospectiva sobre la IA para combatir la desinformación.

EL PAPEL DE LAS PLATAFORMAS DE VERIFICACIÓN. ORGANIZACIONES DE VERIFICACIÓN: DEFINICIÓN, TIPOLOGÍA Y METODOLOGÍA

Organizaciones de verificación

Como explicábamos en la introducción de este capítulo, el crecimiento exponencial de la desinformación en una sociedad digitalizada en la que las redes sociales ocupan un lugar principal en la transmisión de mensajes ha hecho aumentar el número de medios de comunicación dedicados a la verificación. Esto es, la comprobación de la veracidad de las afirmaciones analizadas. Según recuerda la Fundación del Español Urgente de la Agencia EFE y la Real Academia Española (FundéuRAE), “contrastar la información y la veracidad de los hechos y las declaraciones de alguien, especialmente un político, forma parte consustancial de la profesión periodística”. Sin embargo, prosigue, “en la actualidad se dispone de medios tecnológicos que permiten comprobaciones más amplias y veloces”. La verificación, añade la FundéuRAE, es una práctica especializada a la que “se destinan equipos y recursos específicos en los medios de comunicación”. Es, por tanto, una especialidad del periodismo, entroncada con la tradición del oficio, pero con recursos específicos para llevar a cabo una labor más eficiente y ágil en la lucha contra la desinformación.

La verificación como instrumento para detectar falsedades y refutarlas mediante la investigación de los hechos inicia su desarrollo autónomo entre finales del siglo XX y principios del siglo XXI con la fundación de medios de comunicación dedicados en exclusiva a esta práctica en Estados Unidos, con Snopes como decano de los verificadores norteamericanos. Algunos verificadores son organizaciones independientes, como en España Maldita.es, Newtral o Verificat, y en otros casos son equipos o departamentos de medios de comunicación con una actividad más amplia, como ocurre con AFP Factual, EFE Verifica o VerificaRTVE. En este segundo caso, su actividad no solo consiste en examinar la veracidad de publicaciones externas, sino que también incluye una labor de servicio de verificación interna de los medios de comunicación en los que están integrados, previa a la difusión de contenidos.

Otra diferencia en el desempeño de estas organizaciones la establece su adscripción a los códigos de principios de las redes internacionales de verificación. Aquellas que forman parte de la IFCN deben cumplir con las reglas de esta red internacional impulsada por el Instituto Poynter estadounidense. Los medios que forman parte de ella deben pasar evaluaciones anuales que certifiquen el cumplimiento de una serie de compromisos para promover la excelencia, imparcialidad y transparencia en la verificación. Por otra parte, en septiembre

de 2022, 44 medios de 30 países aprobaron el código profesional con requisitos éticos y de transparencia que ha servido de base para la constitución de la red europea EFCSN. Ambos reglamentos ofrecen una garantía del cumplimiento de determinadas normas de calidad por parte de las organizaciones que los suscriben, si bien existen verificadores rigurosos y dignos de credibilidad que no forman parte de estas redes.

En todo caso, la existencia de estos códigos ha establecido un estándar para el conjunto de la comunidad de verificadores, cuyos procedimientos son similares y se basan en principios de no partidismo y equidad, así como la transparencia de las fuentes, la financiación, la organización y la metodología y el uso de correcciones abiertas y honestas. En la selección de los contenidos que analizan, los verificadores priman el riesgo que estos suponen para la sociedad y el volumen de su difusión. En la investigación, las afirmaciones se examinan y contrastan con información pública oficial, la evidencia científica y expertos académicos. En la redacción de las verificaciones, se recopilan las pruebas, los testimonios y las herramientas empleadas para que los lectores sean capaces, si lo desean, de reproducir el camino seguido para llegar a las conclusiones presentadas. Los artículos, además, son revisados por varios miembros de la redacción antes de ser publicados.

Los falsos verificadores

Como parte de las soluciones frente al problema que supone la desinformación, la verificación ha obtenido una confianza creciente por parte de los ciudadanos. Esa credibilidad se ha utilizado también para desinformar, con la publicación de falsas verificaciones, las cuales, en lugar de ofrecer hechos y datos reales, propagan una mentira con apariencia de ser una verdad contrastada. Este tipo de prácticas han sido observadas en la guerra de Ucrania, donde han sido puestas al descubierto por una investigación del Media Forensics Hub de la Universidad de Clemson (Carolina del Sur, Estados Unidos) en colaboración con el medio de comunicación ProPublica. En su estudio se identificaron más de una docena de vídeos creados para refutar supuestos contenidos engañosos ucranianos que no se habían producido y que se grabaron al mismo tiempo que sus falsas verificaciones.

En el ámbito nacional, se ha detectado una tendencia que consiste en replicar el modelo y la imagen de un verificador para tratar de desprestigiar y poner en entredicho este modelo periodístico, así como difundir mensajes engañosos como si fueran el resultado de una verificación. Un ejemplo sería el de una cuenta de Twitter con decenas de miles de seguidores creada en marzo de 2019 con un nombre que pretende emular y contrarrestar el trabajo de Maldita.es y de otros verificadores profesionales.

PARTE 2: RESPONDIENDO A LA DESINFORMACIÓN

¿CÓMO RESPONDER A LAS CAMPAÑAS DE DESINFORMACIÓN Y EN QUÉ CASOS?

Identificación del nivel de riesgo y amenaza

El primer paso para poder afrontar una amenaza y desarrollar mecanismos para reducir sus posibles efectos es generar conocimiento sobre la misma para, utilizando como base dicho conocimiento, poder elaborar un análisis de riesgo para la organización. Además, dada la continua evolución de las técnicas utilizadas por las campañas de desinformación, es necesario mantener una evaluación periódica del riesgo.

Las campañas de desinformación no solo suponen un riesgo para los procesos democráticos y la cohesión social, también pueden presentar riesgos contra la salud y la seguridad públicas. Además, pueden suponer una amenaza contra las empresas, pudiendo menoscabar la imagen corporativa o incluso influir en las decisiones de un consejo asesor o junta de accionistas;¹ así como para las organizaciones no gubernamentales, limitando su financiación o afectando al desempeño de sus proyectos e incluso la seguridad de sus trabajadores.²

A la hora de identificar el riesgo que presentan las campañas de desinformación se pueden definir tres ámbitos de análisis principales:

- Riesgos para el sector o la comunidad: algunas campañas de desinformación pueden ir dirigidas contra organizaciones internacionales, asociaciones sectoriales o las actividades de las misiones humanitarias en una región.

¹ Más información sobre riesgos asociados a las empresas en Petrov P. y Fehon A. (2021): *Disinformation attacks have arrived in the corporate sector. Are you ready?* PriceWaterhouseCoopers. Disponible en: <https://www.pwc.com/us/en/tech-effect/cybersecurity/corporate-sector-disinformation.html>

² Más información sobre riesgos asociados a organizaciones no gubernamentales en Blake J. (31 de julio de 2022): *Disinformation and security risk management for NGOs*. Global Interagency Security Forum. Disponible en: <https://www.gisf.ngo/blogs/disinformation-and-security-risk-management-for-ngos/>

- Riesgos para la organización: en este ámbito se incluirían todas las campañas dirigidas contra la imagen de una organización, sus órganos de gobierno y sus productos o servicios.
- Riesgo para las personas: la desinformación también puede estar dirigida contra ciertas personas (representantes diplomáticos, trabajadores o accionistas) con el fin de limitar su trabajo y credibilidad y, por tanto, los de la organización.

Incrementar la concienciación y formación sobre esta amenaza en el seno de la propia organización

Algunas medidas que podrían adoptarse a nivel interno para concienciar y formar sobre el fenómeno de la desinformación son:

- Sesiones de formación, así como talleres o simulaciones sobre la naturaleza y los peligros de la desinformación. En ellos, se mostrará a los empleados cómo identificarla y evitar su difusión. Estas sesiones pueden ser impartidas tanto por expertos internos (formación entre pares) como por externos (periodistas, académicos o profesionales de la comunicación).
- Boletines informativos y reuniones de equipo sobre esta materia, con inclusión de recomendaciones de buenas prácticas sobre cómo evitar este fenómeno.
- Canales de comunicación interna que permitan a los empleados compartir información de manera segura y fiable y recibir orientaciones sobre cómo verificar la autenticidad de la información: buzón de correo específico, espacio acotado en la intranet corporativa, etc.
- Políticas y guías de fácil lectura que establezcan cómo manejar la información y la comunicación interna en la organización, con especial énfasis en la verificación y el uso de fuentes fiables.
- Ejemplos prácticos de desinformación y de cómo puede afectar al funcionamiento ordinario de la organización con el objetivo de ayudar a los empleados a entender mejor los peligros y motivarlos para adoptar medidas de prevención.
- Evaluaciones periódicas sobre la comprensión y la aplicación de las citadas políticas y guías con el objetivo de identificar posibles brechas y áreas de mejora, mediante la actualización periódica de estadísticas y encuestas de uso interno.
- Liderazgo y cultura de la organización: es imprescindible que los líderes de la organización establezcan una cultura dentro de la organización en la que se valoren la transparencia, la honestidad y la integridad en la comunicación y el intercambio de información como pilares básicos de la organización.

Establecer mecanismos de alerta temprana

Identificar cuanto antes una campaña de desinformación puede limitar su impacto y evitar que se generalice y se asiente. Asimismo, también es esencial en algunos casos, ya referenciados, para establecer mecanismos adecuados que permitan proteger a los trabajadores. A fin de poder establecer los procedimientos de alerta temprana es necesario definir varias cuestiones:

- Qué se quiere identificar: el punto de partida para determinar las amenazas a las que prestar atención es un análisis de riesgo previo. Este debe recoger los principales tipos de riesgos que puedan afectar a la organización.
- Dónde se debe buscar: si bien internet suele ser el principal canal de distribución de las campañas de desinformación, no deben descuidarse otros canales más tradicionales, sobre todo en países donde la penetración de las nuevas tecnologías es limitada.
- Cómo se debe buscar: dependiendo de la complejidad de los riesgos que pueden afectar y la amplitud de los medios en los que buscar (distintas redes sociales, varios idiomas, etc.) se elegirán las herramientas tecnológicas adecuadas: desde simples alertas de motores de búsqueda o herramientas de gestión de imagen corporativa en línea hasta herramientas diseñadas específicamente para la detección de campañas de desinformación.

Establecer procedimientos y protocolos para responder

En tiempos de desinformación cualquier organización puede ser objetivo de campañas orquestadas que busquen afectar a su imagen y reputación (Rodríguez-Fernández, 2019). Una potente herramienta para combatir la desinformación es la comunicación corporativa e institucional. Por ello, las organizaciones tienen que tener en cuenta estas situaciones de riesgo como potenciales escenarios de crisis y anticiparse con tácticas de prevención como las que se indican a continuación:

- **Escucha activa:** la monitorización de contenidos sobre la organización en redes sociales es una de las formas más relevantes para identificar cuestiones sensibles, así como potenciales actores de conflicto. Resulta de vital importancia recopilar periódicamente las conversaciones que se puedan producir y atender a cuestiones sensibles.

Minimizar los mensajes falsos o restarles valor no es una opción, ya que a futuro ello podría convertirse en argumento de desinformación. Si carecen de la suficiente entidad para ser desmentidos, podemos registrarlos, hacer seguimiento de su frecuencia y volumen y localizar los públicos que están prestando atención a dichas cuestiones.

- **Anticiparse a la suplantación de identidad:** aunque la organización no presente actividad comunicativa en todas las redes sociales, resulta conveniente registrar

su nombre para evitar que un tercero pueda acceder fácilmente a dicho registro y suplantar su identidad. Esta medida de cautela debería aplicarse también a los dominios web para evitar que otro actor pueda hablar en nuestro nombre.

Además del nombre de la organización, pueden registrarse otras variantes que ofrezcan credibilidad, como el nombre de la organización y el país (por ejemplo: organización_es). En este sentido, es importante que los principales responsables de la organización (presidentes/as, directores/as generales...) adopten medidas similares, dado que la suplantación de identidades de portavoces y representantes institucionales es cada vez más frecuente. Dicho registro no implica tener actividad comunicativa, sino tener el control de los nombres en estas plataformas.

- **Incorporar la desinformación como potencial escenario de crisis:** las organizaciones suelen recoger en sus manuales de comunicación de crisis actuaciones concretas en función de determinados escenarios de riesgo. Es necesario actualizar estos manuales periódicamente e incorporar la desinformación entre dichos escenarios. Se debe especificar el momento en que debe darse respuesta, así como los miembros de la organización que deben participar en la misma.
- **Trazar escenarios de respuesta para acontecimientos potencialmente problemáticos:** algunas organizaciones, por el carácter de su actividad, deben enfrentar situaciones complejas ante días conmemorativos o como consecuencia de informaciones puntuales que no llegan a derivar en crisis. Es conveniente, cuando sea posible, definir qué días o eventos podrían impulsar una actividad de influencia negativa contra la organización para dar una respuesta rápida.

También pueden acometerse acciones específicas de prebunking, es decir, de comunicación anticipativa, cuya eficacia ha sido demostrada incluso en crisis financieras (Dillingham e Ivanov, 2017).

- **Mantener una comunicación profesional y permanente basada en la verdad que genere confianza en el público:** si la organización tiene una estrategia de comunicación consolidada y presenta altos grados de confianza entre sus diferentes públicos, resultará más sencillo que la verificación funcione y más complejo que esos públicos creen los mensajes desinformativos. En este sentido, no deben olvidarse los públicos internos de la organización, que son grandes prescriptores y validadores de información.

Una vez que la organización ha sido alcanzada por una campaña de desinformación o de influencia negativa, deben establecerse acciones de respuesta. Estas tienen que estar lógicamente asentadas en la verdad, ofreciendo siempre información que permita refutar o esclarecer los contenidos desinformativos, y han de implementarse con la mayor celeridad posible:

- **Desmentir la información en canales propios:** el *Institute for Public Relations* propone que, para defenderse rápidamente de estos ataques, las organizaciones deben centrarse en tres áreas clave. En primer lugar, la atribución. Esto es, la identificando de la fuente emisora. Como se indica, este punto puede ser

complejo, ya que suele tratarse de actores internacionales. En segundo lugar, la motivación. Se trata de comprender el porqué y el objetivo del ataque. En último lugar, la respuesta. En esta área se deben aportar hechos y verdades, realizar conferencias de prensa, crear páginas web para contrarrestar la desinformación o revelar públicamente quiénes están detrás del trabajo desinformativo.

Estas respuestas pueden tener la longevidad que nosotros consideremos pertinente. Algunos bulos tienden a ser cíclicos, por lo que conviene generar contenido que ofrezca indexación a largo plazo. De esta manera, cuando el usuario busque información sobre el mismo, encontrará la respuesta ofrecida por la organización con argumentos y datos que le permitan contrastar. Danone, por ejemplo, recoge en su página web³ todos los argumentos que desmontan el bulo que sostiene que su producto es perjudicial y dañino para la salud, aportando las certificaciones y evidencias correspondientes. Mantener esta información indexada de manera permanente evidencia una clara honestidad y transparencia de la organización con respecto a dicho asunto.

- **El mensaje, el verdadero protagonista:** la corrección de información falsa suele ser más exitosa cuando se acompaña de una explicación distinta al relato difundido a través de la desinformación y cuando se despiertan sospechas sobre su fuente emisora (Lewandowsky *et al.*, 2012). Así, la argumentación debería recoger siempre las posibles motivaciones que han dado lugar a la emisión de ese contenido falso, evitando explicaciones demasiado largas o complejas de comprender.

Ejemplo de esto es el caso de Tesla, que en 2019 fue víctima de un *deepfake*, viralizado a través de Twitter, en el que se mostraba cómo un vehículo autónomo se estrellaba contra un prototipo de robot. El vídeo se hizo viral y la compañía sufrió los efectos del mismo: sus acciones cayeron un 6% en pocas horas. La compañía fue tajante en su respuesta y ofreció una única información, clara y concisa, en la que explicaba que no tenía un vehículo autónomo en producción. Estas declaraciones fueron suficientes para demostrar que el vídeo no era factible, ya que el vehículo no existía, o no, al menos, para tal empresa, provocando la sospecha sobre la fuente emisora (Kropf *et al.*, 2023).

A diferencia de los mensajes empleados en los *storytelling* de las organizaciones, normalmente asentados en la comunicación emocional, la comunicación refutacional tiene que basarse en los datos, el contraste y la generación de confianza racional.

- **Reforzar la comunicación con tu red de apoyo:** resulta complejo identificar el número de personas a las que alcanza un mensaje desinformativo. Por ello, es necesario sumar esfuerzos a la hora de desmentir una información, tratando de incorporar el apoyo de prescriptores y personas de la organización que den confianza a nuestro relato y nos ayuden a difundir nuestra narrativa de confrontación.

³ Conoce toda la verdad sobre el bulo Actimel: <https://actimel.danone.es/novedades/bulo-actimel/>

- **Evaluar y restaurar:** Una vez desmentida la información, corresponde realizar un proceso de evaluación y restauración. En el mismo, se intentará establecer el grado de impacto que ha tenido la campaña de desinformación que ha afectado a nuestra organización, con especial atención a las audiencias que han podido estar expuestas. Esto nos permitirá trazar acciones concretas de restauración de confianza con dichos públicos, ya que es necesario que vuelvan a confiar en la organización si en algún momento han creído los contenidos falsos. Este último proceso requiere de la aplicación de las acciones de prevención trazadas anteriormente: escuchar qué les preocupa y mantener una comunicación permanente, basada en la verdad, que genere confianza.

Generar confianza a través de una comunicación estratégica a largo plazo

En toda sociedad democrática, la lucha contra las campañas de desinformación ha de librarse sin dejar de garantizar la libertad de expresión y la libertad de prensa. En España, además, estas libertades están protegidas constitucionalmente. En este sentido, el incremento de la concienciación, el fomento de la alfabetización mediática y la apuesta por la comunicación proactiva y la transparencia suponen las principales herramientas para limitar el impacto de este fenómeno.

Como mejor se combate la desinformación es informando con datos veraces, contrastados y precisos. A menudo resulta más efectivo prevenir que responder. Se trata de implementar políticas proactivas de lucha contra la desinformación, no de limitarse a acciones reactivas ante este fenómeno. Es importante fomentar la confianza de los ciudadanos en la información que le facilitan las instituciones. Para ello, estas instituciones deben informar sobre su actividad de manera constante, activa, precisa, veraz y transparente para evitar que los vacíos de información sean aprovechados para difundir información inexacta, falsa o malintencionada. Las organizaciones deben convertirse en la principal fuente de información de la ciudadanía sobre ellas mismas y las competencias que ejercen.

Parte de la desinformación puede preverse y contrarrestarse con antelación (por ejemplo, en el caso de procesos electorales o fechas señaladas en las que es posible anticipar este tipo de práctica) mediante un esfuerzo adicional de comunicación que permita minimizar el impacto de este fenómeno. Pero ese esfuerzo ha de iniciarse con antelación. Aquí es donde la comunicación estratégica desempeña un papel fundamental. Una comunicación proactiva permite ocupar parte del espacio informativo que posteriormente intentarán llenar otros con desinformación. El riesgo de no hacerlo, o de hacerlo demasiado tarde, es que sean otros actores los que desplacen el relato que una organización quiere transmitir, contribuyendo a difundir una narrativa distinta a la real que, una vez extendida, es muy difícil de rebatir. Por tanto, es necesario establecer un plan de acción claro y específico, identificar las áreas de mayor riesgo y asignar recursos adecuados y suficientes a esta labor. La

desinformación se combate antes de que aparezca. Una vez aparece esa desinformación sin réplica, la desinformación ya ha ganado, dado que las herramientas y métodos utilizados por los desinformadores son normalmente más complejos y eficaces que los que pueden usar las organizaciones para llevar a cabo su comunicación. Los desinformadores utilizan métodos alegales, cuando no directamente ilegales, que una organización no debe utilizar bajo ningún concepto.

Por tanto, algunas medidas que pueden adoptarse para generar confianza mediante una comunicación estratégica a largo plazo son las campañas de comunicación y concienciación preventivas que permitan atraer la atención de los medios y de los ciudadanos y suscitar su interés. Esas campañas deben narrar casos de éxito y comunicar posibles desafíos. También es recomendable una interlocución privilegiada con los medios de comunicación y los periodistas, relación que puede aportar mayor capacidad de influencia. Además de facilitar información fiable, verificada y transparente, se debe facilitar al máximo el acceso a esta información. Si la información no llega a su público objetivo, no es eficaz. Para ello, lo ideal es utilizar múltiples canales de comunicación, pero adaptando el mensaje a cada canal. De esta forma se llega a un público más amplio, pero, a la vez, se consigue una correcta segmentación de ese público. No funciona igual un mensaje en prensa escrita que en redes sociales, en televisión o en una intranet. Otra forma de generar confianza es fomentar la participación del público en la toma de decisiones y en la generación de ideas, así como escuchar y responder a sus preocupaciones. Las redes sociales desempeñan un papel fundamental en este punto, al permitir una comunicación bidireccional y directa, sin filtros ni intermediarios, con la ciudadanía. También es importante colaborar con líderes de opinión y expertos en cada tema, ya que con ello se puede contribuir a aumentar la credibilidad y la confianza en la información que se comparte. Por último, también puede resultar interesante colaborar con otras entidades que compartan nuestros mismos fines, lo que nos permitirá amplificar el mensaje a la vez que ganamos en credibilidad.

VACUNAS CONTRA LA DESINFORMACIÓN

Desinformación y conducta humana

Actualmente vivimos inmersos en un ecosistema mediático en el que la desinformación y los bulos circulan a través de cualquier canal y de forma casi camaleónica, adaptando sus discursos a los diferentes formatos de comunicación. Además, estos contenidos mediáticos se propagan a una gran velocidad, impidiendo a los receptores reaccionar para razonar y asimilar la información que reciben. Por tanto, es necesario que la ciudadanía adquiera conocimientos básicos sobre cómo funciona la psicología del ser humano en determinadas situaciones. Al respecto, señala el profesor Joan Ferrés (2019) que conocer cómo funciona la conducta humana es fundamental para poder comprender las experiencias mediáticas de las personas. La extensión de los desórdenes informativos a los que estamos asistiendo a diario se ve acelerada por el uso intencional de los sesgos psicológicos (Del Fresno García, 2019). Ser conscientes de estos sesgos es prioritario para reducir nuestra vulnerabilidad ante los desórdenes informativos.

A continuación, se exponen, de forma sucinta, aquellos sesgos asociados al fenómeno desinformativo que afectan a nuestra conducta y, por consiguiente, a nuestra forma de relacionarnos, comportarnos y reaccionar ante diferentes tipos de mensajes.

El efecto de mera exposición

Cualquier estímulo resulta más atractivo cuanto mayor sea el número de exposiciones al mismo. En el caso de la comunicación, se puede afirmar que, bajo este efecto, el receptor puede sentirse atraído por un mensaje simplemente por la familiaridad que ha surgido después de haberlo visto muchas veces. Ahora bien, existen dos limitaciones a la efectividad persuasiva de la repetición. En primer lugar, si se repite el mensaje demasiadas veces, podría desgastarse. Para prevenir este efecto, se suele reiterar la misma información, pero con una presentación diferente. Esto es algo muy habitual en el campo de la publicidad. La segunda limitación tiene que ver con el procesamiento de esta información. Cuanto más superficial es el procesamiento, menor poder persuasivo tiene. Sin embargo, en la mayor o menor profundidad de ese procesamiento, las emociones desempeñan un papel fundamental, de forma que la emoción que se suscita en nosotros puede determinar el nivel de profundización que alcancemos con el mensaje.

La deseabilidad social

Este sesgo surge cuando anticipamos que aquello que vamos a decir u opinar está en consonancia con lo que piensa el resto de la gente que nos rodea. Se trata de uno de los principales problemas que emergen en las encuestas de intención de voto, donde se comprueba que las personas a veces dicen lo que “creen” que suena mejor. La necesidad de pertenencia a un grupo y de conformidad con la mayoría juega un papel importante. En el caso de la desinformación, este sesgo puede aparecer a la hora de compartir con otros ciertas informaciones o contenidos “deseables”, aunque no hayan sido contrastados. Si lo extrapolamos a los adolescentes, el fenómeno es mayor, puesto que esta es la etapa en la que las personas más necesitan reforzar su identidad grupal.

El sesgo de confirmación

Se trata de un sesgo que lleva al sujeto a elegir aquellas evidencias o argumentos que refuerzan sus opiniones o ideas previas (Ash, 1951). Este sesgo es un mecanismo de autodefensa para evitar o reducir la disonancia cognitiva que puede generar el enfrentarnos a una idea o un argumento diferente a lo que creemos o pensamos. Se trata de un fenómeno que en periodismo se observa con respecto a la fidelización a unos medios de comunicación determinados. Escuchamos, vemos o leemos con más frecuencia aquel medio cuya línea editorial refuerza nuestra ideología. Es más cómodo a nivel cognitivo y consolida nuestras ideas previas. En las redes sociales, el sesgo de confirmación deviene en el fenómeno de la cámara de eco (discursos similares), que puede fomentar la polarización respecto a determinados pensamientos.

Filtro burbuja

Las actividades que realizamos en internet van dejando una huella digital que leen los algoritmos de las páginas web para personalizar los resultados de las búsquedas. También, estos algoritmos devuelven al usuario información relacionada con lo que previamente ha buscado en internet. Este fenómeno genera un sesgo que se denomina filtro burbuja (Pariser, 2017) porque nos aísla de otros puntos de vista diferentes a los nuestros.

Teoría de la inoculación y vacunas contra la desinformación

“La dificultad para diferenciar desórdenes informativos intencionales de la información veraz, soportada en un conocimiento más exacto de los sesgos cognitivos, no va a dejar de crecer” (Del Fresno García, 2019). Ante esta realidad, no solo podemos luchar contra la desinformación desde la penalización, sino que es necesario alfabetizar a la ciudadanía para que sea consciente de sus debilidades psicológicas a la hora de consumir información.

La teoría de la inoculación está relacionada con lo que algunos referentes académicos en el mundo de la educomunicación han apuntado sobre la integración. Esto es, que la mejor manera de enfrentarse a los efectos negativos de fenómenos como la desinformación sería incorporarla al mismo proceso de prevención o curación para desarrollar mecanismo de defensa contra ella.

Conceptualización de la teoría de la inoculación

La teoría de la inoculación en psicología social (McGuire, 1970) tiene como objetivo generar inmunidad en las personas respecto a la posible manipulación, al exponerlas previamente a una versión debilitada del argumento. Las vacunas contra la desinformación son un medio de establecer resistencias cognitivas a este tipo de influencia. Esto permite a los receptores identificar esa desinformación, reducir el efecto en sus creencias y convicciones, así como limitar la cadena de distribución de la misma y el efecto de la mera exposición.

La inoculación se puede dirigir a combatir contenidos concretos, por ejemplo, avisando sobre ciertas campañas de desinformación o, sabiendo que hay una campaña en marcha, transmitiendo información veraz que combata la narrativa subyacente (refutación preventiva). También es posible un enfoque más general, entrenando a la ciudadanía en las técnicas y estrategias que se usan en desinformación (polarización de grupo, suplantación, falsas dicotomías...).

Principales iniciativas y actividades

En el contexto de las vacunas contra los desórdenes informativos, se han implementado diferentes iniciativas y actividades que se han dividido en inoculación pasiva y activa. La inoculación pasiva no requiere de una acción del receptor, salvo la lectura o visualización de la información. Un ejemplo de inoculación pasiva es la acción de Twitter durante las elecciones de 2020 en Estados Unidos. En ella, Twitter alertaba de posibles intentos de desinformación y distribuía información verídica sobre algunos ataques previstos, por ejemplo, en el voto por correo. Este tipo de acciones —en formato escrito y orientadas a difundir información verídica— están siendo implementadas por los verificadores.

En cuanto a la inoculación activa, esta conlleva la implicación del destinatario en una acción que facilitará la interiorización de los conceptos. Diversos estudios parecen sugerir que este último enfoque tiene un resultado mejor en combatir la desinformación (preferencia de la experiencia frente al análisis). En la inoculación activa es importante considerar las actividades basadas en juegos. *Bad News Game* fue el primero de varios juegos en línea creados por la Universidad de Cambridge para vacunar contra la desinformación. En él se simula una red social donde el jugador es un experto en desinformación que tiene que

difundir noticias falsas y obtener seguidores. Posteriormente, Cambridge colaboró con el Gobierno británico en el desarrollo de *GoViral!*, sobre la COVID-19, con el respaldo de la Organización Mundial de la Salud (OMS). También desarrolló, junto con el Gobierno de Estados Unidos, *Harmony Square* para evitar desinformación política en las elecciones de 2020. Después vendrían *Join this Group*, en colaboración con Whatsapp, y *Radicalize*, junto a la ONG Nudge Lebanon, sobre el reclutamiento por parte de extremistas. En cada uno de ellos se trabajaban las técnicas concretas de desinformación y los pasos que se usan en el contexto correspondiente. En España tenemos el caso de *La fábrica de bulos*, una escape room desarrollada por Maldita.es en colaboración con Oxfam Intermón orientada a combatir la desinformación relacionada con la migración.

Para los rangos de edad y público objetivo en los que los juegos pueden no ser un elemento de interés, los cuestionarios son otro medio de inoculación activa posible. Investigadores del instituto Max Planck los utilizaron con diferentes usuarios que, tras rellenarlos, recibían de forma automática comentarios sobre un aspecto de su personalidad (extroversión) susceptible de manipulación a través de anuncios dirigidos. Tras recibir esta información, dichos usuarios obtenían mejores resultados que un grupo de control a la hora de detectar anuncios orientados a explotar dicha característica de su personalidad.

Limitaciones y conclusiones

En general, las intervenciones de inoculación realizadas frente a la desinformación parecen haber tenido un efecto suficientemente apreciable (Traberg, 2022). Sin embargo, el uso del método de inoculación para combatir la desinformación tiene sus limitaciones. Algunas de ellas se enumeran a continuación:

- El efecto de la inoculación es temporal y este decae transcurridas entre dos y seis semanas, aunque es algo mayor para la inoculación activa basada en juegos. Esto requiere la aplicación periódica de estos mecanismos.
- Se debe considerar el contexto local, incluidos el nivel de alfabetización digital y mediática, el contexto cultural, los valores locales o las preferencias sociales.
- En el caso de la refutación preventiva, se requiere un esfuerzo importante —el cual no es siempre escalable— de detección, redacción de argumentarios y difusión.

Por ello, se recomienda que su uso se combine con otros elementos, como los desmentidos realizados por los verificadores. Para obtener mejores resultados se plantea la búsqueda de la inmunidad de rebaño. La suposición es que, si un alto porcentaje de la población es sometida a estas acciones preventivas, se puede también evitar que la desinformación llegue a las personas que no fueron vacunadas.

EL PAPEL DE LA ALFABETIZACIÓN MEDIÁTICA EN EL ACTUAL SISTEMA INFORMATIVO: PRINCIPALES INICIATIVAS Y EXPERIENCIAS

Instituciones privadas

Taller Telekids

Taller Telekids es una entidad dedicada a la educación mediática y, también, productora audiovisual. Telekids tiene como objetivo principal la formación de espectadores críticos. Para alcanzar este objetivo, organiza talleres de cine y televisión para niños, adultos y profesorado. Aunque se trata de una propuesta para desarrollar competencias relacionadas con la alfabetización audiovisual, en los últimos tiempos el equipo de Telekids ha tomado partido contra la desinformación. Telekids ha publicado tres manuales con recursos didácticos sobre desinformación para trabajar en el aula, titulados *Educación en el aula sobre fake news*, *Con las fake news no se juega* y *Eva, la pequeña reportera*.⁴

Junior Report (Blue Globe Media)

Junior Report es un proyecto de alfabetización mediática creado en 2017 por la productora Blue Globe Media. El producto estrella de la productora es el periódico digital Junior Report. Esta cabecera se ha creado con el objetivo de que los jóvenes de entre 12 y 18 años adquieran el hábito de informarse a través de diarios convencionales. Los contenidos de la revista están elaborados por el alumnado de colegios e institutos. Tanto los estudiantes como el profesorado que participan en la elaboración de la revista cuentan con el asesoramiento y la revisión de los contenidos por parte de los periodistas de Junior Report. Además, los periodistas educadores de Junior Report imparten talleres de educación mediática y digital en las escuelas asociadas al proyecto. En estos cursos se enseña al alumnado a documentarse, a elaborar contenidos en distintos formatos y a identificar las distintas caras del desorden informativo. Los recursos didácticos para la enseñanza y el aprendizaje son elaborados por periodistas y pedagogos.⁵

⁴ Este material didáctico y sus experiencias de alfabetización mediática se alojan en su página web: <https://tallertelekids.com/>

⁵ Más información en: <https://junior-report.media/>

Vito. Educando en comunicación

Vito. Educando en comunicación es una iniciativa creada en 2019 por la periodista freelance Victoria Contreras. Los proyectos de alfabetización mediática que impulsa Vito pretenden acercar los medios de comunicación a la ciudadanía. Para ello, la entidad organiza talleres de radio, prensa y televisión en centros educativos de enseñanza primaria y secundaria de Andalucía. Estos talleres van dirigidos a toda la comunidad educativa: alumnado, profesorado y familias. La metodología pedagógica que utiliza Vito en los talleres es “aprender haciendo”, otorgando con esta propuesta todo el protagonismo a la práctica. Vito, por otra parte, coordina los proyectos de alfabetización mediática de la Asociación de la Prensa de Málaga: *La prensa en mi mochila* y *La prensa sin edad*.⁶

Mentes AMI (Fundación Atresmedia)

Proyecto creado por la Fundación Atresmedia con la colaboración de Platino Educa y UNIE Universidad. Se trata de una convocatoria anual de premios para promover, reconocer y difundir iniciativas educativas desarrolladas en las aulas que fomenten el pensamiento crítico, impulsen la creatividad audiovisual responsable y promuevan los valores y la convivencia. Para la entrega de estos premios se organiza un macroencuentro educativo, gratuito y dirigido a los docentes con la intención de contribuir al desarrollo de habilidades y competencias relacionadas con la Alfabetización Mediática e Informativa (AMI), así como para proporcionarles herramientas para su aplicación en el aula. También se han producido diferentes vídeos cortos que forman parte de AMIBOX y AMIKIDS. En ellos —a través de *influencers* y otros personajes populares entre los más jóvenes— se informa de forma sencilla y con un lenguaje ameno sobre cómo navegar por la red de una manera responsable y evitar abusos o estafas. Estos vídeos están alojados en la plataforma de ATRESplayer y son de acceso gratuito. Uno de los objetivos es que puedan ser trabajados también en el aula por los propios docentes.⁷

Maldita Educa

Esta iniciativa ha puesto en marcha la fundación de la plataforma de verificación Maldita.es. Desde 2018, esta plataforma realiza actividades de formación dirigidas a diferentes colectivos, como adolescentes o personas mayores, entre otros. El objetivo es fomentar la alfabetización mediática y dotar de herramientas y recursos a la ciudadanía para que pueda hacer frente a los bulos y a la desinformación. Sus acciones de formación llegan a medios de comunicación, universidades, colegios e institutos, ayuntamientos, ONG o directivos de empresas. En los últimos años, Maldita Educa también participa, junto a diferentes universidades, en proyectos europeos de investigación.⁸

⁶ Las experiencias educomunicativas que propone Vito están alojadas en su página web: <https://vitocontreras.es/>

⁷ Más información: <https://fundacion.atresmedia.com/Mentes-AMI/2023/>

⁸ Más información: <https://maldita.es/malditaeduca/>

Instituciones públicas

(In)fórmate (Campus FAD)

(In)fórmate es una iniciativa de Google, Fad (Fundación FAD Juventud) y el Gobierno de España que cuenta con el apoyo de algunos medios de comunicación. Se trata de un proyecto formativo sobre el consumo de medios e información en internet que promueve la alfabetización mediática y el fomento del pensamiento crítico entre la población adolescente de 14 a 16 años que está cursando 3º y 4º de la ESO en centros educativos españoles. El proyecto cuenta con un programa educativo llamado *Surfear la red*. Este programa está dirigido a que el alumnado de educación primaria navegue con seguridad (*Sé Genial en Internet*) y a que el de secundaria detecte y evite bulos en la red (*Eraser*). Desde el propio portal, los centros interesados en implantar la alfabetización mediática en el aula pueden solicitar talleres y formación específica para el propio profesorado.⁹

Educlip

Iniciativa impulsada por la Red ALFAMED, la Agencia Estatal de Investigación del Ministerio de Ciencia, Innovación y Universidades de España y RTVE. La finalidad es ofrecer un portal en el que los más jóvenes puedan ver, subir y compartir vídeos sobre educación y comunicación y, así, conocer cuál es la mirada crítica que tienen hacia los medios de comunicación. Uno de los principales incentivos es que los ganadores del Concurso Educlips pueden ver su vídeo emitido en Televisión Española.¹⁰

Educación mediática para una ciudadanía crítica (Cátedra RTVE-UJI)

El Consejo Audiovisual de la Comunidad Valenciana (CACV), junto a la Cátedra RTVE “Cultura Audiovisual y Alfabetización Mediática” de la Universidad Jaime I, ha producido la serie *Educación mediática para una ciudadanía crítica*, cuyo objetivo es contribuir al desarrollo de un sentido crítico de la ciudadanía ante el sistema de medios de comunicación. Esta serie se compone de siete clips de vídeo que abordan diferentes temáticas en torno al concepto de alfabetización mediática, así como acerca de la naturaleza y estructura del sistema audiovisual español y valenciano.

⁹ Más información en: <https://www.campusfad.org/informate/>

¹⁰ Más información en: <https://www.educlips.es/>

Los materiales producidos en castellano y en valenciano, y con subtítulos en castellano, valenciano e inglés, son de acceso gratuito. Cada videocápsula está acompañada de unidades temáticas para trabajar los contenidos. Se trata de una iniciativa que están desarrollando medios de comunicación públicos de nuestro entorno como France Télévisions (Francia), BBC (Reino Unido), RTBF y VRT (Bélgica), RAI (Italia) o RTP (Portugal), entre otros.¹¹

Programa ComunicA (Consejería de Educación y Deporte de la Junta de Andalucía)

ComunicA es un programa para la innovación educativa, impulsado por la Consejería de Educación y Deporte de la Junta de Andalucía, que tiene como objetivo ofrecer estrategias y recursos metodológicos que puedan complementar el plan de estudios. La línea de intervención relativa a la alfabetización mediática contempla 15 recursos didácticos para educación infantil, primaria, secundaria y bachillerato. De entre las iniciativas ofertadas, destacamos el taller de alfabetización mediática *Desenreda* para luchar contra el desorden informativo. Se trata de un curso teórico-práctico impartido por periodistas, dividido en tres módulos: desinformación, redes sociales y desempeño de la tarea del periodista. Este recurso didáctico se aplica en los cursos de 3º y 4º de la ESO. Entre las competencias que el programa pretende transmitir está el uso y gestión de la comunicación, así como la formación de un criterio propio y fundamentado a partir de los mensajes que llegan a través de los medios.¹²

¹¹ Más información en: <https://catedrartve.uji.es/serie-educacion-mediatica-para-una-ciudadania-critica/>

¹² Más información en www.juntadeandalucia.es/educacion/portals/web/lecturas-bibliotecas-escolares/programa-comunica

Instituciones del tercer sector

Asociación de la Prensa de Málaga (tercer sector)

La labor que realizan las asociaciones de la prensa en materia de alfabetización mediática cobra cada vez más peso. La Asociación de la Prensa de Málaga (APM) ha impulsado varios proyectos destacables. *La prensa en mi mochila*¹³ pretende acercar el mundo del periodismo a los escolares de primaria y secundaria para incentivar el consumo responsable y crítico de los contenidos mediáticos. *La prensa sin edad*¹⁴ pretende, a través de la alfabetización mediática, minimizar la brecha digital y cognitiva que sufren, en especial, las personas mayores de las zonas rurales. En este proyecto también se contempla alfabetizar a los mayores para mejorar la resiliencia frente a la desinformación. *¡Reinventate! Educación mediática contra el desorden informativo*¹⁵ es un taller que tiene como objetivo formar a periodistas en alfabetización mediática. Este taller también está pensado para impulsar la figura del educador mediático como asesor y docente educomunicativo en centros de enseñanza, asociaciones ciudadanas y medios de comunicación.

¹³ Más información en <https://apensamalaga.com/prensa-en-mi-mochila-portada>

¹⁴ Más información en <https://apensamalaga.com/prensa-sin-edad-portada>

¹⁵ Más información en https://apensamalaga.com/prensa_en_mi_mochila/reinventate-educacion-mediatica-contra-el-desorden-informativo-inscribete-ya-en-el-curso#

BIBLIOGRAFÍA

- Alaphilippe, A., Machado, G., Miguel, R. y Poldi, F. (2022). Doppelganger – Media clones serving Russia Propaganda. <https://www.disinfo.eu/doppelganger>
- Humprecht, E. (2018). Where ‘fake news’ flourishes: a comparison across four Western democracies. *Information Communication and Societ*, 21, 1-16.
- Rodríguez-Fernández, L. (2021). Propaganda digital. Comunicación en tiempos de desinformación. UOC.
- Rodríguez-Fernández, L. (2020). Junk News y “medios de desinformación” en la campaña electoral del 10-N. *Revista Mediterránea de Comunicación*, 11(2), 71-83. <https://www.doi.org/10.14198/MEDCOM2020.11.2.19>
- Ash, S. E. (1951). Effects of group pressure upon the modification and distortion of judgement. En *Groups, Leadership and Men*. Carnegie Press.
- Del-Fresno-García, M. (2019). Desórdenes informativos: Sobreexpuestos e infrainformados en la era de la posverdad. *El Profesional de la Información*, 28(3). <https://doi.org/10.3145/epi.2019.may.02>
- Ferrés, J. (2019). Las pantallas y el cerebro emocional. Gedisa.
- Harjani, T., Roozenbeek, J., Biddlestone, M., van der Linden, S., Stuart, A., Iwahara, M., Piri, B., Xu, R., Goldberg, B., y Graham, M. (2022). A Practical Guide to Prebunking Misinformation.
- McGuire, W. J. (1970). Vaccine for brainwash. *Psychology today*, 3(9), 36.
- Pariser, E., y Vaquero, M. (2017). El filtro burbuja: Cómo la red decide lo que leemos y lo que pensamos. *Taurus*.
- Traberg, C. S., Roozenbeek, J., y van der Linden, S. (2022). Psychological inoculation against misinformation: Current evidence and future directions. *The ANNALS of the American Academy of Political and Social Science*, 700(1), 136-151.
- Van der Linden, S. (2023). Foolproof: Why We Fall for Misinformation and How to Build Immunity. Fourth Estate.

Danone (s.f). Conoce toda la verdad sobre el bulo Actimel. <https://actimel.danone.es/novedades/bulo-actimel/>

Dillingham, L. L., y Ivanov, B. (2017). Inoculation messages as a preemptive financial crisis communication strategy with inexperienced investors. *Journal of Applied Communication Research*, 45(3), 274–293. <https://www.tandfonline.com/doi/full/10.1080/00909882.2017.1320571>

Institute for Public Relations (2022). Defending an organization against disinformation. <https://instituteforpr.org/dark-pr-is-infecting-global-business-executives-need-to-respond/>

Kropf B., Wood M. y Parsons K. (2023). Message matters: Correcting organisational fake news. *Computers in Human Behavior*. doi: <https://doi.org/10.1016/j.chb.2023.107732>

Lewandowsky, S. y Van der Linden, S. (2021). Countering Misinformation and Fake News Through Inoculation and Prebunking. *European Review of Social Psychology*, 32(2), 348-384. <https://www.tandfonline.com/doi/full/10.1080/10463283.2021.1876983>

Rodríguez-Fernández, L. (2019). Desinformación y comunicación organizacional: estudio sobre el impacto de las fake news. *Revista Latina de Comunicación Social*, 74, 1714-1728.

Capítulo 2

ESTUDIO DE LA DESINFORMACIÓN RUSA
A NIVEL INTERNACIONAL

Coordinadores:

Nicolás de Pedro

Ministerio de Defensa - Estado Mayor Conjunto de la Defensa - DIVESTRA

Autores y colaboradores:

Juan Luis Chulilla Cano

Borja Díaz-Merry Rivas

Carmen Fernández Torres

Alberto Fierro Aldudo

Carlos Galán Cordero

Juan José Herrera de la Muela

Carla Hobbs

Daniel Iriarte Olalla

Javier Lesaca Esquiroz

Raquel Miguel Serrano

Aurea Molto Sánchez

Blas Moreno

Jonathan Nelson

Jesús Pérez Triana

Alejandro Romero Gallego

Rafael Rubio Núñez

Javier Rubio Donzé

Irene Sánchez Cózar

José Ignacio Torreblanca Payá

Ministerio de Defensa - Centro Superior de Estudios de la Defensa Nacional (CESEDEN)

Ministerio de Defensa - Mando Conjunto del Ciberespacio (MCCE)

Ministerio del Interior - Dirección General de la Policía – (CGI)



INTRODUCCIÓN

Desde hace más de una década, Rusia recurre de forma masiva y sistemática a las campañas de desinformación e influencia como un pilar esencial de su política interior y exterior. El Kremlin concibe un entorno estratégico caracterizado por una permanente “guerra o confrontación de la información” (*informatsionnaya voina o informatsionnoye protivoborstvo*). Es decir, una actividad que se ejecuta tanto en tiempos de paz como de guerra, cuyas fronteras se difuminan, y que, en la concepción rusa, se aplica en un espectro continuo que abarca desde la infraestructura para la transmisión de señales al plano cognitivo. Esto se traduce en la práctica en que las operaciones rusas en el ámbito de la información contra los que percibe como adversarios estratégicos —singularmente los países miembros de la Unión Europea y la OTAN— tienen lugar de forma permanente, incluso en tiempos de paz. Estos ataques pueden ir desde ataques cibernéticos con robo de datos hasta agresivas campañas de desinformación con suplantación de medios de comunicación de referencia o instituciones públicas. Todo ello con vistas a neutralizar, desde dentro, a sus adversarios e impedir, a su vez, ataques similares que el Kremlin y su comunidad estratégica están genuinamente convencidos de sufrir procedentes de Occidente.

RUSIA Y LA “GUERRA DE LA INFORMACIÓN”

El uso y control de la información en un sentido amplio e integral es un pilar fundamental de la Rusia del presidente Vladímir Putin desde su misma llegada al poder en el año 2000. Su régimen no puede entenderse sin incluir la dimensión mediática e informativa. Así, entre sus primeras medidas de calado cabe mencionar la toma de control de los influyentes canales de televisión NTV y ORT. El asunto se presentó públicamente como una mera disputa empresarial privada —justificada como castigo por la corrupción de sus propietarios—, pero resultaba evidente el deseo del nuevo presidente de controlar medios que habían demostrado una significativa influencia y alcance durante el periodo de su predecesor Borís Yeltsin.

A finales del primer mandato de Putin en 2004, el Kremlin ya controlaba, directa o indirectamente, el grueso de la oferta televisiva. A través de los tres grandes canales, Piervy Kanal (antigua ORT), NTV y Rossiya 1, el Kremlin estableció un control firme del flujo informativo y narrativas predominantes en el discurso que llega hasta día de hoy. Esos mismos tres canales concentran en la actualidad el 80% de la audiencia mayor de dieciocho años y son la fuente principal de información para tres cuartas partes de la sociedad rusa. Y el espacio para la crítica se ha ido reduciendo progresivamente desde 2012 y, de manera drástica, desde el inicio de la invasión de Ucrania a gran escala en febrero de 2022. Eso incluye los medios tradicionales y también los digitales. De esta manera, bajo una aparente pluralidad mediática y libertades formales sobre el papel, en Rusia existe una auténtica dictadura informativa.

La conceptualización de la información que se produce en el ámbito estratégico y militar refuerza la acción del Kremlin. La Doctrina de Seguridad de la Información de la Federación Rusa, adoptada tempranamente en septiembre de 2000, ya ofrece una definición de la “esfera de la información” amplia e integral que anticipa certeramente el impacto de la digitalización. De este modo, la define como “el conjunto de información, infraestructura de la información y entidades dedicadas a la recopilación, formación, difusión y uso de la información, así como un sistema que rige las relaciones públicas surgidas de esas condiciones”. Y continúa:

“La esfera de la información, como sistema que conforma la vida social, influye decisivamente en el estado de los componentes —entre otros, políticos, económicos y de defensa— de la seguridad de la Federación Rusa. La seguridad nacional de la Federación Rusa depende sustancialmente del nivel de seguridad de la información, y se espera que esta dependencia aumente con el progreso de la tecnología”. (Presidente de la Federación de Rusia, 2000)

Aquí radica una de las grandes diferencias entre las concepciones occidental y rusa de la información en el ámbito estratégico. Así, a grandes rasgos, en el ámbito occidental la tendencia es adoptar un enfoque restringido a tiempos bélicos, si se trata de operaciones ofensivas, y muy centrado en la protección de la infraestructura física o cibernética. El objetivo es permitir el libre flujo de información y posibilitar los sistemas de mando y control. Sin embargo, en ningún caso se entra a considerar el contenido de la información. Por el contrario, Rusia adopta un enfoque integral que incluye tanto las infraestructuras de transmisión como el contenido de esa información. Esto se basa en la convicción de que la revolución digital permite ejecutar operaciones a gran escala destinadas a moldear, alterar o quebrar el pensamiento y las percepciones de una comunidad política completa.

Asimismo, la desinformación rusa actual es heredera directa de la doctrina soviética y de sus medidas activas, aunque adaptadas a un contexto caracterizado por el auge de las tecnologías de la información y la comunicación digital. Internet es el elemento que transforma radicalmente el contexto informativo con respecto al de la época de la Guerra Fría. Aquí, de nuevo, el contraste con el enfoque predominante en Occidente no podía ser mayor. Así, del lado ruso, la desinformación se concibe como una herramienta estratégica con la que, trazando un símil, se busca una denegación de área, acceso y maniobra en el ámbito informativo y cognitivo en el marco de una permanente “guerra de la información”. Mientras que, del lado de las democracias liberales, el debate público sobre la desinformación se circunscribe con frecuencia a una cuestión de praxis periodística y educación social.

Ahondando en estas diferencias el Ministerio de Defensa de la Federación Rusa publica en septiembre de 2011 un documento con la pretensión de que sea la base para una Convención Internacional sobre la Seguridad de la Información en el que se ofrece ya una definición explícita del concepto de guerra de la información (*information warfare*): “El conflicto entre dos o más Estados en el espacio informativo con el objetivo de infligir daños en los sistemas de información, procesos y recursos, así como en las estructuras de importancia crítica, con el objetivo de socavar los sistemas político, económico y social de otro Estado. Esto se llevará a la práctica con el uso de campañas psicológicas masivas contra la población de otro Estado con el fin de desestabilizar su sociedad y su gobierno y forzar a ese Estado a tomar decisiones en interés de su adversario”. Y tal y como se afirma en ese mismo documento, ese es el contexto en el que estamos y nos vemos abocados si no se adopta esta Convención. (Ministerio de Defensa Ruso, 2011)

Este documento complementa, entre otros, la Doctrina Militar de febrero de 2010 en la que se alerta de que “el rol de la guerra de la información se intensifica”. También se advierte de la “implementación previa de medidas de guerra informativa para alcanzar objetivos políticos sin la utilización de la fuerza militar para, posteriormente, moldear una respuesta favorable de la comunidad internacional ante la utilización de la fuerza militar” (Presidente de la Federación de Rusia, 2010). Esta es la idea o conceptualización popularizada después en el celeberrimo discurso y artículo de febrero de 2013 de Valeri Gerásimov, jefe del Estado Mayor de las Fuerzas Armadas rusas. Conviene recordar, una vez más, que Gerásimov

no inventa o propone un concepto propio sobre la “guerra híbrida”, sino que describe y reflexiona sobre lo que interpreta que Occidente está aplicando contra Rusia. (Gerasimov, 2013)

Es una idea ampliamente asumida en el Kremlin y en la comunidad estratégica rusa que la guerra de la información es el instrumento decisivo para instigar una “revolución de color”, auténtica bestia negra y obsesión rusa. Desde el prisma conspirativo de Moscú el ciclo de protestas en el espacio eurasiático en la primera década de este siglo —conocidas popularmente como “revoluciones de colores”— no son más que golpes de Estado inducidos por Occidente, disfrazados de altruismo democrático, aunque, en realidad, persigue fines geopolíticos y tiene a Rusia como gran objetivo. El propio presidente Putin presentó públicamente las protestas que tuvieron lugar en Moscú entre diciembre de 2011 y marzo de 2012 como un intento de revolución de color instigado por EEUU con el único “objetivo de destruir el Estado ruso y usurpar el poder (Putin, 2011, 2012). Visto así, no es de extrañar que, en palabras del prominente analista ruso: “desde 2014 el Kremlin ha estado de facto operando en modo de guerra y el presidente ruso Vladímir Putin actúa como un líder en tiempos de guerra” (Trenin, 2017).

Y ahí juegan un papel central los medios de comunicación controlados por el Kremlin, concebidos no como medios sino como instrumentos o verdaderas armas para desestabilizar países en tiempos de confrontación abierta. Así, el canal de noticias global RT, (antigua Russia Today), creado en 2005 para, inicialmente, dar una visión más positiva del país fue rebautizada como RT en 2009 debido a lo que se interpretó como un pobre desempeño informativo ruso durante la guerra contra Georgia de agosto de 2008. El canal se mantendría en línea con la agenda del Kremlin, pero ya no se trataba solo de dar noticias sobre Rusia o el punto de vista ruso sobre noticias internacionales, sino, sobre todo, difundir todo aquello que cuestionara y contribuyera a erosionar la legitimidad de los países occidentales. En una entrevista en 2012 la editora en jefe y peso pesado de la maquinaria de propaganda y desinformación rusas, Margarita Simonyan, calificó a RT como un “arma como cualquier otra” y un instrumento concebido para la guerra, porque “es imposible empezar la construcción de un arma cuando la guerra ya ha estallado” (Simonyan, 2012). Es decir, hay que capturar y fidelizar las audiencias antes de iniciar las hostilidades y por eso, RT debe parecer un medio y camuflar su desinformación bajo capas de genuina información.

Así, el objetivo no es ofrecer otros puntos de vista para que —como reza su eslogan— el espectador se pregunte o sepa más, sino para desestabilizar, deslegitimar, confundir y desmoralizar a los que concibe como adversarios. Y para ello explota sus fisuras y debilidades. De ahí que cuanto más entremezcla su actividad con una realidad local genuina y orgánica mayores son sus probabilidades de éxito. Una de las características de la desinformación rusa es que no trata tanto de vender las bondades de la sociedad rusa y su modelo político como de generar cinismo y apatía en las audiencias objetivo en Occidente. Así, la desinformación rusa se convierte en un arma estratégica cuando se emplea para inhibir la intervención de Occidente allí donde los intereses del Kremlin estén en juego. Las narrativas rusas presentan las acciones de su gobierno motivadas por razones nobles y

universales, alineadas con los intereses de los pueblos locales, mientras que las acciones occidentales se presentan como malvadas, egoístas o al menos hipócritas. El objetivo último es generar dudas sobre las verdaderas intenciones de los gobiernos occidentales y, sobre todo, hacer que la audiencia objetivo sienta menos empatía o conexión con las víctimas de las acciones rusas y de sus aliados.

Y conviene no perder de vista que RT o la agencia de noticias Sputnik son solo dos instrumentos dentro de un ecosistema mucho más amplio que se analiza en el siguiente apartado. En él, diversos actores confluyen y se complementan, pero están todos permeados por esa concepción de la guerra de la información. Así, por ejemplo, en febrero de 2017, el ministro de Defensa ruso, Sergei Shoigú, admitía en público la existencia de unas “tropas informativas”, pero sin dar más detalles sobre sus actividades u objetivos (BBC, 2017). Y en enero de 2022, es decir, en plena preparación para lanzar la invasión de Ucrania, en una entrevista concedida a Zvezda, —canal de televisión propiedad del Ministerio de Defensa y que está gestionado directamente por él—, el ministro Shoigú declaró sin ambages que “existe una guerra informativa en todos los frentes y no tenemos derecho a perderla”. (Shoigu, 2022)

En definitiva, esta conceptualización de la información como un arma y su uso con fines estratégicos hostiles ha sido ampliamente asumida por el conjunto de la comunidad estratégica rusa y permea toda la acción estatal de la Federación rusa hacia dentro y hacia fuera de sus fronteras. Así, en el Concepto de Política Exterior adoptado en marzo de 2023, se habla del espacio informativo como una “nueva esfera para la acción militar, donde se difumina la línea entre medios militares y no militares de confrontación interestatal”. Y en ese mismo documento se establece como interés nacional y objetivo estratégico de la política exterior rusa el “desarrollo de un espacio informativo seguro para proteger a la sociedad rusa contra las influencias informativas y psicológicas destructivas [...] con el objetivo de que se forme una percepción objetiva sobre Rusia en el exterior, reforzando su posición en el espacio informativo, contrarrestando la campaña de propaganda antirrusa coordinada de forma sistemática por Estados hostiles y que entraña desinformación, difamación e incitación al odio”. Ese es el entorno estratégico que percibe Rusia y en el que, a modo de pretendido espejo, despliega su actividad hostil contra los miembros de la UE y la OTAN. (Ministerio de Asuntos Exteriores de la Federación Rusa, 2023).

EL ECOSISTEMA INTERNACIONAL DE MEDIOS DE DESINFORMACIÓN DEL KREMLIN

Para desarrollar su actividad informativa nacional e internacional, Rusia ha creado un ecosistema de desinformación y propaganda. Este ecosistema, siguiendo la definición del Global Engagement Center del Departamento de Estado de Estados Unidos, consta de cinco pilares: la comunicación oficial del gobierno ruso, los medios estatales de alcance global, el cultivo de fuentes *proxies* o compañeros de viaje, la utilización de las redes sociales como un arma, y la desinformación habilitada por medios cibernéticos. (Global Engagement Center, 2020). Ese ecosistema se ha desarrollado bajo la premisa conceptual de la permanente guerra de la información descrita en el apartado anterior de la que la UE o la OTAN no han sido plenamente conscientes hasta fechas muy recientes. Y eso que una de las lógicas que guía la creación de este ecosistema es la de infiltrarse en el ecosistema occidental y saturarlo con cientos de narrativas que deslegitimen y erosionen las democracias liberales desde dentro y dificulten su proyección internacional.

Este ecosistema proporciona a Rusia algunas ventajas estratégicas:

- Permite la introducción de numerosas variaciones de las mismas narrativas manipulativas, insidiosas o directamente falsas, lo que hace posible afinar los mensajes a diferentes públicos objetivos. La ventaja es que no es necesaria mantener una coherencia narrativa como la que, a pesar de todo, deben tratar de mostrar los medios oficiales, tales como RT o Sputnik.
- Proporciona una negación plausible para el Kremlin cuando los *proxies* difunden información peligrosa, manipulativa o directamente falsa, lo que le permite desviar las críticas (y las posibles responsabilidades) sin dejar de introducir contenidos dañinos.
- Crea un efecto multiplicador de las narrativas, ya que estos medios suelen nutrirse entre ellos, lo que aumenta su alcance, su resonancia y su impacto.
- Estos medios pueden responder con inmediatez a objetivos políticos distintos o circunstancias precisas, para retornar al statu quo de desprecio a los elementos que Rusia percibe como adversarios una vez concluidas sus acciones. Además, permite que, mientras los diplomáticos y los medios de comunicación estatales pueden abordar una cuestión dada desde una perspectiva “oficial”, los supuestos medios “independientes” podrán mostrar variaciones interesadas al respecto, generando una *infodemia* muy útil para el Kremlin.
- Para prevenir esta intervención hostil de los medios rusos, una de las primeras sanciones a Rusia promulgada por el Consejo de la Unión Europea en febrero de 2022 fue, precisamente, el veto a la difusión de Sputnik y RT en todo el territorio

de la UE mientras dure la invasión rusa de Ucrania. El Consejo considera su actuación “esencial para llevar adelante y apoyar la agresión militar contra Ucrania y para la desestabilización de sus países vecinos”. (Consejo Europeo, 2022). Este veto, sin duda, ha dificultado la acción desinformativa de Rusia, pero conviene no perder de vista la capacidad de adaptación y respuesta de Moscú. Así, Rusia crea constantemente nuevas URL para burlar con éxito el veto, también utiliza otras plataformas para difundir sus vídeos o simplemente cambia el nombre de los servicios en las redes sociales. Rusia explota así las dificultades de la UE para actuar con la suficiente agilidad y rapidez y, más aún, a medida que se diluye el impulso inicial que generó la conmoción de la invasión. La UE, no obstante, en junio de 2023 amplió el veto de emisión en el espacio comunitario a algunos de estos medios no oficiales o *proxies* de Rusia, indicando que se disponía de “pruebas de que estos están bajo el control permanente de la dirección rusa y participan sistemáticamente en la difusión de información errónea y propaganda, que constituye una amenaza para el orden público y la seguridad de la Unión” (Consejo Europeo, 2023).

El ecosistema de propaganda de Rusia es extraordinariamente amplio: las estimaciones oscilan entre doscientos y cuatrocientos medios digitales, a lo que cabe sumar otros muchos actores individuales. En este ecosistema están involucrados medios dirigidos a toda clase de audiencias objetivo, ya que el Kremlin fomenta y apoya la difusión de todas aquellas líneas argumentales que puedan provocar polarización en las sociedades democráticas y liberales. Pero independientemente de la naturaleza, de la línea editorial o del espectro político-ideológico del medio, todos estos medios tienen las siguientes características:

- Apoyan de manera constante a Rusia (o a sus principales dirigentes). No existen narrativas de crítica en estos medios hacia las políticas internas y externas de Rusia. Esto incluye que aquellas figuras, organismos y entidades críticas con el régimen de Putin también son ampliamente atacadas por todos estos medios.
- Siempre apoyan las narrativas oficiales de Rusia en cuestiones clave. Cuanto más relevante es un asunto para el Kremlin, mayor será el volumen de actividad desinformativa rusa. Estas acciones irán desde las campañas más sofisticadas—que involucran actividades clandestinas de los servicios de inteligencia— hasta las más burdas con la saturación de un entorno con docenas de narrativas, incluyendo algunas absurdas o manifiestamente ridículas.
- Es recurrente y generalizada la crítica a la Unión Europea, la OTAN y los respectivos gobiernos nacionales —salvo aquellos que se muestren alineados o comprensivos con el Kremlin— con el objetivo claro de provocar inestabilidad y desmoralización desde dentro amplificando las divisiones y tensiones endógenas.

Como se ha indicado más arriba, Russia Today se rebautizó como RT en 2009. En diciembre de ese mismo año, comenzó a emitir su versión en español. Para Rusia, la lengua española tiene un alto valor estratégico tanto por su volumen de hablantes (unos

500 millones, incluyendo cerca de 50 millones en EEUU) como de peso geopolítico (entre otros, por el voto de los países latinoamericanos en la ONU). Los estrategas rusos suelen tener muy presentes las variables demográficas y, según algunos estudios, la comunidad hispanohablante de EEUU podría convertirse en la primera del país en términos absolutos de población. Así que el interés es claro y, de hecho, en los meses de existencia de este Grupo de Trabajo, se aprecian claros signos de cómo la maquinaria de desinformación rusa presta una creciente atención a la comunidad latina de EEUU de cara a las elecciones presidenciales de noviembre de 2024.

Un error de interpretación frecuente es minusvalorar el peso de RT en español tomando como referencia las limitadas audiencias que alcanza por medios tradicionales. Sin embargo, conviene no perder de vista, por un lado, su difusión fragmentada por medio de unos mil proveedores de cable y satélite en España y América Latina de tamaños y alcances muy diversos. A ello cabe añadir, por un lado, los acuerdos suscritos por RT con canales de países como Cuba o Venezuela para que transmitan sus contenidos durante varias horas al día. Por otro lado, y mucho más relevante, el impacto de su actividad en las redes sociales. RT tiene una enorme capacidad de generar contenido y narrativas y de amplificarlas masivamente en las redes. En el caso concreto de América Latina, esa capacidad está especialmente canalizada a través de Facebook, y ha resultado extraordinariamente efectiva en momentos de crisis para desestabilizar gobiernos en Ecuador (2019), Chile (2019-2020) o Colombia (2021).

ALGUNAS CONSIDERACIONES SOBRE LA DESINFORMACIÓN RUSA CONTRA UCRANIA (2014-2022)

Cuanto más relevante y estratégico es un asunto para el Kremlin, más férreamente están alineados todos los medios e instrumentos del ecosistema y más agresivas son sus campañas. El caso de Ucrania resulta, por ello, paradigmático. Así, por ejemplo, alrededor de seis mil de los casi quince mil casos documentados por la unidad del Servicio Europeo de Acción Exterior dedicada a identificar y exponer campañas de desinformación (East StratCom - EUvsDisinfo), tienen que ver con la cuestión ucraniana. En Ucrania se ha ensayado la desinformación más burda y absurda, pero también la más elaborada y perversa. Y en retrospectiva, resulta claro que Rusia —en línea con las doctrinas y conceptos militares analizadas más arriba— preparó con mucha antelación el entorno informativo antes de iniciar la invasión. Y eso incluye el entorno informativo de Ucrania y Rusia, pero también el de los países miembros de la UE y la OTAN.

La evolución de la intervención rusa —de encubierta o no declarada en 2014 a invasión abierta en 2022— también se ha visto reflejada en su actividad desinformativa. Y así, la importancia del contexto informativamente confuso y moralmente ambiguo puede comprobarse en el desigual resultado de la desinformación rusa acerca de Ucrania en 2014 y 2022. La grave crisis institucional en Ucrania tras las revueltas del Euromaidán se presentó como un golpe de Estado de carácter neonazi, haciendo referencia desde entonces al “régimen de Kiev”. La infiltración de agentes, agitadores y combatientes desde Rusia avivó en Ucrania oriental una revuelta que generó un conflicto armado en el que intervinieron tropas regulares rusas. Pero al contrario que ocho años más tarde, la respuesta occidental fue limitada y apática a pesar de la acumulación de pruebas sobre la presencia militar rusa en suelo soberano ucraniano.

Un caso paradigmático de la desinformación rusa y de su impacto se produjo con el derribo del vuelo MH17 de la aerolínea Malaysia Airlines que sobrevolaba Ucrania el 17 de julio de 2014 mientras hacía la ruta Ámsterdam-Kuala Lumpur. Fallecieron los 283 pasajeros y los 15 tripulantes del avión. El derribo del vuelo MH17 fue anunciado por las agencias de noticias y los medios rusos en primer lugar como el derribo de un avión de transporte militar ucraniano Antonov An-26. Combatientes de la insurgencia rusa se fotografiaron sonrientes con restos del avión malasio. Cuando la identidad del aparato fue descubierta y las autoridades rusas se percataron del error, lanzaron una campaña de desinformación masiva con vistas a ofuscar lo que constituía un crimen de guerra.

Así, el ecosistema de medios rusos puso rápidamente en circulación numerosas hipótesis con el objetivo de diluir la responsabilidad rusa y poner el foco sobre Ucrania. De este modo, según las diferentes versiones rusas, el vuelo MH17 fue derribado por un sistema antiaéreo ucraniano que se coló tras las líneas separatistas. Otra versión atribuía el derribo a una

confusión ucraniana por la similitud de los colores del aparato con el empleado por Vladímir Putin. Una tercera mencionaba concretamente a un aparato Sukhoi Su-25. Ninguna era plausible por diferentes razones técnicas, pero contribuyeron a enturbiar la cobertura informativa del suceso y en España tuvieron un eco significativo.

Este eco tenía que ver, también, con el hecho de que una de las principales operaciones para encubrir la autoría del derribo se lanzó desde RT en español y utilizó un activo desinformativo introducido por los rusos con semanas de antelación. Así, el mismo día del derribo, uno de los presentadores estrella de la cadena, Javier Carrasco, se hacía eco de las supuestas informaciones que transmitía a través de su cuenta de Twitter, Carlos Spainbuca, un presunto controlador aéreo español en la torre del aeropuerto de Boryspil en Kiev. Según este individuo, había al menos dos cazas ucranianos que escoltaban al Boeing malasio instantes antes de su derribo. RT había popularizado y contribuido a dar legitimidad a la declaración de este personaje, al ser entrevistado, semanas antes, por Gonzalo Wancha, otro periodista español por aquel entonces en la nómina de RT. El presunto controlador aparecía con su rostro difuminado para proteger su identidad y, supuestamente, su seguridad y la de su familia que estaba, según él, siendo amenazada por su opinión crítica sobre las protestas del Maidán. Así que, en el momento del derribo, el activo había sido introducido desde mucho antes y resultó clave en la rápida ofuscación y posterior encubrimiento de este crimen de guerra.

El bulo del supuesto del derribo del MH17 por cazas ucranianos sigue circulando a día de hoy por internet. El propio presidente Putin se refirió al “especialista español” que informó de la presencia de “aviones militares en el corredor asignado para la aviación civil” durante sus entrevistas con el cineasta estadounidense, Oliver Stone. El corte no se incluyó en la versión distribuida del documental, pero figura en la versión publicada como libro que incluye la transcripción completa de las varias horas que pasaron juntos (Putin, 2017). Por investigaciones periodísticas posteriores, hoy se sabe que Carlos Spainbuca —ciudadano español cuyo nombre real es José Carlos Barrios Sánchez— ni era controlador ni estaba en Ucrania y, según declaró él mismo, había recibido importantes sumas de dinero procedentes de Rusia e indicaciones de lo que debía publicar en su cuenta de Twitter. Hay que indicar que este punto era ya conocido cuando el presidente Putin mantuvo su encuentro con Oliver Stone por lo que resulta difícil creer que se trate de una simple equivocación.

Se trata, pues, de un caso especialmente grave por tratarse del encubrimiento de un crimen de guerra con 298 víctimas y significativo en cuanto al alcance y dificultades para contrarrestar la desinformación rusa. Así, pese a que investigaciones judiciales, forenses y de fuentes abiertas han identificado, incluso, a la unidad militar rusa responsable del misil que derribó el avión, en el debate público persiste la impresión de que es un asunto sin aclarar. Y ahí está el quid del asunto: el efecto acumulativo y erosionante de la desinformación rusa que busca deslegitimar y minar la confianza y la credibilidad en nuestro debate público. No se trata pues de ruido molesto en las redes sociales, sino de algo más profundo y peligroso.

Conviene insistir que, en el caso de la desinformación rusa, el empleo de diferentes teorías, aunque contradictorias y poco plausibles, tiene la finalidad de saturar el espacio informativo al ser lanzadas con un efecto de “manguera de riego”. Esto” (*firehose of falsehood*) obliga a los verificadores de noticias a atender a una sucesión rápida de narrativas, datos y opiniones. El objetivo es dejar en la opinión pública un poso de desconfianza sobre las versiones presentadas por los medios occidentales bajo la premisa de que de entre tanta versión procedente de Rusia debe haber un trasfondo de verdad.

Ocho años después de la primera invasión rusa de Ucrania, que se saldó con la ocupación de la península de Crimea, Rusia volvió a invadir el país vecino. La preparación de la llamada “Operación militar especial” implicó convencer a la opinión pública occidental, y en especial a la sociedad ucraniana, de que no se iba a producir invasión alguna y de que las advertencias procedentes de los servicios de inteligencia de Estados Unidos y el Reino Unido formaban parte de una campaña de demonización de Rusia. Según los medios rusos y sus *proxies* españoles, todo era fruto de la “histeria” o de la “manipulación anglosajona”. Cuando finalmente la invasión se produjo el 24 de febrero de 2022, diplomáticos y periodistas rusos, junto con opinadores al servicio del Kremlin, fueron puestos en evidencia por la contundencia de los hechos. La agresión rusa a Ucrania fue juzgada severamente en los medios de comunicación. La confusión informativa y la ambigüedad moral que había explotado Rusia en otras ocasiones esta vez había dado paso a una realidad incontestable desde el punto de vista factual que provocó un rechazo moral unánime.

Esta situación dejó, por vez primera en una década, a Rusia en fuera de juego en el ámbito informativo. La credibilidad de sus diplomáticos, medios de comunicación y demás actores e instrumentos del ecosistema de desinformación ruso había quedado seriamente dañada. Sin embargo, transcurridos veinte meses de guerra contra Ucrania, Rusia ha recuperado la iniciativa en el ámbito informativo. De nuevo, lo ha conseguido por medio de la saturación narrativa y la adaptación permanente de sus medios digitales. Ha creado constantemente nuevas páginas web o alimentado nuevos personajes como, por ejemplo, recurrir durante semanas a *influencers* sin experiencia previa en asuntos internacionales. En el denominado Sur Global han resonado con fuerza las narrativas desinformativas y bulos rusos relativos a la cuestión del grano ucraniano. En Europa y EEUU se han asumido algunas premisas como que “Rusia no puede ser derrotada”, “la escalada nuclear puede ser inminente”, etc. que hacen que el Kremlin siga convencido de que solo es cuestión de tiempo que alcance una victoria estratégica en Ucrania por el simple agotamiento o miedo en las capitales de las potencias euroatlánticas.

UN ESTUDIO DE CASO SOBRE SUPLANTACIÓN DE MEDIOS E INSTITUCIONES

La campaña de desinformación e influencia bautizada como *Doppelganger* lanzada en plena guerra de Ucrania es un buen ejemplo para comprender mejor el funcionamiento del ecosistema de propaganda ruso, su impacto en países como Francia o Alemania, su persistencia en el tiempo pese a haber sido expuesta y las medidas —claramente insuficientes— adoptadas para frenarla.

En septiembre de 2022, la organización no gubernamental belga EU DisinfoLab, con el apoyo de la sueca Quriium, exponía y bautizaba con el término *Doppelganger* una campaña de influencia prorrusa que llevaba operando en Europa desde, al menos, el mes de mayo de ese mismo año. Esta campaña suplantaba las páginas web de importantes medios de comunicación de diversos países (Alaphilippe et al., 2022). Varios medios alemanes ya habían alertado de la existencia de falsas URL de prestigiosas cabeceras —como *Spiegel*, *FAZ*, *SZ* o *Die Welt*— circulando en varias redes sociales y publicando contenido de apoyo a Rusia en la guerra de Ucrania. Por ejemplo, reforzaba las acusaciones de nazismo a los dirigentes y ciudadanos ucranianos. También intentaba posicionar a la población alemana en contra de las sanciones a Moscú, del apoyo de Berlín a Kiev, o de la acogida de refugiados, entre otros temas.

Las falsificaciones consistían en la clonación de páginas web en las que se introducía contenido falso de baja calidad en distintos formatos, principalmente encuestas, artículos y vídeos. El sistema permitía acceder a las páginas web originales si se seguía navegando y haciendo clic en otro lugar de la página, lo que les dotaba de mayor credibilidad. Falsas URL simulaban las páginas web auténticas (una práctica conocida como *cybersquatting*), modificando levemente la extensión del dominio y pasaban fácilmente desapercibidos para el ojo del lector desprevenido. Además, la campaña elegía específicamente a sus audiencias con un sistema de bloqueo geográfico incorporado en algunas de las URL que no permitía leer el contenido fuera del país en cuestión.

Según expuso EU DisinfoLab en aquel momento, decenas de nombres de dominios de internet habían sido registrados siguiendo ese mismo patrón y simulando más de diecisiete medios de comunicación de diferentes países —por ejemplo, *Bild*, *20minutes*, *Ansa*, *The Guardian* o *RBC Ukraine*—, aunque la mayoría eran alemanes. Muchos de ellos habían sido ya utilizados en URL concretas difundiendo noticias falsas, pero de muchos otros no se había encontrado rastro. No está claro si ya habían sido utilizados y su huella borrada, o si los responsables de la campaña los guardaban para lanzar contenido falso en un momento posterior.

El contenido falsificado era después amplificado en varias redes sociales, como Twitter o Telegram, pero fue Facebook quien desempeñó un papel clave en su distribución. Esta se vio facilitada por una red de perfiles falsos y anuncios pagados por al menos 115.000 dólares (según cálculos iniciales muy por lo bajo). La propia Meta anunció en septiembre de 2022 la desarticulación de una campaña rusa de comportamiento inauténtico coordinado — la misma *Doppelgänger*— y anunció que actuaría contra los dominios detectados (Nimmo y Torrey, 2022).

Como suele ocurrir en este tipo de operaciones, pese a que en un primer momento numerosos indicios apuntaban a la mano directa de Moscú, la atribución no era sencilla. Una adquisición de dominios coordinada y la producción de vídeos apuntaba a Rusia, pero eran sobre todo las narrativas las que se alineaban con la estrategia de Moscú. Las noticias describían a Ucrania como un estado fallido, corrupto y nazi o negaban masacres como la de Bucha mientras asustaban a ciudadanos alemanes, italianos, franceses, británicos o letones asegurando que las sanciones contra Moscú arruinarían su economía y sus vidas.

Fue la propia empresa Meta la que meses después, en diciembre de 2022, señaló claramente a dos compañías rusas: Structura National Technologies y Social Design Agency (META, 2022). Ambas fueron sancionadas por la Unión Europea en julio de 2023, acusadas de manipulación coordinada y selectiva de la información en el marco de “una campaña híbrida más amplia de Rusia contra la UE y los Estados miembros”. En el comunicado, la UE aseguraba que “organismos gubernamentales o vinculados al Estado ruso” habían participado en la campaña (Consejo Europeo, 2023). También Francia apuntó directamente a Rusia en junio de 2023, cuando la agencia francesa para la lucha contra la injerencia extranjera en internet VIGINUM, expuso nuevas suplantaciones de los medios *Le Monde*, *Le Figaro* y *Le Parisien*, y el Ministerio de Exteriores galo acusó directamente a Rusia de estar implicada, al menos, en la distribución del contenido (VIGINUM, 2023).

Más de un año después de que la campaña fuera expuesta por primera vez y en torno a un año y medio desde que se puso en marcha, muchos de los dominios han sido inhabilitados y están inaccesibles, pero, lejos de cesar, la campaña se ha expandido a otros países, sigue circulando en redes sociales y ha alcanzado un nuevo nivel al conseguir usurpar la identidad ya no solo de medios de comunicación, sino de instituciones públicas nacionales e incluso organizaciones internacionales.

Francia denunció la suplantación de páginas web de instituciones públicas, entre ellas la del Ministerio de Exteriores, mientras en Alemania la suplantación afectó a la cartera del Interior. En agosto, una nueva investigación de la organización Graphika y el último informe trimestral de amenazas de Meta confirmaban incluso que la OTAN había sido víctima de suplantación en el contexto de la cumbre de la Alianza Atlántica celebrada un mes antes en Vilna. Además, las últimas investigaciones apuntan a la expansión de la campaña a medios en Estados Unidos —como el *Washington Post* y *Fox News*- o en Israel. Entre tanto, algunos

medios como el alemán *Süddeutsche Zeitung* o el francés *Le Parisien* han acudido a los tribunales para denunciar lo ocurrido y las autoridades de Francia y Alemania afectadas aseguran haber tomado medidas cuyo alcance se desconoce (Ronzaud et al., 2023; Nimmo et al., 2023).

Mientras, la campaña continúa con links activos circulando en varias redes, entre otras en Twitter, con un complejo sistema de redireccionamiento de enlaces que oculta los dominios suplantados. El contenido falso sigue circulando difundiendo narrativas más actuales y acordes a la evolución de los acontecimientos. Recientemente hemos observado también la distribución de contenido muy similar en forma de capturas de pantalla de noticias de medios falsificados sin que existieran falsas URL creadas con ese fin. Si bien no pueden atribuirse a la misma campaña por la diferencia en el *modus operandi* utilizado, es posible que otros actores se hayan inspirado en *Doppelganger* para difundir narrativas similares de una forma más sencilla sin la necesidad de la estructura web requerida para crear los dominios.

Si bien no se tiene constancia de que la campaña haya llegado a España o haya suplantado medios en español, vista la reciente expansión a otros países y el carácter persistente de la campaña, conviene no bajar la guardia.

Doppelganger permite sacar algunas conclusiones y aprender de las tácticas híbridas de influencia rusa. En primer lugar, es un gran ejemplo de cómo Rusia utiliza su propaganda no solo para reforzar sus narrativas en el interior del país, sino para buscar apoyos en el exterior y, más aún, para polarizar y sembrar inestabilidad en países que percibe como hostiles. Para ello, explota narrativas nacionales y busca aliados internos en esos países -a veces aliados directos y otros indirectos- que refuercen su estrategia. En Alemania, por ejemplo, las noticias falsas creadas no solo tocaban de forma muy directa el tema de la guerra en Ucrania, reforzando las acusaciones de nazismo a los ucranianos y posicionándose en contra de las sanciones a Rusia, sino que explotaban especialmente temas sensibles en el país como la inmigración o la crisis económica y energética que afectan de forma muy tangible al bienestar y el bolsillo de los ciudadanos.

La suplantación de medios de comunicación es otra prueba de la siembra del descrédito institucional como estrategia central de Rusia a la hora de minar las democracias liberales. Si los medios constituyen órganos centrales para el correcto funcionamiento de las democracias, suplantarlos tiene un doble efecto que posiciona sí o sí a los actores maliciosos como ganadores de la partida. En primer lugar, explota su credibilidad labrada como mediadores de los acontecimientos para dotar de mayor veracidad al contenido falso difundido. En segundo lugar, mina esa credibilidad con un discurso polarizador, radical e incluso con los errores y la baja calidad del contenido falso insertado. La estrategia se multiplica en el caso de la suplantación de instituciones.

Doppelganger ilustra también cómo campañas que pueden no tener un gran impacto en un primer momento, manteniéndose con un perfil relativamente bajo en redes sociales, pueden expandirse y persistir en el tiempo y minar la opinión pública de forma más paulatina, aunque no está claro que esta haya sido la estrategia elegida intencionalmente por los responsables de la campaña. Desde su creación, *Doppelganger* se ha convertido en una campaña internacional, plurilingüe y multiplataforma que, pese a no tener un gran impacto en términos de viralidad, no desaparece.

BIBLIOGRAFÍA

Alaphilippe, A., Machado, G., Miguel, R., y Poldi, F. (2022). *Doppelganger: Media clones serving Russian propaganda*. Obtenido de <https://www.disinfo.eu/wp-content/uploads/2022/09/Doppelganger-1.pdf>

BBC. (23 de Febrero de 2017). Russian military admits significant cyber-war effort. Obtenido de <https://www.bbc.com/news/world-europe-39062663>

Consejo Europeo. (2022). DECISIÓN (PESC) 2022/351 DEL CONSEJO. Obtenido de <https://eur-lex.europa.eu/legal-content/ES/TXT/PDF/?uri=OJ:L:2022:065:FULL>

Consejo Europeo. (2023). Decisión (PESC) 2023/1566 del Consejo de 28 de julio de 2023 relativa a medidas restrictivas respecto de acciones que menoscaban o amenazan la integridad territorial, la soberanía y la independencia de Ucrania. Obtenido de <https://www.boe.es/buscar/doc.php?id=DOUE-M-2023-81102>

Consejo Europeo. (23 de Junio de 2023). Russia's war of aggression against Ukraine: EU adopts 11th package of economic and individual sanctions. Obtenido de <https://www.consilium.europa.eu/en/press/press-releases/2023/06/23/russia-s-war-of-aggression-against-ukraine-eu-adopts-11th-package-of-economic-and-individual-sanctions/>

Gerasimov, V. (27 de Febrero de 2013). The Value of Science Is in the Foresight. (R. Coalson, Trad.) Military-Industrial Kurier. Obtenido de https://www.armyupress.army.mil/portals/7/military-review/archives/english/militaryreview_20160228_art008.pdf

Global Engagement Center. (2020). *Pillars of Russia's Disinformation and Propaganda Ecosystem*. Department of State. Obtenido de https://www.state.gov/wp-content/uploads/2020/08/Pillars-of-Russia%E2%80%99s-Disinformation-and-Propaganda-Ecosystem_08-04-20.pdf

META. (2022). *Recapping Our 2022 Coordinated Inauthentic Behavior Enforcements*. Obtenido de <https://about.fb.com/news/2022/12/metax-2022-coordinated-inauthentic-behavior-enforcements/>

Ministerio de Defensa Ruso. (2011). Conceptual Views Regarding the Activities of. (N. C. COE, Trad.) Obtenido de https://ccdcoe.org/uploads/2018/10/Russian_Federation_unofficial_translation.pdf

Nimmo, B., y Torrey, M. (2022). *Taking down coordinated inauthentic behavior from Russia and China*. Obtenido de https://about.fb.com/wp-content/uploads/2022/11/CIB-Report_-China-Russia-Sept-2022.pdf

Nimmo, B., Torrey, M., Franklin, M., Agranovich, D., Milam, M., Hundley, L., y Flaim, R. (2023). *Adversarial Threat Report: Second Quarter*. Obtenido de https://scontent-mad2-1.xx.fbcdn.net/v/t39.8562-6/10000000_180063885098584_6098733693167598956_n.pdf?_nc_cat=110&ccb=1-7&_nc_sid=b8d81d&_nc_ohc=5xodtcIndrEAX8K-Y1a&_nc_ht=scontent-mad2-1.xx&oh=00_AfB_Lnl6Om6OFfW4Z1ch0NILj4eJKh55ytBKTRATYClg&oe=6548F8AE

Presidente de la Federación de Rusia. (9 de Septiembre de 2000). Doctrina de seguridad de la información de la Federación de Rusia. Obtenido de <https://base.garant.ru/182535/>

Presidente de la Federación de Rusia. (2010). Doctrina Militar de la Federación de Rusia. Obtenido de <https://web.archive.org/web/20110504070127/http://www.scrf.gov.ru/documents/33.html>

Putin, V. (Diciembre de 2011). Russia's Prime Minister accuses Secretary of State Hillary Clinton of inciting election protests. (N. Y. Times, Trad.) Obtenido de <https://www.nytimes.com/video/world/europe/100000001214703/timescast--putin-blames-sec-clinton.html?smid=url-share>

Putin, V. (Marzo de 2012). Vladimir Putin celebrates return to Russian presidency. Moscú. Obtenido de <https://youtu.be/4rfJpEngvnc?feature=shared>

Putin, V. (2017). The full transcripts of the Putin interviews : Oliver Stone interviews Vladimir Putin. (O. Stone, Entrevistador)

Ronzaud, L., Carter, J. A., y Williams, T. (2023). Summit Old, Summit New: Russia-Linked Actors Leverage New and Old Tactics in Influence Operations Targeting Online Conversations About NATO Summit. Obtenido de https://public-assets.graphika.com/reports/graphika_report_summit_old_summit_new.pdf

Shoigu, S. (9 de Enero de 2022). Sergei Shoigu claims Russia is the victim in information war. (M. Acceptance, Entrevistador) TZ Zvezda. Obtenido de <https://www.youtube.com/watch?v=iQ0nvZbOX8k>

Simonyan, M. (7 de Abril de 2012). "No hay objetividad". (A. Gabuev, Entrevistador) *kommersant.ru*. Obtenido de <https://web.archive.org/web/20210427193514/https://www.kommersant.ru/doc/1911336>

Trenin, D. (Agosto de 2017). Demands on Russian Foreign Policy And Its Drivers: Looking Out Five Years. Obtenido de <https://carnegiemoscow.org/commentary/72799>

VIGINUM. (2023). RRN: *Une campagne numérique de manipulation de l'information complexe et persistante*. Obtenido de <https://www.sgdsn.gouv.fr/publications/maj-19062023-rrn-une-campagne-numerique-de-manipulation-de-linformation-complexe-et>

Capítulo 3

EFFECTOS PSICOLÓGICOS,
RADICALIZACIÓN Y AMPLIFICACIÓN DE
LOS CONFLICTOS SOCIALES CAUSADOS
POR LA INJERENCIA Y MANIPULACIÓN
EXTRANJERA DE LA INFORMACIÓN:
ANÁLISIS Y PREVENCIÓN



Coordinadores:

Sara Degli Esposti

Ministerio del Interior - Oficina de Coordinación de Ciberseguridad (OCC)

Autores y colaboradores:

Amalio Blanco Abarca

Maria Alberta Chulvi Ferriols

Sergio Hernández García

Nuria Lorenzo-Dus

Juan Antonio Pérez Pérez

Ministerio del Interior - Centro de Inteligencia contra el Terrorismo y el Crimen Organizado (CITCO)

Ministerio de Defensa - Centro Superior de Estudios de la defensa Nacional (CESEDEN)



INTRODUCCIÓN

El principal objeto de este capítulo es analizar el efecto de la desinformación en la amplificación de los conflictos sociales y la radicalización, a fin de encontrar pautas para tratar de prevenirla. Se entiende por desinformación la «información verificablemente falsa o engañosa que se crea, presenta y divulga con fines lucrativos o para inducir a error deliberadamente a la población, y que puede causar un perjuicio público» (HLEG 2018) como consecuencia de la injerencia y manipulación procedente del exterior. Este tipo de (des)información “supone una amenaza para los valores constitucionales, los procesos democráticos, las instituciones democráticamente constituidas y, por ende, para la seguridad nacional” (Ministerio de la Presidencia, Relaciones con las Cortes y Memoria Democrática, 2022, p. 7). Aquellos que recurren a esas técnicas de control de la información pretenden influir en las percepciones, opiniones, creencias, actitudes y comportamientos de los ciudadanos para lograr algún fin (Wardle y Derakhshan, 2017). La (des)información puede manifestarse en forma de bulos expresados a través de bromas, exageraciones, descontextualizaciones o engaños (Salaverría et al., 2020). En este capítulo trataremos solo tangencialmente el tema de las múltiples formas de las que se reviste, se materializa y se difunde la desinformación, para centrarnos en el análisis de los mecanismos psicosociales que nos permitan entender lo que lleva a determinados grupos y colectivos a creer verse influidos por informaciones sesgadas, manipuladas, descontextualizadas o alejadas de hechos objetivos y verificables.

A la hora de abordar el problema de la desinformación en España deberíamos tener en cuenta que, de acuerdo con el Standard Eurobarometer 96 (Comisión Europea, 2022b), el 81% de los encuestados en febrero de 2022 declaraba encontrarse a menudo con “noticias o informaciones que tergiversan la realidad”, y el 82% consideraba la “desinformación un problema para la democracia”. Además, solamente el 54% pensaba estar algo o muy seguros de poder identificar las noticias falsas. Otra encuesta realizada por el Instituto Reuters destaca la falta de confianza en los medios de comunicación en España; tan solo un 32% de los encuestados declaraba confiar casi siempre en ellos, y un 56% decidió utilizar las redes sociales como fuente de información (Newman et al., 2022).

Los resultados de otra encuesta realizada en España (febrero de 2021) en el marco del proyecto H2020 TRESCA,¹ con una muestra representativa de 1009 personas de todas las edades, indican que el 40% de los encuestados declaraba confiar en los medios de comunicación, mientras que un 33% afirmaba depositar su confianza en las redes sociales. De cara al análisis de la desinformación, llama la atención que el 28% de los participantes declarase estar parcial o totalmente de acuerdo con la opinión de que el ser humano no aterrizó en la luna en julio de 1969, sino que las imágenes difundidas provenían de una grabación realizada en un estudio de televisión. Esta idea era especialmente compartida entre los menores de 24 años. En relación también con en esta misma encuesta, un 82% de los participantes manifestaba la convicción de que mucho de lo que sucede en el mundo lo decide un pequeño grupo de personas que actúan en secreto, típica creencia conspirativa; el 90% llegaba a opinar que a menudo “las grandes empresas o los grupos de interés sobornan a los políticos”, mientras que el 47% se mostraba favorable a que los grupos en el poder “tomen decisiones de forma unilateral en lugar de intentar alcanzar amplios acuerdos”.

Resultados similares se encuentran en la encuesta de la FECYT (2022) realizada en noviembre de 2022 a través de entrevistas telefónicas a 2.100 personas mayores de 15 años residentes en España. En ella, el 88% cree que nunca se informa sobre muchas de las cosas importantes que suceden en el mundo. Además, el 68% piensa que “existen organizaciones secretas que ejercen una notable influencia en las decisiones políticas”. Esta encuesta nos permite además identificar grupos con creencias pseudocientíficas: hay un 13% de los encuestados que sostienen que el sol gira alrededor de la tierra; por su parte, un 16% opina que el cambio climático es una consecuencia del agujero de la capa de ozono (y no de la acumulación de gases de efecto invernadero), y un 16% piensa que las noticias y el contenido que una persona visualiza en Facebook se selecciona aleatoriamente. Como veremos a lo largo del capítulo, es importante no subestimar las creencias y comportamientos insólitos de determinados grupos, porque pueden terminar influyendo en las opiniones y conductas de los demás. Es decir, los juicios e ideas de una minoría pueden cambiar la forma de pensar de los miembros de una sociedad hasta convertirse en pensamiento mayoritario.

Es sobradamente sabido que negar que la Tierra era una esfera inmóvil rodeada por otros cuerpos celestes esféricos que giran a su alrededor (teoría heliocéntrica de Nicolás Copérnico) llevó Galileo Galilei a ser condenado por herejía en 1633. Hoy en día la ciencia ha rechazado dichas creencias; sin embargo, eso no ha impedido que se sigan generando afirmaciones pseudocientíficas. Además, a pesar de profundos cambios en los procesos y prácticas sociales —incluidas las comunicativas—, las dinámicas identitarias y grupales siguen influyendo en la invención de creencias y su transformación en verdades normativas en el seno de un determinado grupo social. Más de cien años de investigación en el campo de la psicología social nos permiten afirmar que el grupo es el principal y el más poderoso

¹ Véase en la página web <https://trescaproject.eu/>

escenario de influencia en el terreno de los valores, en el de las creencias, en el de las emociones y, a resultas de todo ello, en el del comportamiento de sus miembros. En muchas situaciones a lo largo de su vida, una persona “siente, piensa y se comporta en términos de su identidad social y en términos de la identidad social de los demás”, bien considerados a título personal o como miembros de grupos o categorías sociales (Tajfel, 1984, p. 52). Como veremos más adelante hablando de las dinámicas grupales, cualquier excusa es válida para promover un conflicto intergrupal, y la pertenencia a un determinado grupo ha sido históricamente una de las más frecuentes.

Dentro del vasto mundo de comportamientos colectivos que la desinformación es capaz de favorecer, podemos distinguir dos tipos de procesos atendiendo a sus objetivos: por un lado, la desinformación que pretende influir en la opinión pública y, por otro, la desinformación cuyo objetivo es contribuir a la radicalización de personas y de grupos. Ambos procesos preocupan a las instituciones democráticas. El primero, porque disuelve la confianza en dichas instituciones y obstaculiza el debate plural centrado sobre los contenidos. El segundo, porque abre puertas a comportamientos personales y acciones colectivas que llevan la marca de la violencia y del terror.

Este capítulo se centra, especialmente, en el análisis de los factores psicosociales vinculados a la desinformación que pretenden influir en la opinión pública, aunque dedicamos los primeros apartados a analizar los contextos que permiten conectar las estrategias de la desinformación con análisis más amplios en el marco de la seguridad nacional y de la radicalización de las personas o de los grupos. En ambos escenarios, la desinformación parece estar jugando un papel clave, y está captando la atención tanto de las instituciones dedicadas a la investigación como de aquellas que protagonizan el marco operativo de lucha contra la radicalización. Este texto trata de conjugar ambas aproximaciones.

En el ecosistema comunicativo actual hay elementos pertenecientes a la forma de comunicar que seguramente han cambiado a raíz de la proliferación de las redes sociales y de la transición al entorno digital; junto a ellos, también existen elementos de continuidad, especialmente en las dinámicas grupales y en las fisuras histórico-sociales, que contribuyen al éxito de las campañas de desinformación. La relevancia de estas dinámicas se deriva principalmente de la satisfacción de necesidades psicológicas tan vitales como la afiliación, el sentimiento de pertenencia, la identidad social, el apoyo y la protección emocional.

En estas páginas se analiza, en primer lugar, las posibilidades que la desinformación ofrece a la injerencia extranjera en el nuevo contexto de guerra cognitiva. Se revisan, brevemente, las aportaciones que tratan de comprender la radicalización de individuos y grupos en ese marco, sobre todo aquellas que permiten conectar esos procesos más extremos de radicalización con las dinámicas grupales que alientan la circulación de la desinformación. A continuación, se analizan los factores psicosociales que explican la vulnerabilidad a la desinformación y la propensión a confiar en sus contenidos, haciendo hincapié en las estrategias de respuesta más eficaces.

EL CONTEXTO INTERNACIONAL Y EL *SHARP POWER*

En el mundo actual, el hecho de que los conflictos locales se pueden globalizar influye gravemente en las agendas políticas internacionales. Los Estados cooperan en algunos asuntos y compiten en otros en el contexto de una intrincada madeja de intereses y relaciones. Esto dificulta las alianzas, pero también la guerra, pues hace muy forzado el alineamiento schmittiano amigo-enemigo y la ruptura total entre las partes que éste demanda. A ello se suma que la globalización también complica la distinción entre dentro y fuera, entre lo interior y lo exterior. Hoy resulta complejo deslindar la seguridad nacional de la internacional, pues actualmente la amenaza se presenta como “interméstica” —palabra que pretende subrayar su naturaleza internacional y doméstica al mismo tiempo— y se caracteriza por “la ausencia de amenazas en la frontera y la ausencia de fronteras para la amenaza”. Este planteamiento añade aún más dificultades a la conceptualización de seguridad (Martínez y Tulchin, 2006, p. 39).

A fin de describir esta situación, en 2017 se introdujo el término *Sharp Power* (Milosevich-Juaristi, 2017),² cuyo precedente son las llamadas “medidas activas” o “guerra política”, una estrategia de la antigua Unión Soviética y aún anterior. La estrategia del *Sharp Power* supone una redefinición del terreno de enfrentamiento. Esto significa la inclusión de nuevos dominios de colisión, como sería el *Dominio Cognitivo*, o el referido a la ciberseguridad, que se añaden a los ya existentes (tierra, aire y marítimo). La actividad esencial en el *Dominio Cognitivo* es la influencia, que se obtendría mediante la gestión de la información (García Servet y Calvo Albero, 2022). El *Dominio Cognitivo* —tal y como se conceptualiza en la literatura sobre seguridad— sería aquel que incluye las percepciones, creencias, comportamientos y toma de decisiones de los seres humanos y la influencia externa que se puede ejercer sobre estos aspectos para modificarlos, todo ello considerando al ser humano en su doble vertiente, individual y social. La respuesta en este plano, la *Seguridad Epistémica*, se refiere a la protección y salvaguarda de verdades y valores sostenidos por individuos y sociedades en general. Esto supone asegurar los procesos de producción y distribución de información fiable, de modo que haga posible su correcta valoración contextual en los procesos de decisión. De ahí que el acceso a una información fiable sea crucial para las sociedades democráticas, convirtiéndose en una cuestión de Seguridad Nacional.

En este contexto, el *Sharp Power* viene a describir la utilización de las herramientas de poder por regímenes autoritarios que se sirven de los procedimientos del *Soft Power* (la atracción

² El término apareció en un artículo de la revista *Foreign Affairs* de la mano de Christopher Walker y Jessica Ludwig, miembros ambos del *Think Tank National Endowment for Democracy*, y posteriormente sería validado por Joseph Nye.

por la cultura y el sistema político), así como de otros que no son propiamente coercitivos, pero tampoco amistosos. Algunos autores lo denominan “la guerra no observable”, subrayando con ello la naturaleza no violenta, aunque hostil de su proceder. Otros ejemplos de herramientas típicas del *Sharp Power* son también la corrupción o la coerción con el suministro energético. Con estas actuaciones no se trata solo de condicionar a la opinión pública o a los líderes, sino también de influir para que no se emprenda una acción concreta. Además, el establecimiento de una relación de dependencia puede traer aparejada la inducción de una vulnerabilidad a corto o medio plazo sobre la que progresar hasta que adquiera valor estratégico.

Las praxis del *Sharp Power*, en lo que se ha venido en llamar *Dominio Cognitivo*, consisten frecuentemente en la instrumentación de los medios de comunicación social, además de otras acciones encubiertas y ejecutadas muchas veces por organizaciones sin una ligazón clara con la organización principal (Prida, 2017). Así, en Estados no completamente democráticos, aparecen instituciones y organismos oficiales junto a otros que no lo son y que se presentan como entes privados, por más que actúen coordinados o en connivencia con las propuestas políticas de los primeros. También se observa cómo empresas surgidas de estos países ocupan no pocas veces sectores dedicados a la tecnología y la información, resultando difícil acreditar su naturaleza o vinculación estatal y atribuir responsabilidades ante ciertas prácticas que pueden comprometer la seguridad, especialmente en el nebuloso mundo de las redes e internet.

En muchas ocasiones, el *Sharp Power* sitúa a su objetivo frente a sus propias contradicciones internas, con lo que contribuye a resaltarlas y a debilitar el sistema. Utiliza torticeramente el marco normativo y cultural establecido sirviéndose del pluralismo para deshacer los equilibrios sociales y provocar que los grupos internos colisionen entre sí. Los efectos de este proceder son claros: manipulación, decepción y engaño. Es decir, los elementos característicos de la guerra de la información, cuyo objetivo es socavar el poder de las instituciones provocando así su desarticulación y debilitamiento. Las vulnerabilidades de nuestras sociedades que instrumenta el *Sharp Power* son reales: derrotar al enemigo desde dentro fomentando el disenso y la desunión, o haciéndole adoptar decisiones inadecuadas, es tan antiguo como la guerra. No obstante, una estrategia de influencia social no es una guerra; se diferencia de una agresión física en que ésta es un hecho fácilmente objetivable y se sustancia de otra manera. Este tipo de estrategias de influencia social resultan difíciles de afrontar, porque no son propiamente amistosas, pero tampoco abiertamente hostiles. La desinformación comparte estas características; por eso, desde las esferas que se preocupan por la seguridad nacional es vista como un factor que, sin llegar a constituir una acción militar, busca, sin embargo, la colonización mental de un grupo o sector concreto. La dificultad de combatir estas amenazas en un país democrático es evidente.

A través de las redes sociales, la ciudadanía está sobreexpuesta a la influencia de un mayor número de actores con intereses particulares que instrumentan en beneficio propio las oportunidades de expresión y comunicación que brinda internet. La información como

exceso (sobrecarga informativa) define dos neologismos, ya aceptados: el de “infoxicación”³ y el de “infodemia”.⁴ La Organización Mundial de la Salud (OMS) define este último como “una sobreabundancia de información, en línea o en otros formatos, que incluye los intentos deliberados por difundir información errónea para socavar la respuesta de salud pública y promover otros intereses de determinados grupos o personas. La información errónea y falsa puede perjudicar la salud física y mental de las personas, incrementar la estigmatización, amenazar los valiosos logros conseguidos en materia de salud y espolear el incumplimiento de las medidas de salud pública, lo que reduce su eficacia y pone en peligro la capacidad de los países de frenar la pandemia”. Dado que en las sociedades digitales la difusión de contenidos con el fin de desinformar se ha vuelto incontrolable, la instrumentación de las crisis informativas en beneficio de determinados intereses puede hacer descarrilar a una sociedad en su conjunto y generar repercusiones aún mayores, laminando a las instituciones del país —implicadas o no— y su cohesión social.

Un momento particularmente notable de crisis y perturbación, por las consecuencias que implica, son los procesos electorales. En estos periodos, las sociedades y las instituciones que surgen de ellas, el binomio Estado-Sociedad, experimentan una exacerbación de las diferencias representadas por los distintos candidatos y candidatas. La puesta en duda de sus resultados implica el debilitamiento y condicionamiento del nuevo liderazgo, pero también —y eso es lo más grave— del propio marco normativo-estructural. No se trata tanto de al candidato que se considera más favorable, que también, sino, ante todo, de cuestionar el marco, hacer dudar a la comunidad, resaltar las contradicciones e insuficiencias del sistema que la sostiene, ponerlo en tela de juicio, someterlo a presión y hacer dudar de las propias referencias.

Desinformación y radicalización violenta

En este escenario cada vez son más las voces que desde distintos ámbitos vinculan la desinformación con la radicalización violenta. El término «radicalización» ha suscitado importantes controversias en el mundo académico en los últimos años, en la medida en que su sentido original se ha visto alterado al vincular su análisis con la acción terrorista. La radicalización definida como “el proceso social y psicológico por el que se alcanza un compromiso cada vez mayor con la ideología política o religiosa extremista” (Horgan y Braddock, 2010, p. 152) no tiene por qué ser violenta ni conducir a acción alguna; sin embargo, en la actualidad el concepto se utiliza para definir “el conjunto de procesos que provocan un cambio de actitud que conduce al uso de la violencia” (Neumann y Rogers, 2011, p. 6), de ahí que hablemos de radicalización violenta (para una revisión de la bibliografía, ver Winter et al., 2021).

³ Intoxicación por información.

⁴ Epidemia de mala información.

En ese marco, los expertos que luchan contra el terrorismo han identificado la desinformación como un elemento clave en el proceso que conduce a la radicalización, estimando que la desinformación facilita la radicalización y la radicalización incrementa la vulnerabilidad a la desinformación (Rivas, 2021). Así se plantea en un informe de la Red de Sensibilización frente a la Radicalización (RAN), creada por la Comisión Europea como un recurso de apoyo a los Estados en su lucha contra la radicalización de individuos y grupos. También el *Plan Estratégico Nacional de Prevención y Lucha Contra la Radicalización Violenta* (PENCRV), coordinado por el Centro de Inteligencia contra el Terrorismo y el Crimen Organizado (CITCO), establece en una de sus esferas —Comunicación Estratégica y Redes Sociales contra la Propaganda Radical Violenta— la necesidad de adoptar medidas dirigidas a la lucha contra la desinformación.

Desde la perspectiva de las instituciones preocupadas por la seguridad nacional, resulta cada vez más necesario entender las dinámicas psicosociales y comunicativas para responder a nuevos y emergentes procesos de radicalización que en algunos casos, pueden ser explotados en el marco de las amenazas híbridas. En este ámbito, la *amenaza híbrida* se entiende como resultado de una combinación de ataques encubiertos que una potencia extranjera puede desplegar contra un país soberano sin tener que declarar la guerra. La Estrategia de Seguridad Nacional de 2021 (ESN21) identifica las campañas de desinformación como un riesgo y una amenaza para la seguridad nacional, al entrar a formar parte de estrategias híbridas caracterizadas por actuaciones situadas al margen del principio de buena fe entre Estados que, pese a alterar notablemente la paz, no cruzan los umbrales que justificarían una respuesta armada.

En la misma línea, el Reglamento (UE) 2021/784 del Parlamento Europeo sobre la lucha contra la difusión de contenidos terroristas en línea (29 de abril de 2021), ha delimitado con claridad qué puede ser considerado un contenido perseguible propaganda terrorista y qué no. En opinión de los expertos en la lucha antiterrorista, los grupos radicales acostumbrados a utilizar estrategias clásicas de propaganda se vuelcan ahora en la utilización de campañas de desinformación que les permiten sortear esas limitaciones legales y alcanzar sus objetivos en el terreno de la radicalización (Rivas, 2021). La ventaja que la desinformación ofrece a estos grupos, frente a la propaganda que defiende abiertamente la violencia, radica en el hecho de que las plataformas establecen normas claras para retirar los mensajes ciertamente atribuibles a grupos violentos, pero tienen mucha más dificultad para retirar contenidos que no están indudablemente relacionados con dichos grupos pese a ser evidente que esos mensajes contribuyen a generar la desconfianza hacia los medios de comunicación y las instituciones gubernamentales. Sin embargo, esta relación directa entre desinformación y radicalización es más una hipótesis de trabajo derivada de la observación de los operativos que luchan contra el terrorismo que un resultado contrastado con datos empíricos.

Comprender por qué algunos individuos están dispuestos a sacrificarse por un grupo, por sus valores o por un líder es clave a la hora de estudiar la radicalización violenta. Sin embargo, la investigación empírica sobre los mecanismos en que se apoya este proceso

es muy compleja (Gómez et al., 2023b). Entre las principales barreras identificadas por los expertos en este campo podemos destacar dos: las dificultades para acceder a muestras de individuos en riesgo de radicalización y a miembros de grupos radicales violentos, y la tendencia a sobrestimar la similitud entre diferentes grupos de esta naturaleza.

Existen distintos modelos teóricos que abordan el proceso de radicalización. Uno de ellos es el modelo 3N (Kruglanski et al., 2019), que integra factores relacionados con la necesidad de significación personal, las redes sociales en las que están inmersas las personas, y las narrativas que hacen circular esas redes sociales. La desinformación como estrategia discursiva encajaría en el rol que las narrativas tienen en este modelo teórico. En parte, sobre todo en la dimensión relativa a las redes, el modelo de las 3N está basado en otra propuesta teórica, avalada por la investigación empírica, que aborda el proceso de radicalización: la teoría de la fusión de identidad (Gómez et al., 2020; Swann et al. 2012). Sus resultados son altamente relevantes para entender la realización de conductas extremas relacionadas con la violencia.

La fusión de identidad se desarrolló originalmente como un sentimiento de conexión visceral con un grupo capaz de predecir la disposición a realizar sacrificios extremos por dicho grupo (Gómez et al., 2020; Swann et al., 2012). El modelo de la fusión de identidad plantea que la dedicación y las acciones de violencia extrema perpetradas por los terroristas en favor del grupo y/o en defensa de sus convicciones indican una alineación personal con los objetivos del grupo —para los individuos altamente fusionados, el grupo y ellos mismos, constituyen una sola entidad— que promueve un comportamiento extremo, muy por encima de las preocupaciones por la seguridad y el bienestar individual. Entre los principales factores que explican por qué los individuos fuertemente fusionados están dispuestos a comportamientos extremos por el grupo se encuentran el sentimiento de agencia personal, la percepción de invulnerabilidad, los lazos familiares y el compromiso emocional (Gómez et al., 2020). Y algunos de los mecanismos más relevantes que amplifican los efectos de la fusión de la identidad son los siguientes: (i) cualquier tipo de amenaza al grupo o al propio individuo; (ii) la activación fisiológica, como cuando se les apremia a que tomen una decisión rápida en una situación hipotética de vida o muerte de algún miembro de su grupo; (iii) la activación o saliencia de características relevantes compartidas con los miembros del grupo, ya sean biológicas (p. ej., genes) o sociales (p. ej., valores importantes). Lo que puede llevar, pues, al individuo fusionado a inmolarse por el grupo es un compromiso emocional, no solamente el reconocimiento de un imperativo moral. En otras palabras, un complejo conjunto de factores contribuye a fenómenos de radicalización violenta que se pueden observar en casos específicos.

Un reciente metaanálisis —un tipo de investigación que consiste en un método de análisis sistemático para sintetizar los resultados de numerosos y diferentes estudios empíricos sobre el efecto de una variable independiente (aquí, la fusión) en una serie de variables dependientes (la disposición a realizar sacrificios costosos por el objeto de la fusión, es decir, el grupo con el que se está fusionado)—, confirma la relación positiva entre la fusión de identidad y las orientaciones grupales extremas, destacando que dichos efectos

disminuyen cuando aumenta la edad media de los individuos (Varmann et al. 2023). Según esta teoría, las fronteras entre el «yo personal» y el «yo social» son permeables; ambas identidades se influyen mutuamente y llevan a las personas a hacer por el grupo lo que harían por ellas mismas, hasta el punto de que, tal y como muestran los datos procedentes de la investigación, la fusión de la identidad predice la disposición a luchar y a morir por el grupo (p.e. Gómez et al., 2011; Swann et al., 2012).

En relación con este modelo, la teoría sostiene que algunas personas tienen una conexión visceral con un valor —la religión, el honor, la democracia o la libertad, por poner algunos ejemplos— que consideran sagrado, y que predice la disposición a hacer sacrificios costosos en su defensa (Gómez et al. 2016, 2022, 2023a). Un modelo teórico basado en la teoría de la fusión es el de los *actores devotos* (Gómez et al., 2017, Vázquez et al., 2020), el cual postula que quienes están fusionados con un grupo cuyos miembros comparten un valor que consideran sagrado, y al que no renunciarían por ningún tipo de intercambio material o inmaterial, estarán dispuestos a hacer sacrificios excepcionalmente costosos y extremos por sus creencias y/o por su grupo.

LA DESINFORMACIÓN Y LAS DINÁMICAS GRUPALES

Las dinámicas grupales juegan también un importante papel en la circulación de la desinformación en contextos donde los individuos no están radicalizados, es decir, en lo que consideraríamos las dinámicas cotidianas de la conformación de la opinión pública. Para ello es necesario aproximarse a la difusión de contenidos considerando las relaciones sociales que estos mensajes crean o fortalecen. Para abordar ese espacio de conexión entre las dinámicas grupales y la difusión de información es útil la distinción teórica entre los géneros de «difusión», «propagación» y «propaganda» propuesta por Serge Moscovici a mediados de los años 70 (Moscovici, 1976). Dicha distinción nos permite identificar con mayor claridad las dinámicas sociales que subyacen a distintas estrategias de difusión de contenidos.

En el primero de los géneros identificado por Moscovici, la «difusión», el contenido de la información se plantea con el objetivo de llegar al máximo número de personas. Para ello la información ha de circular por encima de las categorías sociales y del conflicto intergrupar. Su objetivo es crear opinión entre las mayorías. El segundo género, la «propagación», tiene como propósito cohesionar al propio grupo. Su contenido se supedita a la existencia de un conflicto intergrupar, y persigue validar el pensamiento de grupo y evitar las divisiones internas. Finalmente, el tercer género, la «propaganda», pretende promover el conflicto intergrupar maximizando las diferencias entre grupos y empujando a la acción contra el exogrupo.

Como puede observarse, la propagación es un fenómeno que se ancla en procesos de comunicación intragrupo, y está dirigido a garantizar la cohesión de sus integrantes en torno a un mensaje. Por su parte, la propaganda es un fenómeno que maximiza el conflicto intergrupar en un marco de relaciones sociales excluyentes, con el objetivo de poner en circulación mensajes que faciliten o promuevan esa confrontación. Estas diferentes estrategias de comunicación son utilizadas por grupos sociales e instituciones de características muy distintas. Las huellas de las relaciones sociales que un mensaje pretende promover pueden ser trazadas tanto en los discursos mayoritarios como en los minoritarios.

Que los contenidos de la desinformación sean sostenidos por grupos minoritarios no debe conducir a menospreciar su capacidad de influencia. Desde la perspectiva interaccionista sabemos que todo individuo o grupo (sea mayoritario o minoritario) es una fuente potencial de influencia. Cuando el comportamiento de una minoría crea un conflicto socio-cognitivo y es capaz de mantenerlo con estilo consistente, sincrónico o diacrónico, puede llegar a conseguir la difusión de sus posiciones innovadoras y participar activamente en el cambio social (Pérez y Mugny, 1985; Moscovici et al., 1991). Dicho cambio puede ser positivo para la sociedad en la medida que integra nuevas propuestas que permiten avanzar en derechos

o libertades, pero también puede ser lesivo para la colectividad cuando quienes lo fomentan son agentes que actúan contra los derechos humanos y minan las bases de la convivencia democrática, o bien agentes externos que operan en el interés de otras naciones.

Para crear ese conflicto sociocognitivo capaz de captar la atención de la audiencia, muchas veces se aprovechan *fisuras histórico-sociales* (en inglés “cleavage”) basadas en diferencias relacionadas con las lenguas, las etnias, las creencias religiosas o la cultura, que han servido y sirven de excusa para justificar un conflicto social presentado como un choque entre diferentes visiones del mundo. Estas fisuras pueden derivar de la (re)interpretación de acontecimientos de la historia reciente que siguen ocupando un espacio en la memoria colectiva.

Continuidad y cambio en el entorno digital

Esta capacidad de influencia de las minorías mediante la articulación del conflicto no es un fenómeno nuevo; ha existido mucho antes de nuestra era digital. Sin embargo, el advenimiento de las actuales tecnologías de comunicación ha provocado cambios en el modo en que se produce, consume y distribuye la información (Bakir y McStay, 2018). En esta era digital los ciudadanos no son meras víctimas de la información manipulada, la selección motivada de noticias, la omisión de información importante o la distorsión de la verdad; también son actores importantes en su creación y difusión a una velocidad y a una escala sin precedentes históricos. De ahí que Bruns (2005) haya acuñado el concepto de *producers*, resaltando el papel activo que desempeñan los usuarios generadores de contenidos.

Esta nueva capacidad de actuación de los ciudadanos como creadores, consumidores y difusores de información mediante un uso masivo de las redes sociales ha producido una amplificación de los fenómenos de polarización afectiva (Törnberg, 2022). Es decir, las ya mencionadas dinámicas clásicas de la propagación y la propaganda (Moscovici, 1976) encuentran nuevas y variadas posibilidades en las sociedades digitales que no solo amplifican el funcionamiento clásico del prejuicio y otros sesgos sociocognitivos, sino que crean un fenómeno de naturaleza nueva (Rosso y Chulvi Ferriols, 2021). Una de las principales novedades de estas sociedades digitales es que existen muchas más posibilidades de marcar socialmente los contenidos. En la actualidad, cualquiera tiene posibilidad de crear un mensaje de recepción masiva, pero eso no supone que se interactúe en el espacio digital de forma individual; esa multiplicidad de emisores se adhiere a comunidades ideológicas y a grupos sociales, precisamente para hacer inteligible su mensaje.

Otra de las características de estas sociedades digitales es la utilización intensiva de una *comunicación anónima o basada en seudónimos*: el ecosistema comunicacional en el que vivíamos con anterioridad al establecimiento de las sociedades digitales no permitía el anonimato para llegar a públicos masivos, tal y como es posible hoy a través de las redes sociales.

Una tercera característica, también relevante para la comprensión de las dinámicas grupales que alimentan la desinformación, es la desaparición de las fronteras físicas. Al permitir la interacción de las personas sin proximidad física, las redes sociales facilitan la conexión de individuos radicales que de otra manera verían muy limitada su posibilidad de conocimiento mutuo. Los expertos señalan que uno de los rasgos definitorios del actual *renacimiento del extremismo político* en el mundo consiste en despojar de la categoría de marginalidad a un grupo y normalizar su discurso (Mudde, 2021), dos aspectos que las redes sociales facilitan y promueven.

Una vez desaparecidos los límites geográficos como condicionantes de la interacción humana, deja de importar que el grupo de pertenencia que legitimará el odio contra el diferente se encuentre en el mismo barrio o disperso por todo el planeta (Kaufman, 2021). La normalización se produce también porque personas que no estarían dispuestas a recibir en sus casas a grupos extremistas entran en contacto por casualidad o por curiosidad con *expresiones verbales o visuales de odio* en las redes sociales, sin que ello tenga un alto coste para su identidad o su espacio de convivencia. Finalmente, el fenómeno de la *repetición ad infinitum* de los mensajes en las redes sociales provoca una *falsa sensación de consenso* que permite que posiciones minoritarias adquieran el poder de organizar el debate social en torno a determinados temas.

La desinformación también se nutre de “contenidos sensacionalistas o provocativos” que son amplificadas por distintos mecanismos característicos de los medios digitales. Algunos de ellos son la personalización de los contenidos a través de filtros y sistemas de recomendación, la tendencia de los usuarios a leer solamente los primeros resultados que aparecen al usar un buscador como Google, o el uso de titulares engañosos (*clickbait* en inglés), típicos de la prensa amarilla, para aumentar el tráfico hacia una dirección web. La desinformación también se beneficia de muchas otras características de los servicios digitales y del diseño de interfaz, que intentan prolongar indefinidamente el tiempo del usuario en la plataforma digital e incentivar su interacción en número de visitas, número de clicks, número de visualizaciones, etc.

Algunos temen que la comunicación personalizada pueda dar lugar a burbujas de filtrado o «filter bubble» (Pariser, 2011). Por ejemplo, un canal informativo podría considerar conveniente personalizar los contenidos que se muestran a cada usuario de acuerdo con sus preferencias, lo cual generaría una mayor exposición a los medios conservadores o liberales en función de los (supuestos) intereses políticos del usuario. Como resultado, los usuarios quedarían expuestos solo a una gama limitada de ideas políticas que convalidaría sus posiciones ideológicas.

De momento, los estudios académicos no han encontrado evidencias que soporten esta tesis (Zuiderveen Borgesius et al., 2016). La exposición selectiva a la información no produce siempre los mismos resultados. Por ejemplo, en un estudio con 408 participantes residentes en España en abril de 2015, se observa que Facebook aumenta la exposición selectiva entre los que se declaran simpatizantes de izquierdas y la disminuye entre los que se declaran de derechas; al mismo tiempo, Google la disminuye entre los de izquierdas y la aumenta entre los de derechas (Cardenal et al., 2019).

La construcción y descalificación del adversario

Los sesgos informativos de naturaleza ideológica, junto a circunstancias de otro orden, se encuentran en el origen del conflicto y de la confrontación entre personas y grupos. Desde los años 60, la teoría del grupo mínimo (Tajfel, 1974) muestra que el agrupamiento y clasificación de personas en torno a un rasgo arbitrario e insignificante (categorización social) que tenga sentido para el individuo se traduce en dos efectos inmediatos: aceptación incondicional de todo lo que hagan los nuestros, y distanciamiento y rechazo de lo que hagan los otros; algo que, en su manifestación más extrema, podría conducir a un conflicto intergrupalo. La categorización “ayuda a estructurar la comprensión causal del ambiente social y por tanto sirve de guía para la acción” (Tajfel 1984, p. 61). Varios estudios sobre el prejuicio hacia las minorías muestran que la discriminación intergrupalo empieza por la adopción de un criterio de clasificación como, por ejemplo, la dimensión natura-cultura que sitúa a determinados grupos fuera del mapa social (Pérez et al., 2002). Una mera clasificación, sin amenazas o discriminación negativa, sirve de base para una clasificación social a partir de la cual se potencia el desarrollo de comportamientos hostiles o amigables hacia el exogrupo.

En la relación entre endogrupo y exogrupo juega un papel fundamental la identidad social de la persona. Las teorías de la identidad social nos ayudan a entender mecanismos de contagio de las opiniones y discriminación intergrupalo en episodios de polarización. La idea central de la Teoría de la Identidad Social es que un individuo tenderá a permanecer como miembro de un grupo, o a buscar la pertenencia a nuevos grupos, si estos tienen alguna contribución que hacer a los rasgos positivos de su identidad social, es decir, a aquellos aspectos de los que obtiene alguna satisfacción (Tajfel, 1984). Aunque nuestra identidad social podría alimentarse de múltiples pertenencias —como ya hemos visto al hablar de la radicalización—, la ilusión y la tiranía de una identidad única (Gómez et al., 2016) que niega la pluralidad de nuestras filiaciones al tiempo que alimenta de manera falaz la ilusión de un destino colectivo al que es necesario sacrificar los objetivos personales, pueden estar abonando la radicalización a partir de lo que algunos autores han descrito como identidades asesinas (Maalouf, 2012; Sen, 2007): “de hecho, muchos de los conflictos y las atrocidades se sostienen en la ilusión de una identidad única que no permite elección. El arte de crear odio se manifiesta invocando el poder mágico de una identidad supuestamente predominante que sofoca toda otra filiación y que, en forma convenientemente belicosa, también puede dominar toda compasión humana o bondad natural que, por lo general, podamos tener. El resultado puede ser una rudimentaria violencia a nivel local o una violencia y un terrorismo globalmente arteros” (Sen, 2007, p. 15).

Los procesos conducentes a la radicalización inician su camino en la diferenciación intergrupalo y acaban desembocando en la deshumanización, es decir, en negar a determinados grupos y a las personas pertenecientes a ellos las cualidades que nos definen como humanos (Kelman y Hamilton, 1989; Haslam, 2006). Deshumanizamos o infrahumanizamos a otras personas “negándoles la posibilidad de experimentar sentimientos, es decir, privándoles de una de las potencialidades básicas del ser humano” (Rodríguez Torres et al., 2003, p. 77). En este proceso juega un papel decisivo el debilitamiento de las reglas, normas,

estándares y obligaciones morales que nos permiten actuar de manera hostil y hasta cruel respecto a determinadas personas a través de un proceso de desconexión moral (Bandura, 1999). Eso fue lo que ocurrió bajo el efecto de desindividuación (anonimato, difusión de la responsabilidad, liberación de restricciones sociales y morales) en el conocido “Experimento de la Prisión de Stanford” (Zimbardo et al., 1986). A quienes se les asignó de manera aleatoria el papel de “prisioneros” les fue arrebatada su intimidad y su identidad, y ello dio pie a que los “guardianes” actuaran liberados de toda convención moral, de todo control y de cualquier mecanismo de auto-regulación: “cuando se considera que ciertas personas o grupos están fuera de la esfera de la humanidad, los agentes deshumanizadores suspenden la moralidad que normalmente regiría sus actos hacia sus congéneres” (Zimbardo, 2007, p. 404).

Bar-Tal (1996) identifica tres estrategias básicas de *descalificación y deshumanización del adversario*: (i) su proscripción al calificarlo como violador de las normas sociales; (ii) la caracterización de rasgos de personalidad, u otras particularidades, que provocan rechazo social; (iii) el uso de rótulos políticos, es decir, comparaciones con otros grupos sociales o políticos que son negativamente evaluados por la ciudadanía. Estas descripciones demoledoras del otro suelen estar acompañadas por discursos de legitimación de la violencia que recurren a una caracterización del adversario basada en rasgos y características socialmente condenables.

Un ejemplo claro de estas estrategias se encuentra en los discursos sobre «Occidente» de los grupos terroristas yihadistas Al Qaeda y del llamado Estado Islámico que entre 2010 y 2015, en sus respectivas revistas de propaganda *Inspire* y *Dabiq*, hablan de Occidente como un «otro» extraño y aberrante (Lorenzo-Dus y Macdonald, 2018). Estrategias similares se encuentran en el estudio de Sabucedo et al. (2004) que analiza los discursos de las Fuerzas Armadas Revolucionarias de Colombia –Ejército del Pueblo (FARC-EP— y de las Autodefensas Unidas de Colombia (AUC). Tanto la FARC-EP como la AUC proscriben al adversario señalándolo como un enemigo público; sin embargo, hay diferencias en la categoría «uso de rótulos políticos». Las FARC describen a las AUC como mercenarios, paramilitares y sicarios (y también fascistas, imperialistas, latifundistas, oligarcas y terratenientes), mientras que las AUC describen a las FARC como bandidos, guerrilla y subversivos (y también izquierdistas, mesiánicos y totalitarios).

Cualquier circunstancia, real o imaginaria, que incremente la percepción de competición intergrupal aumentará también la percepción de amenaza exogrupal, independientemente de que la escasez de recursos sea real o simplemente percibida. La amenaza realista implica la percepción de competición entre endogrupo y exogrupo por recursos escasos, como la sanidad o el trabajo. La amenaza simbólica, por su parte, se produce porque los miembros del endogrupo perciben que el exogrupo posee diferentes valores, creencias o actitudes que no están en consonancia con los propios y, por tanto, éstos corren el peligro de ser alterados por los del exogrupo. Es aquí donde cabe recordar, con la ayuda de Daniel Bar-Tal, la existencia de creencias grupales como “convicciones que los miembros de un grupo son conscientes de compartir y a las que consideran definitorias de su pertenencia grupal”

(Bar-Tal 1996, p. 256). Esas creencias a) “sirven como base para la existencia del grupo” (p. 279) y para su cohesión; b) son la base sobre las que se erigen las creencias personales; c) definen las condiciones de pertenencia; d) establecen el marco de comparación respecto a otros grupos; e) satisfacen la necesidad de claridad cognitiva, y f) marcan la ruta a seguir hacia el interior del grupo y, de manera particular, en relación con los otros grupos.

Cabe esperar, por tanto, que la exageración de las diferencias (o falta de similitudes) percibidas entre el endogrupo y el exogrupo constituyan un elemento fundamental en la percepción de amenaza exogrupal y, por tanto, en la actitud negativa hacia el otro grupo social. En esta línea de razonamiento, tendemos a pensar que es el sentimiento de amenaza el que precede al sentimiento de discriminación y lo explica. Sin embargo, investigaciones recientes han mostrado que es la discriminación basada en la diferenciación categorial la que precede al sentimiento de amenaza (Pérez et al., 2023). Es decir, ser el autor de una acción discriminatoria hacia la minoría hace que, posteriormente, esa minoría sea percibida como amenazadora. Para medir la percepción de amenaza exogrupal (EPAE) integral percibida hay disponibles escalas adaptadas al contexto español (Soledad Navas et al., 2012).

La exageración de las diferencias intergrupales va asociada a un fuerte proceso de expresión emocional que se define como *polarización afectiva*. Dicha *polarización* se expresa en reacciones de adhesión/afecto hacia quienes comparten nuestras ideas, y de rechazo/antipatía hacia quienes defienden ideas opuestas a las nuestras. Por ejemplo, hay estudios que muestran que las preferencias secesionistas condicionan altos niveles de polarización afectiva, en el sentido de que los partidarios y los contrarios a la independencia tienen fuertes opiniones negativas unos de otros (Balcells y Kuo, 2022). A esta exageración de las diferencias puede contribuir un discurso político cuyas ganancias electorales reposen en la contraposición de los intereses grupales como vía para movilizar al electorado. Los discursos polarizantes encuentran un eco muy especial en situaciones dominadas por la percepción de amenaza y el miedo, en las que la necesidad de claridad cognitiva nos anima a buscar información capaz de dar respuestas, no importa de qué tipo, a la situación de incertidumbre (Blanco, 2022). El miedo es una emoción básica en cuyo entorno se desata una suerte de tormenta psicológica donde se dan cita el estrés, la ansiedad y la indefensión. En la ruta mental del miedo al otro están implicados los mismos procesos que en el conflicto, la fusión de la identidad o la radicalización, a saber, la categorización, la diferenciación categorial, la comparación social y la identidad (Blanco, 2022, p. 11-17); todos ellos se juegan en un espacio multigrupal y multicategorial en el que la protección y la seguridad que encontramos en los nuestros se corresponde con el peligro y la amenaza, imaginaria a veces, proveniente de los otros. La paz, la fraternidad y la lealtad forman parte de las relaciones con los nuestros; la guerra, el odio y el desprecio pertenecen a las relaciones con los otros (Sumner, 1906, p. 13).

Pero no solo es el miedo el que entra en juego en todo este proceso. Ignacio Martín-Baró, en sus estudios sobre la guerra que asoló El Salvador en la década de los ochenta —y de la acabó siendo una de sus víctimas—, plantea una estrecha correspondencia entre

polarización y violencia a través de la construcción de la imagen del enemigo y de la mentira; ambas, dice, se encuentran en el origen de la guerra civil (Martín-Baró, 2003). La mentira deforma la realidad, convierte a los perpetradores en víctimas, inventa hechos inexistentes, ocasiona una penumbra psicosocial donde resulta difícil distinguir lo real de lo ficticio, reduce la compleja realidad social a esquemas e imágenes rígidas y simplistas cargadas de emotividad “donde los fantasmas terminan imponiendo su ley al conocimiento, hasta el punto de que algunas personas y grupos llegan a creerse las mentiras que ellos mismos han fabricado” (p. 318). Con todo, “lo más grave es su tergiversación moral. No se trata solo de que se falsee la forma como actúan las personas o los grupos; se trata, sobre todo, de su denigración” (Martín-Baró, 2003, p. 214).

La presión por la conformidad en el seno del grupo

Las dinámicas de propagación y propaganda pueden ser especialmente efectivas en sociedades polarizadas donde las personas son menos propensas a considerar diferentes puntos de vista. Las personas con fuertes identidades partidistas suelen ser más proclives al sesgo de confirmación (sobre ello se hablará en un epígrafe posterior) y al escepticismo motivado (p. ej., Flynn et al., 2017; Taber y Lodge, 2006), lo que significa que buscarán y aceptarán información que respalde sus creencias partidistas y rechazarán información que las contradiga. Esta tendencia puede estar relacionada con el deseo de mantener una autoimagen positiva y la cohesión del grupo (Tajfel y Turner, 1986).

Entre los factores que contribuyen a que los grupos radicalizados mantengan una alta unión interna a fin de prevenir la mínima desviación del credo central del grupo, cabe mencionar la inexistencia de interacción social con entornos de ideología diferente y la creación de un enemigo o una amenaza externa. Figura, además, la llamada influencia normativa dentro del grupo, que induce al individuo a comportarse de acuerdo a lo que imagina que se espera de él, lo que viene a ser una de las estrategias más eficaces de presión. Este mecanismo es altamente frecuente en los entornos radicalizados en los que se da una permanente preocupación por no desviarse del corpus ideológico del grupo, llegando incluso a mostrar comportamientos del tipo *primus inter pares* (“ser más papista que el papa”, vid., Codol, 1975). Digamos que el individuo siente una alta ansiedad por saber qué tiene que opinar para ser aceptado como un miembro ferviente y leal al grupo. El miedo al rechazo está presente en los procesos que conducen a la conformidad con la mayoría en el seno de los grupos (Asch, 1969). Esta influencia reposa en una distribución real o simbólica de premios (“likes”, comentarios positivos, reenvío de sus mensajes) y castigos (“dislikes”, críticas, cancelaciones) según que el individuo se comporte o no como agrada al resto. En grupos más cerrados se da tal apoyo social a sus miembros que les hace sentirse especiales y valorados, siempre que su opiniones y comportamientos se ajusten a las expectativas normativas comunes.

Estas dinámicas intragrupo que acabamos de describir están cada vez más presentes en las investigaciones sobre la difusión de la desinformación. Recientemente, un interesante trabajo en las ciencias sociales de la computación (Phadke et al., 2021) ha encontrado evidencia empírica que apoya la importancia de los factores sociales en el compromiso con las teorías de la conspiración. Los resultados de un amplio estudio de las comunidades conspiracionistas de Reddit apoyan la idea de que la cantidad de interacciones diádicas con usuarios conspiracionistas es el elemento más importante para predecir si los usuarios se unirán a estas comunidades, incluso después de controlar factores individuales como la ira, la tristeza, la ansiedad, así como la tendencia a adherirse a fuentes de información aisladas y sesgadas. Por otra parte, la aparición de comunidades conspiracionistas y su creciente participación en los procesos políticos, incluida la violencia, apuntan a que el clásico estereotipo del “teórico de la conspiración” solitario que apoya creencias irracionales debe ser abandonado. En este sentido, Wagner-Egger et al. (2022) proponen abordar el problema de la adhesión a las teorías de la conspiración desde la perspectiva de las dinámicas intragrupo: “si su principal objetivo es mostrar lealtad a una visión del mundo o a una comunidad y reforzar y coordinar la acción, es evidente que deben considerarse algo más que meras creencias” (p. 3).

Este proceso en el que el individuo queda atrapado en una comunidad conspiranoica ha sido descrito recientemente como “síndrome de la madriguera del conejo” (Sutton y Douglas, 2022). Dicho síndrome sugiere que las personas se adhieren a creencias conspirativas casi de forma inadvertida, pero el proceso se acelera de forma recursiva y, en la fase final, resulta muy difícil salir de él. En la misma línea, un estudio cualitativo sobre creyentes en conspiraciones propone una tipología en la que una cosmovisión conspirativa es el final de un “viaje espiritual” progresivo, donde se empieza por cuestionar las ortodoxias sociales y políticas y se acaba desarrollando un fuerte sentimiento de pertenencia a una comunidad (Bradley et al., 2017). En conclusión, las dinámicas de conformidad en el seno del grupo necesariamente juegan un papel clave, desde el momento en que la adhesión a una mentalidad conspirativa se considera como un proceso y no solo como un estado de compromiso intenso con una narrativa alternativa.

El concepto de verdad y las teorías de la conspiración

Una articulación reciente de cómo funcionan en el discurso de desinformación las dinámicas intergrupales ya descritas es la realizada por Blitvich y Lorenzo-Dus (2022). Estas autoras extraen tres elementos fundamentales en la construcción de una Teoría de la Conspiración (TC): (i) la presencia de conocimientos factuales y conceptuales; (ii) conocimientos procedimentales pseudo-científicos, y (iii) conocimientos metacognitivos asociados con identidades grupales. Cabe destacar el recurso al concepto de verdad en los tres tipos de conocimientos. La categoría de conocimiento factual-conceptual hace constantes referencias a la verdad en oposición a las versiones oficiales (por defecto falsas) de los hechos. Para sonar veraces, las TC se exponen en términos racionales, no emocionales, y se

describen recurriendo a un vocabulario pseudo-científico. Este conocimiento procedimental crea un conjunto de herramientas y un marco analítico *sui generis* en el que se da prioridad a la generación de ideas (formación de hipótesis y revisión de la bibliografía) en lugar de enfatizar el análisis de datos empíricos (confirmación de hipótesis). En cuanto a las categorías metacognitivas implicadas en la formación de identidades, la verdad es también el elemento principal: los conspiracionistas buscan la verdad a pesar de la resistencia de agentes (tanto humanos como no humanos) cuyo principal objetivo es ocultar dicha verdad en contra del interés del público.

Aunque parezca contraintuitivo, recurrir al concepto de verdad a la hora de desacreditar conspiracionistas puede ser totalmente contraproducente. Al proponerse como activistas de la verdad, los conspiracionistas segregan la verdad dentro de una dinámica de contraposición entre grupos. El discurso polarizado que suele presentarse como un «nosotros contra ellos» funciona según lo que van Dijk (1991) denomina un «cuadrado ideológico». Éste comprende cuatro estrategias:

- i. Enfatizar las buenas propiedades/acciones del endo-grupo.
- ii. Enfatizar las malas propiedades/acciones del exo-grupo.
- iii. Ocultar las malas propiedades/acciones del endo-grupo.
- iv. Ocultar las buenas propiedades/acciones del exo-grupo.

Un grupo puede favorecer las estrategias ii y iv hasta tal punto que su discurso se vuelve abiertamente discriminatorio hacia uno o más exogrupos. En este caso, se trata más bien de una «alteridad» (en inglés *othering*) que de una mera exclusión. Este proceso de *othering* respecto a personas y grupos significa representarlos discursivamente no solo como radicalmente diferentes del propio grupo, sino también, y de forma crucial, como «excéntricos y distantes» y como «extraños incompatibles» (Baumann 1991, 66). El término *othering* se ha examinado ampliamente en las ciencias sociales, sobre todo en el marco de las relaciones entre Occidente y Oriente (véase, por ejemplo, Said, 1997, 2003). Dentro del análisis del discurso, y adaptando taxonomías de alteridad lingüística desarrolladas por Coupland (2010) a contextos de radicalización en línea, Lorenzo-Dus (2023) identifica cinco estrategias comunicativas de alterización que se solapan parcialmente:

- i. Homogeneización: uso de un discurso que niega a las personas su individualidad, recurriendo a menudo a estereotipos sociales.
- ii. Represión y silenciamiento: referirse a un conjunto limitado de características del otro grupo que se adaptan a las agendas y prioridades del propio grupo.
- iii. Representación cero: convertir en invisibles a determinados grupos sociales, haciéndolos desaparecer del discurso.

- iv. Empeoramiento: uso de etiquetas y atributos para mostrar a otros individuos y grupos bajo un prisma muy negativo.
- v. Exhibición de liberalismo: afirmación discursiva de que se defienden orientaciones “liberales” hacia los grupos marginados. Un ejemplo clásico es una representación racista acompañada de renunciaciones personales al estilo “No soy racista, pero...” (véase van Dijk, 1993).
- vi. Subversión de la tolerancia: “trabajo discursivo [que] muestra que el liberalismo es demasiado idealista o ingenuo o aburrido o anticuado” (Coupland 2010, p. 253). El humor es una estrategia muy utilizada para subvertir la tolerancia en el discurso.

Los grupos extremistas invierten mucho tiempo y esfuerzo en construir comunidades o “espacios de afinidad” digitales (Gee, 2005), a través de los cuales poder consolidar su identidad grupal. Furlow et al. (2014) argumentan que los esfuerzos del llamado Estado Islámico por restaurar el Califato en 2014 se basaban en la construcción de una comunidad política imaginaria que ofrecía cohesión social a sus miembros en torno a su «musulmanidad», independientemente de que se conocieran entre sí o no, y también de su nacionalidad y etnia. Según los autores, esta construcción vaga y genérica se basaba en estructuras retóricas del tipo “nosotros contra ellos” (Van Dijk 2016), que permitían a públicos diversos identificarse con el Califato y facilitaban el énfasis en sus diferencias irreconciliables con el grupo exterior. Nouri y Lorenzo-Dus (2019) también consideraron que el modelo de Anderson (2006) de «comunidades políticas imaginadas» es aplicable a la creación de comunidades digitales por parte de la derecha radical, concretamente los grupos *Britain First* y *Reclaim Australia*. Lorenzo-Dus (2023) muestra que los miembros de estos grupos (ya sean de derecha radical o de ideología yihadista) adoptan un estilo discursivo que se caracteriza por tres posicionamientos:

- i. Una amplia experiencia, lo que permite opinar sobre cualquier circunstancia, idealmente de actualidad, utilizando cualquier tema para presentar favorablemente la ideología del grupo de pertenencia.
- ii. La apertura emocional tóxica, que se basa en la mitopoiesis, concretamente en la narración en primera persona que resalta emociones estereotipadamente masculinas como la ira frente la injusticia, la victimización percibida y una visión favorable a adoptar una lógica de represalia.
- iii. Consecuentemente, surge la avidez impaciente, es decir, una llamada apremiante a la acción, que puede llevar a comportamientos violentos o antisociales.

Los intentos de contrarrestar la captación ideológica/radicalización en línea tienden a encajar en uno o más de los siguientes enfoques (Henschke y Reed, 2021):

- i. Redireccionamiento: se dirige a los usuarios de internet en los que se observa una búsqueda de material nocivo en línea con contenidos, incluidos anuncios, que promueven alternativas no nocivas y están vinculados a las búsquedas originales.
- ii. Interrupción: busca reducir la oferta del contenido eliminándolo y/u ocultándolo para impedir el acceso al mismo, consciente de que esta estrategia puede producir la migración de estos contenidos y de los grupos que los promueven a otras plataformas o redes sociales menos accesibles.
- iii. Diálogo cara a cara: detectar a los productores de contenidos nocivos en línea y ponerse en contacto con ellos ofreciéndoles entablar un diálogo en línea.
- iv. Educación y contenidos alternativos: tratar de reducir la demanda de contenidos nocivos en línea difundiendo mensajes que contrarresten el impacto de dichos contenidos.

VULNERABILIDAD A LA DESINFORMACIÓN Y LA PROPENSIÓN A CREERLA

Como ya se ha mencionado, la desinformación no es un fenómeno nuevo; ha existido mucho antes de nuestra era digital. Sin embargo, la amplitud de estos fenómenos ha llevado a muchos a calificar nuestra época nada menos que como la “era de la posverdad” (p. ej., McIntyre, 2018). Alejándose de los ideales de la Ilustración, ahora la verdad en sí parece en múltiples casos menos importante que promover el propio relato de los hechos y la creación idiosincrásica de realidades. Cualquiera puede decir lo primero que le venga a la mente, mientras que los hechos objetivos o los argumentos de los especialistas no logran contrarrestar la influencia que las emociones y las creencias populares ejercen sobre la opinión pública. Uno de los aspectos más inexplicados de este predominio de la posverdad en el conjunto de las ciencias sociales es que, incluso cuando las personas saben que la información es errónea, todavía la creen y la difunden en su entorno. Así, la “posverdad” no es solo un problema de falta de conocimiento, desinformación o falta de información; parece estar relacionada con el pensamiento social, con una interferencia de los prejuicios sociales en la manera en que se procesa e interpreta la información.

La mayoría de las veces, la manipulación de la información tiene como principal objetivo activar emociones fuertes, como el miedo, la ira, la humillación, la injusticia, la exclusión social, que pueden tener más poder persuasivo que los datos objetivos y los argumentos científicamente fundamentados. En sus trabajos, Moghaddam (p. ej., 2005) muestra que estas emociones desempeñan un papel crucial en el proceso de radicalización, y señala que los mensajes de los grupos extremistas como el Estado Islámico, por ejemplo, no solo buscan atraer personas al contenido de una ideología extremista, sino que también aspiran a satisfacer necesidades emocionales y psicológicas como pueden ser la identidad y el sentimiento de pertenencia grupal. Se recopila información sobre los individuos y se personalizan los mensajes para atraerlos progresivamente de manera más efectiva.

Varios estudios muestran que la información que genera emociones negativas se propaga más rápido y llega a más personas que la información “aséptica” que describe con objetividad los hechos (Vosoughi et al., 2018; Pennycook y Rand, 2018). Además, hay casos en los que la desacreditación explícita incrementa su propagación (p. ej., Lewandowsky et al., 2017). A veces, al intentar negar la credibilidad a una información, lo que se produce es un resurgimiento del interés por ella, especialmente si se trata de un tema controvertido o emocionalmente intenso (p. ej., Moscovici et al., 1985). Las personas tienden a difundir rumores y noticias falsas si creen que hacerlo es socialmente aceptable o que con ello agradarán a su grupo o a su comunidad (Cialdini y Goldstein, 2004).

Dada la magnitud de este fenómeno, no cabe sino preguntarse cómo combatirlo, cómo contrarrestar su impacto en la sociedad. ¿Qué resistencias se pueden movilizar ante tales ataques? Las ciencias sociales y políticas no disponen de respuestas cabales sobre tales cuestiones. No hay una fórmula mágica para contrarrestar esas amenazas. La psicología de la influencia social que se logra con la difusión de información manipulada es muy compleja. En la mayoría de los casos no responde a la lógica más elemental, sino que se caracteriza por tener su propia lógica. A continuación, resumiremos algunos procesos cognitivos y psicosociales involucrados en el procesamiento de la información, sea ésta verídica o no. Posteriormente resumiremos algunas propuestas de resistencia a la influencia social que cuentan con mayor aval empírico y que pueden implementarse en casos concretos. De cara a los objetivos del presente capítulo, y entre una amplia gama de procesos cognitivos y psicosociales involucrados en el procesamiento de la información, nos centraremos en los tres siguientes: el sesgo de confirmación, la exposición selectiva a la información, y el heurístico de disponibilidad. Estos sesgos pueden aumentar la vulnerabilidad de las personas a la desinformación.

El sesgo de confirmación y la exposición selectiva

Los sesgos de confirmación y exposición selectiva se encuentran estrechamente relacionados. Ambos influyen en cómo las personas hacen conjeturas, procesan la información y toman decisiones, quedando afectada así su percepción de la realidad y sus interacciones sociales.

Con el llamado sesgo de confirmación se hace referencia a la tendencia de las personas a interpretar, prestar atención y recordar sobre todo la información que confirme su *wishful thinking*, sus creencias, prejuicios o hipótesis previas, y a ignorar o desacreditar la información que las contradice (p. ej., Wason, 1960; Nickerson, 1998). No se debe tanto a eventuales limitaciones cognitivas del procesamiento de la información cuanto a la más pura naturaleza del pensamiento mágico: es verdad porque deseo que sea verdad.

Complementariamente, la exposición selectiva es el proceso por el cual las personas tienden a exponerse a aquella información que resulte coherente con sus creencias, opiniones y actitudes preexistentes, al tiempo que procuran evitar la que es disonante. Esto puede perpetuar estereotipos y prejuicios, ya que las personas propenden a no prestar la mínima atención a la información que los desafía (Crawford y Pilanski, 2014). Hay una tendencia a procesar información de manera más eficiente cuando esta información es consistente con nuestras creencias (Kunda, 1990). Además, las personas se inclinan a evitar información disonante que les generaría un intenso malestar (Frey, 1986) y prefieren la información que les permita mantener la coherencia cognitiva entre sus creencias, actitudes y conductas (Festinger, 1957). Esta pulsión a la coherencia sociocognitiva puede llevar a las personas a rechazar o desacreditar la información que desafía su identidad partidista y sus principios ideológicos, por ejemplo.

Estos dos procesos (el sesgo de confirmación y la exposición selectiva) pueden tener notables consecuencias a la hora de la toma de decisiones y la formación de opiniones. Por ejemplo, el sesgo de confirmación puede conducir a perpetuar y exacerbar la polarización de opiniones y la formación de las llamadas cámaras de eco (*echo chamber*). Las cámaras de eco son un fenómeno en el que la información, opiniones y creencias se refuerzan y amplifican dentro de un grupo o comunidad cerrada, limitando la exposición a perspectivas diferentes (p. ej., Kathleen Hall y Cappella, 2008). Este fenómeno ha sido ampliamente estudiado en el ámbito de la psicología social, y particularmente en relación con las redes sociales y la polarización política (p. ej., Sunstein, 2001; Pariser, 2011). En un gran estudio, Del Vicario et al. (2016) analizaron la difusión de información en redes sociales y observaron que los usuarios tienden a agruparse en comunidades cerradas donde comparten y promueven información que refuerza sus creencias previas, lo que contribuye al aislamiento ideológico y a la polarización. Flaxman et al. (2016) encontraron que los algoritmos de personalización *online* pueden exacerbar la polarización política al limitar la exposición a puntos de vista diferentes. Sin embargo, algunos estudios han matizado esos efectos de las cámaras de eco o no han logrado confirmar su existencia. Por ejemplo, Barberá et al. (2015) y Dubois y Blank (2018) sugieren que, aunque las cámaras de eco pueden existir en ciertos entornos *online*, su prevalencia y efectos podrían estar siendo exagerados, ya que los usuarios también se exponen a diversas fuentes de información a través de otros medios. Los estudios que no confirman las cámaras de eco tienen un interés particular, ya que nos señalan vías para contrarrestarlas. Por ejemplo, Haidt (2012) subraya que lo ideal es fomentar la exposición a perspectivas diversas y la interacción con personas de diferentes orígenes y opiniones.

El heurístico de disponibilidad

El heurístico de disponibilidad se refiere a la tendencia de las personas a estimar la frecuencia o probabilidad de un evento basándose en la facilidad con la que pueden recordar ejemplos similares, en lugar de proceder a una mera estimación o evaluación probabilística (Tversky y Kahneman, 1973). Este sesgo puede afectar a la percepción y a la toma de decisiones en una amplia variedad de temas y contextos. Por ejemplo, Lichtenstein et al. (1978) encontraron que las personas sobreestiman la probabilidad de eventos espectaculares, pero poco comunes, como accidentes aéreos, debido a la mayor disponibilidad de estos eventos en la memoria.

Aplicado al tema que nos ocupa, la exposición repetida a noticias falsas puede llevar a terminar creyendo que son verdaderas (Pennycook y Rand, 2018). Este fenómeno, conocido como la verdad ilusoria, se debe a que las personas utilizan indicios periféricos de veracidad tales como la *familiaridad* (señal de que un mensaje ya nos resulta conocido), la *fluidez* de procesamiento (una señal de que un mensaje ha sido codificado o recuperado sin esfuerzo) y la *cohesión* (señal de que los elementos de un mensaje contienen referencias en la memoria que son internamente coherentes). La fuerza de estos indicios aumenta

con la repetición (para una revisión véase, Brashier y Marsh, 2020). Además, el estudio de Effron y Raj (2020) muestra que la difusión y exposición repetida a la desinformación hace que decrezca la sensación de inmoralidad, confirmando así el efecto de la mera exposición (Zajonc, 1968): la repetida visión de un objeto (una información en nuestro caso) hace que, al familiarizarnos con él, pierda los rasgos de extrañeza o incluso de rechazo que nos pudiera haber causado en un primer momento.

La injerencia extranjera suele aprovechar este sesgo inundando las redes sociales con una misma desinformación o noticia falsa, lo que aumenta la disponibilidad de estos eventos en la mente de las personas y altera su valoración de los hechos. Además, las personas pueden basar su identificación con un partido político en la facilidad con la que recuerdan eventos o información que respaldan sus creencias y prejuicios (Lodge y Taber, 2013). Al proceder así y desestimar información contradictoria, las personas terminan radicalizándose (Westfall et al., 2015). Un ejemplo notorio de este fenómeno fue la intervención de Rusia en las elecciones presidenciales de Estados Unidos en 2016, donde se llevaron a cabo numerosas campañas de desinformación *online* para influir en la opinión pública y fomentar la polarización política (p. ej., Jamieson, 2018; Allcott y Gentzkow, 2017).

DESTREZAS PARA DETECTAR LA DESINFORMACIÓN

Se han propuesto varias estrategias de intervención para ayudar a los ciudadanos a desarrollar habilidades y actitudes que les permitan reconocer, evaluar y resistir a la desinformación y la manipulación informativa en sus distintas formas.

Programas de alfabetización mediática

Los programas llamados de alfabetización mediática, que buscan enseñar a las personas a pensar críticamente sobre la información que consumen y a evaluar la credibilidad y veracidad de las fuentes, pueden ser una forma efectiva de prevenir la difusión de desinformación y noticias falsas (Kahne y Bowyer, 2017). Estos programas incluyen actividades y ejercicios prácticos tales como la identificación de sesgos y prejuicios que tergiversan la interpretación de la información, el análisis de la estructura y contenido de las noticias, y el uso de herramientas y recursos *online* para comprobar la veracidad de la información. Además, la alfabetización mediática debe abordar la importancia de la diversidad de fuentes de información y la exposición a perspectivas alternativas, así como el papel de las emociones y la identidad partidista en la percepción y difusión de la desinformación. La inclusión de la alfabetización mediática en el currículo escolar puede ayudar a desarrollar habilidades de pensamiento crítico en los estudiantes desde una edad temprana (Hobbs, 2010). Por ejemplo, en un estudio realizado por Mihailidis y Viotty (2017) se observó que los estudiantes que participaron en un programa de alfabetización mediática mostraron una mayor capacidad para identificar noticias falsas y una mayor confianza en sus habilidades para evaluar la información *online*.

Otra de las estrategias para contrarrestar la desinformación fomentando la participación ciudadana y la colaboración público-privada es a través de campañas de sensibilización dirigidas a la adopción de prácticas de verificación de las fuentes de información y de sus contenidos (Arroyo Guardado et al., 2023). Es el ciudadano, desde su criterio y juicio, el que va a determinar la veracidad de la información a través de este mecanismo de doble verificación. En este supuesto no hay una censura previa ni un tercero que determine la veracidad de la información. A nivel operativo, estas campañas de sensibilización podrían definirse a partir del modelo de las “4Ds: Detener, Denunciar, Desmontar, Difundir” propuesto por Rivas (2021). Con el fin de reducir el volumen de contenidos falsos o manipulados en las redes sociales, este modelo propone cuatro tipos de respuestas y fases de implicación por parte de los usuarios:

- Primer nivel de respuesta: detener la propagación. Prevé la inacción o la no difusión de la noticia detectada como posiblemente falsa.

- Segundo nivel de respuesta: denunciar para impulsar la retirada de contenidos. Promueve que los usuarios denuncien las noticias falsas a los proveedores de contenido para facilitar su bloqueo o retirada.
- Tercer nivel de respuesta: desmontar la desinformación. Fomenta que los usuarios verifiquen los contenidos, desarticulando los distintos elementos que componen una noticia falsa y haciendo público el resultado del análisis y las evidencias que apuntan a su falsedad.
- Cuarto nivel de respuesta: difundir los contenidos verificados y fomentar una cultura de prevención de la desinformación. Promueve la denuncia del suceso entre los contactos del usuario (familiares, amigos compañeros de trabajo) a fin de ayudarles a ser más conscientes de la presencia de desinformación en su entorno.

La difusión de herramientas de verificación de acceso libre, de manuales y cursos puede ayudar a que los usuarios sean capaces de detectar y descartar contenidos engañosos. Es de especial relevancia la disponibilidad de material divulgativo de acceso libre para los más jóvenes (véase, por ejemplo, la guía “¿Cómo protegerme de la desinformación?”).⁵

Verificación de información

La verificación de información (*fact-checking*) por parte de agencias especializadas es otra estrategia para contrarrestar la desinformación. La velocidad de respuesta frente a las noticias falsas se plantea como un hito para evitar que estas avancen ni siquiera como *leyendas urbanas*. La comprobación rápida es imprescindible para luchar contra la desinformación. Los desmentidos deben también viralizarse, lo que requiere de contenidos ricos y adaptados al entorno al que van dirigidos. El proceso de *fact-checking* es especialmente relevante en el contexto de las redes sociales y las plataformas de noticias *online*, donde la información se difunde de manera rápida y a menudo sin verificación previa. Algunos estudios (p. ej., Vosoughi et al., 2018) han mostrado que, cuando se les presenta información corregida, las personas son capaces de cambiar sus opiniones y creencias erróneas. No obstante, hay que recordar que esto posiblemente sea menos efectivo en el caso de personas con fuertes identidades partidistas, ya que pueden interpretarlo como una amenaza a su existencia, mostrando entonces un “efecto *boomerang* o *backfire*” y aferrándose aún más a sus creencias erróneas (Nyhan y Reifler, 2010).

La prensa es el cuarto poder, y su existencia contribuye al equilibrio del conjunto del sistema democrático. Los medios son los principales interesados en luchar contra las noticias falsas. El periodismo como marca y fuente de información veraz y de calidad debe ser potenciado.

⁵ Arroyo Gaurdeño et al., 2023.

Cuando una noticia falsa se cuela en un ciclo informativo no solo alcanza altas cotas de veracidad, sino que contamina al conjunto del sistema. La fortaleza de una cadena queda consignada a su eslabón más débil. Para reforzarla, el periodista profesional ha de disponer de una formación y un bagaje que le distinga sensiblemente del llamado “periodismo cívico”, para poder acometer así un cribado eficaz y convertirse en referencia social. Las grandes empresas de comunicación y todo el mundo del periodismo deben ir más lejos y compartir la información para facilitar el abordaje sinérgico de esta problemática. Además, es conveniente la adhesión voluntaria de las empresas a códigos de buenas prácticas y a esquemas internos que impulsen la responsabilidad social corporativa.

Diversidad de fuentes de información y autoregulación profesional

Fomentar el hábito de recurrir a la diversidad de fuentes de información y la exposición a perspectivas alternativas también puede ayudar a reducir la dependencia de las redes sociales y otras fuentes de información no verificadas o sesgadas, al ofrecer una variedad de medios y contenidos informativos de alta calidad (Valverde-Berrocoso et al., 2022). Igualmente, la diversidad de fuentes de información y la exposición a perspectivas alternativas puede contribuir a combatir la conformidad con la desinformación y la manipulación informativa, promoviendo la independencia de pensamiento y la autonomía en la formación de opiniones y creencias (p. ej., Garrett, 2009). De esta manera hay posibilidad de contrarrestar los efectos de la polarización y la influencia normativa, al proporcionar un contexto más amplio y equilibrado para la evaluación de la información y las opiniones.

El periodismo, como parte del pensamiento crítico, no se puede ni se debe regular. Los costos en términos de legitimidad son muy superiores a cualquier eventual beneficio. Pero se debe promover una suerte de código deontológico que favorezca un profesionalismo auto regulativo y atento frente a eventuales actitudes poco éticas de los miembros de la profesión. Fortalecer el periodismo pasa por fortalecer financieramente a las empresas que lo llevan a cabo, y en última fortalecer la marca. Ya en un informe elaborado para la Comisión Europea por un grupo de expertos de alto nivel se recomendaba apoyar financieramente a grupos informativos independientes, libres de potenciales interferencias de autoridades públicas y de compañías tecnológicas que pudieran verse tentadas de utilizar dichos proyectos como escaparates en sus relaciones públicas (Jiménez Cruz et al., 2018).

Educación cívica

La educación cívica —es decir, el de proceso de instruir a las personas sobre sus derechos, responsabilidades y roles como ciudadanos en una sociedad democrática— puede ayudar a los individuos a resistir la desinformación, al aumentar su comprensión de los procesos políticos y sociales (Miles, 2021). Es una vía para contrarrestar los efectos de la identidad partidista y la polarización ideológica, promover la tolerancia y el respeto por las opiniones y creencias de los demás y fomentar al mismo tiempo la responsabilidad cívica y la colaboración en la solución de problemas comunes (Kaufman, 2021). Por tanto, otro de los ejes para luchar contra la desinformación se sitúa en el humanismo, en la educación de la sociedad. De hecho, todo proyecto político de calado pasa por la educación. La educación refuerza la transversalidad y evita la fragmentación, la disgregación de la sociedad, y su debilitamiento. Una sociedad precisa de referentes intelectuales, de faros en lo alto de la montaña, personajes u organismos encumbrados a los que poder seguir. Pero también de figuras críticas que, permanentemente y desde la buena fe, cuestionen el modelo vigente mientras proponen universos distintos y nuevos sueños. Las sociedades se crecen en sus contradicciones.

ESTRATEGIAS DE RESISTENCIA A LA DESINFORMACIÓN

Planteamos finalmente una revisión de las distintas estrategias que se han desarrollado para combatir la desinformación. Una de las dificultades para la eliminación de la desinformación en las redes —a pesar de su contenido dañino e incluso cuando su incumplimiento de los Términos de Servicio (ToS) de las empresas— reside en la protección legal que ampara a los contenidos. Actualmente, los esfuerzos se centran en el código de buenas prácticas: unas normas de autorregulación para luchar contra la desinformación que la industria ha acordado, de forma voluntaria. En este código (Comisión Europea, 2022a) se recoge un conjunto de mejores prácticas firmadas por Facebook, Google, Twitter, Mozilla (2018), Microsoft (2019) y TikTok (2020), entre otros.

Prevención y respuesta frente al *Sharp Power* y la desinformación

La cooperación internacional y la acción conjunta es otra de las claves ante esta clase de desinformación que participa de las estrategias del *Sharp Power*, tal y como se describieron al inicio del capítulo. En no pocas ocasiones, los problemas son comunes y requieren de la actuación concertada, la coordinación y el intercambio de información y de experiencia que aseguren un efecto sinérgico beneficioso para todos y en diferentes planos, desde el nivel político hasta el táctico. El auxilio y la implicación de la comunidad internacional supone además una confirmación del consenso alrededor de una causa. Así, no es igual que un solo país decida sobre la prohibición de un medio de comunicación a que lo haga la comunidad internacional en su conjunto o, en concreto, la Unión Europea.

La clave de la lucha contra la desinformación no es la censura; lo importante es la legitimidad del marco institucional, que es precisamente lo que en el fondo la desinformación pretende cuestionar. El objetivo principal es proteger los valores atacados, que deben permanecer en el centro del sistema, y el futuro plural y democrático que pretende la sociedad para sí misma. Este lugar no puede ser ocupado por el problema, pues eso desorienta ante cualquier posible respuesta, priva de la iniciativa a quien debe responder y, finalmente, somete a las sociedades a la dictadura del agente perturbador. La democracia y la libertad se basan en la evidencia y la verdad; por eso la educación, junto al periodismo y la justicia, son elementos claves para su defensa, y deben reforzarse como parte de la respuesta. Promover una ciudadanía responsable es un proyecto a largo plazo que pasa necesariamente por una educación lo más sólida posible. En cualquier caso, es imprescindible actuar con prevención, y no militarizar amenazas que no lo son, así como espacios y ámbitos que pertenecen a la sociedad civil. Nuestros valores no se defienden cambiándolos, por más que puedan y deban realizarse las modificaciones organizacionales y de autoridad precisas para atender a las demandas de respuesta que estas nuevas amenazas plantean; pero

ese proceso nunca ha de actuar contra las esencias de lo que se trata de defender. Como recuerda el sociólogo Todorov (2014), “La democracia tiene mucho más que temer de las perversiones o desvíos del proyecto democrático que vienen del interior.”

Finalmente, la ciberseguridad y las implicaciones del uso de tecnologías son campos en los que hay que continuar profundizando, por la relevancia estratégica de muchos de los aspectos considerados y las vulnerabilidades que introducen en el sistema. El uso del marco conceptual DISARM ayuda a establecer un diálogo entre el mundo de la ciberseguridad y el mundo del periodismo y de los verificadores. DISARM es una herramienta de trabajo que deriva de la metodología ATTCK *framework* de MITRE,⁶ una base de conocimientos de acceso global sobre tácticas, técnicas y procedimientos (TTP) de adversarios basada en observaciones del mundo real. Terp y Breuer (2022) propone el Adversarial Misinformation and Influence Tactics and Techniques (AMITT), cuyo fin es combatir y detectar la desinformación a través del estudio de las tácticas y técnicas utilizadas por los distintos actores involucrados en una campaña de desinformación.

Los efectos no deseados de la censura

Como ya se ha comentado desde la perspectiva de la ética y la filosofía política, la utilización de la censura atenta contra los valores democráticos. Nos detenemos ahora un poco más en el fenómeno de la censura desde un enfoque psicosocial, ya que, aunque no lo parezca, hablamos de un hecho complejo y multifacético, que puede adoptar diversas formas y afectar diferentes aspectos de la vida social y política. Se conceptualiza aquí como el instrumento utilizado por gobiernos, instituciones y otros actores para limitar o controlar el flujo de información y los contenidos en diversos medios de comunicación. Varios estudios sobre los efectos de la censura como resistencia a la influencia social muestran que puede generar una sensación de escasez de información y, por lo tanto, incrementar el interés y la curiosidad por ella (Bushman, 1998; Zillmann y Brosius, 2012). Por ejemplo, Metzger et al. (2015) encontraron que la censura de noticias *online* relacionadas con temas políticos en China llevó a un aumento de búsqueda y consumo de información alternativa. Asimismo, Tsfati et al. (2018) observaron que la censura de contenidos en las redes sociales en Israel incrementó la percepción de escasez de información y la curiosidad por la información censurada, lo que al final resultó en una mayor exposición a la desinformación. Por otra parte, la censura puede provocar una reacción negativa en el público, ya que existe el riesgo de que sea percibida como una violación de sus derechos a la libertad de expresión e información (Bollinger, 1986), llevando a las personas a desafiar la censura y buscar activamente la información prohibida (Jowett y O'Donnell, 2012). Además, puede aumentar la desconfianza en las fuentes de información oficiales o en las autoridades que la administran (Tsfati y Ariely, 2014), fomentando la creación de cámaras de eco y comunidades *online* que se toman como un reto descubrir, compartir y difundir esa información censurada (Sunstein, 2009).

⁶ Véase en la página web: <https://www.mitre.org/>

La reactancia psicológica

La reactancia psicológica es una respuesta emocional negativa que surge cuando las personas perciben que su libertad personal está siendo amenazada o restringida (Brehm, 1966). Aunque los individuos tal vez estén dispuestos a mostrar conformidad con las normas del grupo, también pueden experimentar una fuerte motivación para resistirse a esa influencia cuando perciben que su autonomía y libertad personal están siendo coaccionadas. Para recuperar su sentimiento de libertad, las personas reaccionan contra todo aquello que perciben como una imposición. La reactancia puede ser especialmente pronunciada cuando las personas se sienten altamente identificadas con sus creencias y valores (Hornsey, 2005).

Las campañas de manipulación de la información para influir en las opiniones sobre temas políticos o sociales verán decrecer su eficacia en la misma medida en que las personas a las que van dirigidas sientan amenazada su libertad para formar sus propias opiniones (Miller et al., 2007). La reactancia también desempeña un papel en la polarización política y el extremismo. Investigaciones recientes han mostrado que las personas que se sienten amenazadas por la manipulación de la información pueden adoptar posiciones opuestas más extremas (Lantian et al., 2018).

Para mejorar la resistencia a la desinformación en contextos de identidad partidista y polarización, habría que implementar estrategias que fomenten la exposición a fuentes de información diversas. Esto puede incluir el fomento del consumo de medios de comunicación no partidistas y la promoción de espacios de debate e intercambio de ideas (Levendusky, 2018). Estas estrategias pueden aumentar la resiliencia de una comunidad, es decir, su capacidad de (i) absorber la presión a través de su resistencia o adaptación; (ii) gestionar o mantener ciertas funciones y estructuras básicas durante contingencias, y (iii) recuperarse después de un evento adverso.

En resumen, es importante que las estrategias de intervención en una u otra dirección tengan presente la posibilidad de generar reactancia. Es necesario que la eventual información correctiva sea presentada de manera que no sea percibida como una amenaza a la libertad de las personas o a su identidad. Esto puede incluir el uso de fuentes creíbles y confiables, la presentación de información de manera equilibrada, y el evitar un enfoque de confrontación que pueda aumentar la reactancia (Hornsey et al., 2018).

Teoría de la inoculación

Una notable preocupación —sobre todo en el ámbito militar— en las décadas de 1950 y 1960 era cómo resistir a la propaganda enemiga. A raíz de la guerra de Corea (1950-1952), el periodista Hunter (1951) introdujo el término *brainwashing* para explicar los actos inesperados de traición cometidos por algunos soldados estadounidenses que habían sido

hechos prisioneros. Se supo que un buen número de ellos cooperaron voluntariamente con el enemigo. Fue un fiasco impactante para los mandos militares, y varios científicos sociales intentaron explicar, o al menos comprender, lo que había sucedido. Inicialmente se especuló con que el enemigo había utilizado una combinación ingeniosa de tortura y castigo para convertirlos a sus planteamientos políticos e ideológicos. Sin embargo, la documentación disponible sugería que las sesiones de lavado de cerebro apenas incluían tortura, sino que consistían en un debate prolongado entre el soldado capturado y un interrogador habilidoso. Ese debate giraba en torno a América y a las creencias estadounidenses sobre la libertad, la democracia o la igualdad. Aunque los soldados creían, por ejemplo, que la democracia era la mejor forma de gobierno, resultó llamativo comprobar que tenían serias dificultades para argumentarlo. Durante los interrogatorios se atacaban esas creencias y se les llevaba a que dudaran de su validez. Una vez convertidos, el camino hacia la “traición” estaba allanado.

Por esa misma época, Lifton (1961) publicó también su libro con las entrevistas que realizó a medio centenar de occidentales apresados en China que habían sido sometidos al programa de “reforma del pensamiento” empleado bajo el mandato de Mao Zedong para conseguir la conversión de la mentalidad occidental a la ideología comunista. Allí se describen toda una serie de técnicas psicológicas de manipulación de las creencias, actitudes y comportamientos de la persona pensadas para lograr dicha conversión.

En ese contexto, especialistas en la persuasión y cambio de actitudes de la Escuela de Yale (p. ej., Lumsdaine y Janis, 1953) empezaban a demostrar que proporcionar a una persona únicamente argumentos que respaldasen sus creencias, sin mencionar posibles argumentos en contra, resultaba poco eficaz para lograr la resistencia a la contrapersuasión. Tras un balance de esos estudios, McGuire (1961) formuló la teoría de la inoculación. La hipótesis inicial partía del postulado de la “exposición selectiva”: las personas tienden a defender sus creencias evitando la exposición a contrargumentos. Dado así el entorno ideológico “aséptico” resultante, la persona tiende a mantenerse altamente confiada en sus creencias, pero también en una posición altamente vulnerable si resulta que tiene que enfrentarse a contraargumentos trabados.

La teoría plantea que preexponer a la persona a pequeños contraargumentos que atacan sus creencias, seguidos de una refutación detallada de ellos, produce una inmunidad considerable a la contrapersuasión. Se esperaría que estas preexposiciones, análogas a la inoculación con un virus debilitado de una persona criada en un entorno libre de gérmenes, estimulen la defensa de sus creencias, mejorando su capacidad de resistir la exposición masiva subsiguiente. Este análisis es particularmente apropiado cuando las creencias involucradas son tópicos culturales, que McGuire llama *truismos*. Los define como una creencia ampliamente aceptada, o evidente por sí misma, que generalmente se da por asentada dentro de una cultura o sociedad en particular. Los truismos se perciben a menudo como incuestionablemente verdaderos y requieren poco o ningún debate o justificación. Pueden abarcar nociones de sentido común, principios morales o normas culturales que están profundamente arraigadas en la mente de las personas, y son compartidas por la mayoría de la población. En el ámbito de la verificación, la teoría de la inoculación ha

adquirido una forma específica en el llamado *prebunking*, voz inglesa construida a partir del prefijo de anterioridad, *pre*, y el sustantivo *bunk*, que significa mentira o patraña, y por contraste al verbo *debunk*, desmentir en español. El *prebunking* tiene por objeto combatir la desinformación proporcionando ejemplos inofensivos de desinformación, para así generar resistencia futura ante afirmaciones falsas similares (Roosenbeek et al., 2022).

En resumen, el proceso de inoculación es un enfoque de resistencia a la persuasión basado en la idea de que amenazar a las personas con argumentos débiles en contra de sus creencias o actitudes puede fortalecer su resistencia a argumentos más fuertes y persuasivos en el futuro. Así, un mensaje de inoculación está diseñado para provocar una amenaza que actúa como catalizador motivacional, inspirando al individuo a apuntalar sus defensas de cara a los próximos desafíos actitudinales. La investigación ha demostrado que, cuanto más activamente se defiende el receptor contra el ataque, más fuerte se volverá la actitud existente. La estrategia de inmunización activa, que consiste en exponer a las personas a ejemplos de noticias falsas para que aprendan a refutarlas y a identificar los métodos comunes utilizados en la manipulación de la información, puede ser una herramienta útil en la lucha contra la desinformación *online*, y especialmente útil en contextos políticos y sociales polarizados, donde la desinformación y la propaganda son elementos habituales (Guess et al., 2019).

CONCLUSIÓN

Las sociedades del siglo XXI se definen así mismas como “de la información”. Éstas consumen grandes cantidades de información, pues tanto las tecnologías de edición como las de distribución de la información facilitan el acceso. El problema no es tanto la información como su selección. La desinformación, en términos generales, es todo aquello que impide el uso correcto de la información disponible. No se sustancia necesariamente en la mentira y el engaño, sino en la alteración de la agenda informativa en la escala considerada; no se trata solo de la verdad y su cualidad, sino también de la adecuada ponderación de los hechos en su contexto por parte de grupos en contraposición.

La desinformación afecta a los cimientos de la democracia, pues esta se sustenta sobre las decisiones pretendidamente libres e incondicionadas de los ciudadanos que, de este modo, se ven afectadas en sus criterios de adopción. Aspira a generar un clima de desconfianza, amenaza, miedo o polarización en la sociedad, socava la confianza en las instituciones democráticas y enrarece la facultad de las sociedades para comunicarse y tomar decisiones que estén fundadas en un análisis racional de los hechos, o en principios morales y democráticos (p. ej., McKay y Tenove, 2020). Finalmente, puede fomentar el extremismo y la radicalización, lo que representa una amenaza para la seguridad nacional e internacional.

En consecuencia, la desinformación tiene graves consecuencias sociales, políticas y económicas, y plantea, según se ha visto, un grave riesgo para la seguridad nacional, al tensionar tanto a la sociedad como al aparato que la soporta. Contrarrestar la desinformación obliga a la adopción de fórmulas estables que sirvan a una conciliación entre la libertad de información y la seguridad en el marco cultural y democrático, frente a los desafíos que las dinámicas tecnológicas aplicadas a las redes sociales van a generar. La resistencia a la información manipulada es cardinal para proteger la democracia (Lewandowsky et al., 2017).

En este capítulo se han analizado las dinámicas de la desinformación en el marco de las relaciones sociales y de los intereses grupales a los que sirven. Hemos visto que la desinformación resulta funcional a las dinámicas de la propagación y de la propaganda (Moscovici, 1976). Además, hemos hablado de la relación entre desinformación y radicalización en el contexto geopolítico actual, identificando en la segunda parte estrategias de prevención y de resistencia a la desinformación.

BIBLIOGRAFÍA

- Allcott, H., y Gentzkow, M. (2017). Social media and fake news in the 2016 election. *Journal of Economic Perspectives*, 31(2), 211-236.
- Anderson, B. (2006). *Imagined communities: Reflections on the origin and spread of nationalism*. Verso Books.
- Arroyo Guardado, D., Degli-Esposti, S., Gómez-Espés, A., Palmero-Muñoz, S., y Pérez-Miguel, L. (2023). *On the Design of a Misinformation Widget (MsW) Against Cloaked Science*. Proceedings of the 9th International Symposium on Security and Privacy in Social Networks and Big Data, University of Kent, Canterbury, UK, August 14-16. Springer.
- Asch, S. E. (1969). A reformulation of the problem of associations. *American Psychologist*, 24(2), 92.
- Bakir, V., y McStay, A. (2018). Fake news and the economy of emotions. *Digital Journalism*, 6(2), 154-175.
- Balcells, L., y Kuo, A. (2022). Secessionist conflict and affective polarization: Evidence from Catalonia. *Journal of Peace Research, Peace Research Institute Oslo*, 60(4), 604-618
- Bandura, A. (1999). Moral disengagement in the perpetration of inhumanities. *Personality and social psychology review*, 3(3), 193-209.
- Barberá, P., Jost, J. T., Nagler, J., Tucker, J. A., y Bonneau, R. (2015). Tweeting from left to right: Is online political communication more than an echo chamber? *Psychological Science*, 26(10), 1531-1542.
- Bar-Tal, D. (1996). Development of social categories and stereotypes in early childhood: The case of "the Arab" concept formation, stereotype and attitudes by Jewish children in Israel. *International journal of intercultural relations*, 20(3-4), 341-370.
- Blanco, A. (2022) *El miedo al otro*. Editorial Sanz Torres.
- Blitvich, P.G.C., y Lorenzo-Dus, N., (2022). Go ahead and 'debunk' truth by calling it a conspiracy theory: The discursive construction of conspiracy theories in online affinity spaces. In V. Massimiliano Demata, V. Zorzi y A. Zottola (Eds), *Conspiracy Theory Discourses* (pp. 71-98). John Benjamins. <https://doi.org/10.1075/dapsac.98.04gar>

- Bradley, F., Bangerter, A., Bauer, M. W., Hall, M., y Noort, M. C. (2017). Beyond “monologicality”? Exploring conspiracist worldviews. *Frontiers in psychology*, 8, 861. <https://doi.org/10.3389/fpsyg.2017.00861>
- Brashier, N. M., y Marsh, E. J. (2020). Judging truth. *Annual Review of Psychology*, 71, 499-515.
- Brehm, J. W. (1966). *A theory of psychological reactance*. Academic Press.
- Bruns, A. (2005). *Gatewatching: Collaborative online news production*. Peter Lang Inc., International Academic Publishers.
- Bushman, B. J. (1998). Effects of warning and information labels on consumption of full-fat, reduced-fat, and no-fat products. *Journal of Applied Psychology*, 83(1), 97-101.
- Cardenal, A. S., Aguilar-Paredes, C., Galais, C., y Pérez-Montoro, M. (2019). Digital technologies and selective exposure: How choice and filter bubbles shape news media exposure. *The international journal of press/politics*, 24(4), 465-486. <https://doi.org/10.1177/1940161219862988>
- Cialdini, R. B., y Goldstein, N. J. (2004). Social influence: Compliance and conformity. *Annual Review of Psychology*, 55, 591-621. <https://doi.org/10.1146/annurev.psych.55.090902.142015>
- Codol, J. P. (1975). «Effet PIP» et conflit de normes. *Année Psychologique*, 75(1), 127-145.
- Comisión Europea (2022a). *The Strengthened Code of Practice on Disinformation 2022*. <https://digital-strategy.ec.europa.eu/en/library/2022-strengthened-code-practice-disinformation>
- Comisión Europea (2022b). *Standard Eurobarometer 96. Annex: Fieldwork: January – February 2022*. Directorate-General for Communication. <https://europa.eu/eurobarometer/api/deliverable/download/file?deliverableId=81059>
- Coupland, N. (2010). ‘Other’ Representation. In J. Jaspers, J.O. Östman y J. Verschueren (Eds.), *Society and Language Use* (242-260), John Benjamins. <https://doi.org/10.1075/hop.5.oth1>

Crawford, J. T., y Pilanski, J. M. (2014). The differential effects of right-wing authoritarianism and social dominance orientation on political intolerance. *Political Psychology*, 35(6), 557-576.

Degli-Esposti, S., y Arroyo, D. (2023) ¿Cómo protegerme de la desinformación? La Aventura de Aprender. Instituto Nacional de Tecnologías Educativas y de Formación del Profesorado. Ministerio de Educación y Formación Profesional.

Del Vicario, M., Bessi, A., Zollo, F., Petroni, F., Scala, A., Caldarelli, G., Stanley, H.E., y Quattrocioni, W. (2016). The spreading of misinformation online. *Proceedings of the National Academy of Sciences*, 113(3), 554-559. <https://doi.org/10.1073/pnas.1517441113>

Dubois, E., y Blank, G. (2018). The echo chamber is overstated: the moderating effect of political interest and diverse media. *Information, Communication & Society*, 21(5), 729-745.

Effron, D. A., y Raj, M. (2020). Misinformation and Morality: Encountering Fake-News Headlines Makes Them Seem Less Unethical to Publish and Share. *Psychological Science*, 31(1), 75-87.

FECYT (febrero, 2022). *Encuesta de Percepción Social de la Ciencia y la Tecnología 2022*. <https://www.fecyt.es/es/noticia/encuestas-de-percepcion-social-de-la-ciencia-y-la-tecnologia-en-espana>

Festinger, L. (1957). *A theory of cognitive dissonance*. Stanford University Press.

Flaxman, S., Goel, S., y Rao, J. M. (2016). Filter bubbles, echo chambers, and online news consumption. *Public Opinion Quarterly*, 80(S1), 298-320. <https://doi.org/10.1093/poq/nfw006>

Flynn, D.J., Brendan N., y Reifler, J. (2017). The nature and origins of misperceptions: Understanding false and unsupported beliefs about politics. *Political Psychology*, 38(S1), 127-150. <https://doi.org/10.1111/pops.12394>

Frey, D. (1986). Recent research on selective exposure to information. *Advances in Experimental Social Psychology*, 19, 41-80. [https://doi.org/10.1016/S0065-2601\(08\)60212-9](https://doi.org/10.1016/S0065-2601(08)60212-9)

Furlow, E., Fleischer, K., y Corman, S.R. (2014). *De-romanticizing the Islamic State's vision of the Caliphate* (Informe No. 1402). Centre for Strategic Communication. . <https://csc.asu.edu/wp-content/uploads/pdf/csc1402-deromanticizing-islamic-state-caliphate.pdf>

García Servet, R., y Calvo Albero, J. L. (2022). *El dominio cognitivo en las operaciones multidominio: concepto y problemático*. Academia de las Ciencias y las Artes Militares. <https://www.acami.es/wp-content/uploads/2022/05/dominio-cognitivo-operaciones-multidominio-web.pdf>

Garrett, R. K. (2009). Echo chambers online? Politically motivated selective exposure among Internet news users. *Journal of Computer-Mediated Communication*, 14(2), 265-285. <https://doi.org/10.1111/j.1083-6101.2009.01440>

Gee, J. P. (2005). Semiotic social spaces and affinity spaces. En D. Barton y K. Tusting (Eds), *Beyond Communities of Practice: Language Power and Social Context* (pp. 214-232). Cambridge University Press <https://doi.org/10.1017/CBO9780511610554.012>

Gómez, A., Brooks, M. L., Buhrmester, M. D., Vázquez, A., Jetten, J., y Swann, W. B. (2011). On the nature of identity fusion: insights into the construct and a new measure. *Journal of personality and social psychology*, 100(5), 918-933. <https://doi.org/10.1037/a0022642>

Gómez, Á., López-Rodríguez, L., Vázquez, A., Paredes, B., y Martínez, M. (2016). Morir y matar por un grupo o unos valores. Estrategias para evitar, reducir y/o erradicar el comportamiento grupal extremista. *Anuario de Psicología Jurídica*, 26(1), 122-129. <https://doi.org/10.1016/j.apj.2016.04.001>

Gómez, A., López-Rodríguez, L., Sheikh, H., Ginges, J., Wilson, L., Waziri, H., Vázquez, A., Davis, R., y Atran, S. (2017). The devoted actor's will to fight and the spiritual dimension of human conflict. *Nature Human Behaviour*, 1, 673-679. <https://doi.org/10.1038/s41562-017-0193-3>

Gómez, A., Chinchilla, J., Vázquez, A., López-Rodríguez, L., Paredes, B., y Martínez, M. (2020). Recent advances, misconceptions, untested assumptions and future research agenda for identity fusion theory. *Social and Personality Psychology Compass*, 14(6). <https://doi.org/10.1111/spc3.12531>

Gómez, A., Atran, S., Chinchilla, J., Vázquez, A., López-Rodríguez, L., Paredes, B., Martínez, M., Blanco, L., Alba, B., Bautista, H., Fernández, S., Pozuelo-Rubio, F., González-Álvarez, J.L., Chiclana, S., Valladares-Narganes, H., Alonso, M., Ruíz-Alvarado, A., López-Novo, J.L., y Davis, R. (2022). Willingness to sacrifice among convicted Islamist terrorists versus violent gang members and other criminals. *Scientific Reports*, 12(1), 1-15. <https://doi.org/10.1038/s41598-022-06590-0>

Gómez, A., Vázquez, A., Atran, S. (2023a). Transcultural pathways to the will to fight. *Proceedings of the National Academy of Sciences*, 120(24). <https://doi.org/10.1073/pnas.2303614120>

Gómez, Á., Vázquez, A., Chinchilla, J., Blanco, L., Alba, B., Chiclana, S., y González-Álvarez, J. L. (2023b). Why is it so difficult to investigate violent radicalization? *Spanish Journal of Psychology*, 26(7). <https://doi.org/10.1017/SJP.2023.2>

Guess, A., Nagler, J., Tucker, J. (2019). Less than you think: Prevalence and predictors of fake news dissemination on Facebook. *Science advances*, 5(1). <https://doi.org/10.1126/sciadv.aau4586>

Haidt, J. (2012). *The righteous mind: Why good people are divided by politics and religion*. Vintage.

Haslam, N. (2006). Dehumanization: An integrative review. *Personality and social psychology review*, 10(3), 252-264. https://doi.org/10.1207/s15327957pspr1003_4

Henschke, A., y Reed, A. (2021). Toward an ethical framework for countering extremist propaganda online. *Studies in Conflict & Terrorism*, 1-18.

HLEG (12 de marzo, 2018). *A multi-dimensional approach to disinformation. Report of the independent High -Level Group on fake news and online disinformation*. Luxembourg: Publications Office of the European Union. <https://digital-strategy.ec.europa.eu/en/library/final-report-high-level-expert-group-fake-news-and-online-disinformation>

Hobbs, R. (2010). *Digital and Media Literacy: A Plan of Action. A White Paper on the Digital and Media Literacy Recommendations of the Knight Commission on the Information Needs of Communities in a Democracy*. Aspen Institute.

Horgan, J., y Braddock, K. (2010). Rehabilitating the terrorists? Challenges in assessing the effectiveness of de-radicalization programs. *Terrorism and Political Violence*, 22(2), 267-291. <https://doi.org/10.1080/09546551003594748>

Hornsey, M. J. (2005). Why being right is not enough: Predicting defensiveness in the face of group criticism. *European Review of Social Psychology*, 16(1), 301-334.

Hornsey, M. J., Harris, E. A., y Fielding, K. S. (2018). The psychological roots of anti-vaccination attitudes: A 24-nation investigation. *Health psychology*, 37(4), 307-315. <http://dx.doi.org/10.1037/hea0000586>

Jamieson, K. H. (2018). *Cyberwar: How Russian hackers and trolls helped elect a president: What we don't, can't, and do know*. Oxford University Press

Kathleen Hall, J., y Cappella, J. N. (2008). *Echo chamber: Rush Limbaugh and the conservative media establishment*. Oxford University Press.

Jiménez Cruz, C., Mantzarlis, A., Nielsen, R. N., y Wardle, C. (2018). Six Key Points from the EU High Level Group's New Report on Disinformation. First Draft March 12. <https://firstdraftnews.org/articles/eu-report-disinformation/>

Kahne, J., y Bowyer, B. (2017). Educating for democracy in a partisan age: Confronting the challenges of motivated reasoning and misinformation. *American Educational Research Journal*, 54(1), 3-34. <https://doi.org/10.3102/0002831216679817>

Kaufman, C. (2021). Civic education in a fake news era: Lessons for the methods classroom. *Journal of Political Science Education*, 17(2), 326-331.

Kelman, H. C.; Hamilton, V. L. (1989). *Crimes of obedience: Toward a social psychology of authority and responsibility*. Yale University Press New Haven.

Kruglanski, A. W., Fernandez, J. R., Factor, A. R., y Szumowska, E. (2019). Cognitive mechanisms in violent extremism. *Cognition*, 188, 116-123. <https://doi.org/10.1016/j.cognition.2018.11.008>

Kunda, Z. (1990). The case for motivated reasoning. *Psychological Bulletin*, 108(3), 480-498.

Lantian, A., Muller, D., Nurra, C., y Douglas, K. M. (2018). "I know things they don't know!" The role of need for uniqueness in belief in conspiracy theories. *Social Psychology*, 48(3), 160-173. <https://doi.org/10.1027/1864-9335/a000306>

Levendusky, M. S. (2018). When efforts to depolarize the electorate fail. *Public Opinion Quarterly*, 82(3), 583-592. <https://doi.org/10.1093/poq/nfy036>

Lewandowsky, S., Ecker, U. K. H., y Cook, J. (2017). Beyond misinformation: Understanding and coping with the "post-truth" era. *Journal of Applied Research in Memory and Cognition*, 6(4), 353-369. <https://doi.org/10.1016/j.jarmac.2017.07.008>

Lichtenstein, S., Slovic, P., Fischhoff, B., Layman, M., y Combs, B. (1978). Judged frequency of lethal events. *Journal of Experimental Psychology: Human Learning and Memory*, 4(6), 551-578. <http://dx.doi.org/10.1037/0278-7393.4.6.551>

Lifton, R. J. (1961) *Thought Reform and the Psychology of Totalism: A Study of 'Brainwashing' in China*. William Clowes and Sons.

Lodge, M., y Taber, C. S. (2013). *The rationalizing voter*. Cambridge University Press.

Lorenzo-Dus, N. (2023). *Digital Grooming: Discourses of Manipulation and Cyber-crime*. Oxford University Press.

Lorenzo-Dus, N. y Macdonald, S. (2018). Othering the West in the online Jihadist propaganda magazines Inspire and Dabiq. *Journal of Language Aggression and Conflict*, 6(1): 79-106. <http://dx.doi.org/10.1075/jlac.00004.lor>

Lumsdaine, A. A., y Janis, I. L. (1953). Resistance to "counterpropaganda" produced by one-sided and two-sided "propaganda" presentations. *Public Opinion Quarterly*, 17, 311-318. <https://doi.org/10.1086/266464>

Maalouf, A. (2012). *Identidades asesinas*. Traducido por Fernando Villaverde Landa. Alianza Editorial.

Martín-Baró, I. (2003). *Poder, ideología y violencia*. Trotta.

Martínez, R., y Tulchin, S. J. (Eds.) (2006). *La seguridad desde las dos orillas: un debate entre Europa y Latinoamérica*. Ediciones Bellaterra.

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

McGuire W. J. (1961). The effectiveness of supportive and refutational defenses in immunizing and restoring beliefs against persuasion. *Sociometry* 24, 184–197.

McKay, S., y Tenove, C. (2020). Disinformation as a Threat to Deliberative Democracy. *Political Research Quarterly*, 74(3). <https://doi.org/10.1177/1065912920938143>

Metzger, M. J., Flanagin, A. J., Eyal, K., Lemus, D. R., McCann, y R. M. (2015). Credibility for the 21st century: Integrating perspectives on source, message, and media credibility in the contemporary media environment. *Annals of the International Communication Association*, 27(1), 293-335. <https://doi.org/10.1080/23808985.2003.11679029>

Mihailidis, P., y Viotty, S. (2017). Spreadable spectacle in digital culture: Civic expression, fake news, and the role of media literacies in "post-fact" society. *American Behavioral Scientist*, 61(4), 441-454. <https://doi.org/10.1177/0002764217701217>

Miles, J. (2021). The ongoing crisis and promise of civic education. *Curriculum Inquiry*, 51(4), 381-388.

Miller, C. H., Lane, L. T., Deatrck, L. M., Young, A. M., y Potts, K. A. (2007). Psychological reactance and promotional health messages: The effects of controlling language, lexical concreteness, and the restoration of freedom. *Human Communication Research*, 33(2), 219-240.

Milosevich-Juaristi, M. (10 de enero, 2017). *El poder de la influencia rusa: la desinformación*. Real Instituto Elcano. <https://www.realinstitutoelcano.org/analisis/el-poder-de-la-influencia-rusa-la-desinformacion/>

Moghaddam, F. M. (2005). The Staircase to Terrorism: A Psychological Exploration. *American Psychologist*, 60(2), 161-169.

Moscovici, S. (1976). *Social influence and social change*. Academic Press. S. Moscovici, G. Mugny, J. A. Pérez (Eds.), *La influencia social inconsciente: estudios de psicología social experimental*, Anthropos.

Moscovici, S., Mugny, G., y Pérez, J. A. (1985). Les effets pervers du déni (par la majorité) des opinions d'une minorité. *Bulletin de Psychologie*, 38(18), 803-812. [http://dx.doi.org/10.3406/bupsy.1985.12751](http://dx.doi.org/10.3406/buppsy.1985.12751)

Moscovici, S., Mugny, G., y Pérez, J. A. (Eds.) (1991). *La influencia social inconsciente: estudios de psicología social experimental*. Anthropos: Editorial del Hombre.

Mudde, C. (2021). Populism in Europe: an illiberal democratic response to undemocratic liberalism (The Government and Opposition/Leonard Schapiro Lecture 2019). *Government and Opposition*, 56 (4), 577-597. doi:10.1017/gov.2021.15

Neumann, P. R., y Rogers B. (2011). *Recruitment and Mobilisation for the Islamist Militant Movement in Europe*. International Centre for the Study of Radicalisation. https://icsr.info/wp-content/uploads/2008/10/1234516791ICSREURResearchReport_Proof1.pdf

Newman, N., Fletcher, R., Robertson, C. T., Eddy, K., and Nielsen, y R. K. (2022). *Digital News Report 2022*. Reuters Institute for the Study of Journalism. https://reutersinstitute.politics.ox.ac.uk/sites/default/files/2022-06/Digital_News-Report_2022.pdf

Nickerson, R. S. (1998). Confirmation bias: A ubiquitous phenomenon in many guises. *Review of General Psychology*, 2(2), 175-220. <https://doi.org/10.1037/1089-2680.2.2.175>

Nouri, L., y Lorenzo-Dus, N. (2019). Investigating reclaim Australia and Britain first's use of social media: developing a new model of imagined political communities online. *Journal for deradicalization*, 18, 1-34.

Nyhan, B., y Reifler, J. (2010). When corrections fail: The persistence of political misperceptions. *Political Behavior*, 32(2), 303-330. <https://doi.org/10.1007/s11109-010-9112-2>

Pariser, E. (2011). *The filter bubble: How the new personalized web is changing what we read and how we think*. Penguin.

Pennycook, G., y Rand, D. G. (2018). The Implied Truth Effect: Attaching Warnings to a Subset of Fake News Stories Increases Perceived Accuracy of Stories Without Warnings. *Management Science*, 66(11), 4944-4957. <https://doi.org/10.1287/mnsc.2019.3478>

Pérez, J. A., Ghosn, F., Chulvi, B., y Molpeceres, M. (2023) Does threat cause discrimination or does discrimination cause threat? *International Journal of Social Psychology*, 38(2), 279-303. <https://doi.org/10.1080/02134748.2022.2158589>

Pérez, J. A., y Mugny, G. (1985). Categorización e influencia minoritaria. *Anuario de Psicología*, 32, 97-116.

Pérez, J. A., Moscovici, S., y Chulvi, B. (2002). Natura y cultura como principio de clasificación social. Anclaje de representaciones sociales sobre minorías étnicas. *International Journal of Social Psychology*, 17(1), 51-68, <http://dx.doi.org/10.1174/021347402753408668>

Phadke, S., Samory, M., y Mitra, T. (2021). What makes people join conspiracy communities? Role of social factors in conspiracy engagement. *Proceedings of the ACM on Human-Computer Interaction*, 4(CSCW3), 1-30. <https://doi.org/10.1145/3432922>

Prida, E. (2 de mayo, 2017). *Las medidas activas*. Primavera digital en Cuba. <https://primaveradigital.org/cubaprimaveradigital/las-medidas-activas/>

Rivas, F. (2021). *Disinformation and Radicalisation*. Radicalisation Awareness Network (RAN). Luxembourg: Publications Office of the European Union.

Rodríguez Torres, R., Rodríguez Pérez, A., y Leyens, J. P. (2003). ¿Sienten los miembros de los otros grupos igual que nosotros? En J.F. Morales y C. Huici (Dirs.), *Estudios de psicología social* (pp. 55-82). UNED.

Roozenbeek, J., van der Linden, S., Goldberg, B., Rathje, S., y Lewandowsky, S. (2022). Psychological inoculation improves resilience against misinformation on social media. *Science Advances*, 8(34). <https://doi.org/10.1126/sciadv.abo6254>

Rosso, P. y Chulvi Ferriols, B. (2021). Los prejuicios en las sociedades digitales: identificación automática de estereotipos y 'hate speech'. *Archiletras científica: revista de investigación de lengua y letras*, 6, 171-188.

Sabucedo, J. M., Barreto, I., Borja, H., López López, W., Blanco, A., De la Corte, L., y Durán, M. (2004). Deslegitimación del adversario y violencia política: el caso de las FARC y las AUC en Colombia. *Acta Colombiana de Psicología*, 12, 69-85.

Said, Edward W. 1997. *Covering Islam: How the media and the experts determine how we see the rest of the world*. Vintage.

Said, E. W. 2003. *Orientalism*. Penguin.

Salaverría, R.; Buslón, N., López-Pan, F.; León, B., López-Goñi, I., y Erviti, M. C. (2020). Desinformación en tiempos de pandemia: tipología de los bulos sobre la Covid-19. *El profesional de la información*, 29(3). <https://doi.org/10.3145/epi.2020.may.15>

Sen, A. (2007). *Identity and violence: The illusion of destiny*. Penguin Books India.

Soledad Navas, M., Cuadrado Guirado, M. I., y López-Rodríguez, L. (2012). Fiabilidad y evidencias de validez de la Escala de Percepción de Amenaza Exogrupal (EPAE). *Psicothema*, 24(3), 477-482.

Sumner, W. G. (1906). *Folkways a Study of the Sociological Importance of Usages Manners Customs Mores and Morals*. New York: Ginn and Company.

Sunstein, C. R. (2001). *Echo chambers: Bush v. Gore, impeachment, and beyond*. Princeton University Press.

Sunstein, C. R. (2009). *Republic.com 2.0*. Princeton University Press.

Sutton, R. M., y Douglas, K. M. (2022). Rabbit Hole Syndrome: Inadvertent, accelerating, and entrenched commitment to conspiracy beliefs. *Current Opinion in Psychology*, 48, <https://doi.org/10.1016/j.copsyc.2022.101462>

Swann, W. B. Jr., Jetten, J., Gómez, Á., Whitehouse, H. y Bastian, B. (2012). When group membership gets personal: A theory of identity fusion. *Psychological Review*, 119(3), 441-456. <https://doi.org/10.1037/a0028589>

Taber, C. S., y Lodge, M. (2006). Motivated Skepticism in the Evaluation of Political Beliefs. *American Journal of Political Science*, 50(3), 755-769.

Tajfel, H. (1974). Social identity and intergroup behaviour. *Social science information*, 13(2): 65-93.

Tajfel, H., y Turner, J. C. (1986). The Social Identity Theory of Intergroup Behavior. En S. Worchel y L. W. Austin (Eds.), *Psychology of Intergroup Relations* (pp. 7-24). Nelson-Hall.

Tajfel, H. (1984). *Grupos humanos y categorías sociales*. Herder.

Terp, S. J. y Breuer, P. (2022). DISARM: a Framework for Analysis of Disinformation Campaigns, *2022 IEEE Conference on Cognitive and Computational Aspects of Situation Management (CogSIMA)*, Salerno, Italy, 2022, pp. 1-8, <https://doi.org/10.1109/CogSIMA54611.2022.9830669>

Todorov, T. (2014). *The inner enemies of democracy*. John Wiley & Sons.

Törnberg, P. (2022). How digital media drive affective polarization through partisan sorting. *Proceedings of the National Academy of Sciences*, 119(42). <https://doi.org/10.1073/pnas.2207159119>

Tsfati, Y., y Ariely, G. (2014). Individual and contextual correlates of trust in media across 44 countries. *Communication Research*, 41(6), 760-782.

Tsfati, Y., Ein-Gar, D., y Zukerman, G. (2018). Perceived information scarcity, curiosity, and news content choice. *Communication Research*, 45(6), 934-955.

Tversky, A., y Kahneman, D. (1973). Availability: A heuristic for judging frequency and probability. *Cognitive Psychology*, 5(2), 207-232. [https://doi.org/10.1016/0010-0285\(73\)90033-9](https://doi.org/10.1016/0010-0285(73)90033-9)

Valverde-Berrocoso, J., González-Fernández, A., y Acevedo-Borrega, J. (2022). Disinformation and multiliteracy: A systematic review of the literature. *Comunicar*, 30(70), 97-110.

Van Dijk, T. (1991). *Racism and the Press*. London: Routledge

Van Dijk, T. (1993). *Elite Discourse and Racism*. London: Sage.

Van Dijk, T. A. (2016). Discourse and racism: Some conclusions of 30 years of research. En A. Capone y J. L. Mey (Eds), *Interdisciplinary studies in pragmatics, culture and society* (pp. 285-295.). Springer

Varmann, A. H., Kruse, L., Bierwaczzonek, K., Gómez, A., Vázquez, A., y Kunst, J. R. (2023). How identity fusion predicts extreme pro-group orientations: A meta-analysis. *European Review of Social Psychology*. <https://doi.org/10.1080/10463283.2023.2190267>

Vázquez, A., López-Rodríguez, L., Martínez, M., Atran, S., y Gómez, Á. (2020). Threat Enhances Aggressive Inclinations Among Devoted Actors Via Increase in Their Relative Physical Formidability. *Personality and Social Psychology Bulletin*, 46(10), 1461–1475. <https://doi.org/10.1177/0146167220907466>

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

Wagner-Egger, P., Bangerter, A., Delouvée, S., y Dieguez, S. (2022). Awake together: Sociopsychological processes of engagement in conspiracist communities. *Current Opinion in Psychology*, 47. <https://doi.org/10.1016/j.copsyc.2022.101417>

Wardle, C., y Derakhshan, H. (27 de septiembre, 2017). Information disorder: *Toward an interdisciplinary framework for research and policymaking*. Consejo de Europa.

Wason, P. C. (1960). On the failure to eliminate hypotheses in a conceptual task. *Quarterly Journal of Experimental Psychology*, 12(3), 129-140. <https://doi.org/10.1080/17470216008416717>

Westfall, J., Van Boven, L., Chambers, J. R., y Judd, C. M. (2015). Perceiving Political Polarization in the United States: Party Identity Strength and Attitude Extremity Exacerbate the Perceived Partisan Divide. *Perspectives on Psychological Science*, 10(2), 145-158. <https://doi.org/10.1177/1745691615569849>

Winter, C., Neumann, P., Meleagrou-Hitchens, A., Ranstorp, M., Vidino, L., y Fürst, J. (2021). Online extremism: Research trends in internet activism, radicalization, and counter-strategies. *International Journal of Conflict and Violence*, 14(2), 1-20. <https://doi.org/10.4119/ijcv-3809>

Zajonc, R.B. (1968). Attitudinal effects of mere exposure. *Journal of Personality and Social Psychology*, 9(2, Pt.2), 1-28. <https://psycnet.apa.org/doi/10.1037/h0025848>

Zillmann, D., y Brosius, H. B. (2012). Exemplification theory of media influence. En J. Bryant, D. Zillmann y M. B. Oliver (Eds.), *Media effects: Advances in theory and research* (pp. 245-267). Routledge.

Zimbardo, P., Haney, C., Banks, W., y Jaffe, D. (1986). La psicología del encarcelamiento: privación, poder y patología. *Revista de psicología social*, 1(1), 95-105. <http://dx.doi.org/10.1080/02134748.1986.10821547>

Zimbardo, P. (2007). *The Lucifer effect: Understanding how good people turn evil*. Random House.

Zuiderveen Borgesius, F. J., Trilling, D., Möller, J., Bodó, B., de Vreese, C. H., y Helberger, N. (2016). Should we worry about filter bubbles? *Internet Policy Review*, 5(1). <https://doi.org/10.14763/2016.1.401>

Capítulo 4

MAPA DE LAS CAPACIDADES
DE INVESTIGACIÓN EN MATERIA
DE DESINFORMACIÓN EN LAS
UNIVERSIDADES Y CENTROS DE
INVESTIGACIÓN ESPAÑOLES

Coordinadores:

Leticia Rodríguez Fernández

Presidencia de Gobierno - Departamento de Seguridad Nacional (DSN)

Autores y colaboradores:

Juan Miguel Aguado Terrón

Ana Almansa Martínez

Sergio Arce García

Ignacio Blanco Alfonso

Leyre Burguera Ameave

Eva Campos Domínguez

Javier Cantón Correa

María José Establés Heras

David García Marín

Salvador Gómez García

Belén Puebla Martínez

María José Rementería Núñez

Jordi Rodríguez Virgili

Ramón Salaverría Aliaga

Astrid Wagner



INTRODUCCIÓN

La desinformación se ha convertido en una amenaza relevante que requiere de investigación especializada que nos permita entender su arquitectura, efectividad y capacidad de influencia y que contribuya en la generación de potenciales soluciones.

El siguiente capítulo “Mapa de las capacidades de investigación en materia de desinformación en las Universidades y centros de investigación españoles” nace como una propuesta para identificar los esfuerzos de las instituciones académicas españolas en el análisis y comprensión de la desinformación. Una propuesta promovida desde la vocalía de la Conferencia de Rectores de las Universidades Españolas (CRUE), representada por el profesor Antonio Díaz, en el Foro contra las Campañas de Desinformación en el ámbito de la Seguridad Nacional.

El principal objetivo de este capítulo es analizar la contribución de los investigadores españoles a la comprensión de las diferentes dimensiones que definen el fenómeno de la desinformación. Para ello, el trabajo traza cuatro objetivos específicos: el análisis de la producción investigadora publicada, la identificación de los grupos de investigación especializados en esta materia, el análisis de la financiación recibida, y la proyección estratégica de la aportación de la universidad de cara a una futura estrategia nacional contra las campañas de desinformación. Cada uno de los cuatro análisis para alcanzar este objetivo se corresponde con cada una de las cuatro secciones que configuran este capítulo.

Para trabajar en su consecución, el 15 de diciembre de 2022 se constituyó un grupo de expertos en el seno del Foro contra la Desinformación en el Ámbito de la Seguridad Nacional. Los 17 investigadores e investigadoras, pertenecen a 13 universidades y centros de investigación españoles y al Departamento de Seguridad Nacional de la Presidencia del Gobierno, y cuentan con una dilatada experiencia investigadora en los diferentes elementos que configuran el fenómeno de la desinformación. Junto al conocimiento experto de sus integrantes, en la composición de este grupo se cuidó de la existencia de una representatividad heterogénea de universidades y centros de investigación, con universidades tanto públicas como privadas, de distintas comunidades autónomas, incluyendo también universidades de modalidad a distancia. El grupo quedó constituido

por un total de 17 expertos, 9 hombres y 8 mujeres, que se describirán en el desarrollo del presente capítulo. En último lugar, en la composición del equipo de trabajo también se atendió a la paridad de género de sus componentes, así como a la participación de distintos perfiles, como investigadores postdoctorales, profesores contratados doctores, profesores titulares y catedráticos de universidad. El objetivo era contar con la mayor pluralidad de expertos y enfoques, de forma que enriquecieran el trabajo y los debates mantenidos en el seno del equipo de trabajo.

APROXIMACIÓN A LA INVESTIGACIÓN SOBRE DESINFORMACIÓN REALIZADA POR LOS ACADÉMICOS DE UNIVERSIDADES Y CENTROS DE INVESTIGACIÓN ESPAÑOLES

El primer objetivo de este trabajo es analizar la investigación que realizan los académicos con filiación en universidades y centros de investigación españoles en materia de desinformación. En concreto, dentro de este primer objetivo se busca conocer cuáles son los temas más investigados, en qué áreas de conocimiento se está trabajando este objeto de estudio y qué universidades españolas y centros de investigación presentan mayor especialización.

Para llevar a cabo estos análisis se realizó una búsqueda exhaustiva dentro del repositorio Web of Science (WOS) que da acceso a un amplio conjunto de textos académicos como artículos publicados en revistas científicas o capítulos de libros, entre otros materiales. Se ha decidido acotar el estudio a los artículos publicados en revistas científicas porque presentan unos procesos de revisión y evaluación más rigurosos que los de otras publicaciones, como revisión ciega por al menos dos evaluadores anónimos, garantizando por extensión una mayor calidad y relevancia científica de sus contenidos. Así, cuando un académico o un grupo de investigadores envían los resultados de un trabajo a una revista se inicia un proceso de varias fases. En primer lugar, el equipo editorial de la revista revisa el texto y valora si es de interés para su publicación. Si este es aceptado comienza la fase de revisión por pares en la que, manteniendo el anonimato tanto del autor como de los revisores, estos últimos determinan si el texto se publicará y ofrecen aquellas recomendaciones que consideran oportunas para ello. Estas recomendaciones pueden recoger cuestiones muy diversas que abarcan desde la propia investigación hasta la redacción del texto y que resultan en ocasiones determinantes para que el manuscrito pueda finalmente ser publicado.

Esta metodología garantiza la calidad de los trabajos publicados y aporta, a su vez, rigor editorial, pues las propias revistas también son categorizadas por su calidad en base a determinados índices e indicadores bibliométricos. Cuanto más relevante es la posición de la revista, más estrictos son sus procesos, obteniendo trabajos de mayor calidad y de mayor impacto.

Siguiendo con nuestro objetivo, se realizó una búsqueda en Web of Science, repositorio bibliográfico internacional para todas las áreas de conocimiento, de aquellos artículos publicados en revistas científicas por investigadores con filiación en universidades y centros de investigación españoles. La búsqueda se limitó a través de varias palabras clave,

relacionadas con el ámbito de la desinformación y acordadas por el equipo de trabajo¹. El rango de fechas de los resultados obtenidos abarca desde 1991, año del que se obtienen los primeros registros, hasta febrero de 2023, fecha en la que se realiza esta labor de recopilación. Esta búsqueda fue realizada por Sergio Arce García, profesor contratado doctor de la Universidad Internacional de La Rioja (UNIR) y se obtuvieron un total de 820 artículos que fueron tratados y analizados para su estudio mediante software estadístico R.

La información obtenida fue revisada y categorizada por un equipo de académicos especializados en desinformación. Participaron en este trabajo Ignacio Blanco Alfonso, catedrático en la Universidad CEU San Pablo; Ana Almansa Martínez, profesora titular en la Universidad de Málaga (UMA); Salvador Gómez García, profesor titular en la Universidad Complutense de Madrid (UCM); Belén Puebla Martínez, profesora titular en la Universidad Rey Juan Carlos (URJC); Leyre Burguera Ameave, profesora contratada doctora en la Universidad Nacional de Educación a Distancia (UNED) y David García Marín, profesor ayudante doctor en la URJC. Tras el análisis y extracción de conclusiones, la representación gráfica de la misma fue realizada por los investigadores Sergio Arce García y Javier Cantón Correa, investigador de la Universidad de Granada (UGR).

Análisis de los artículos científicos publicados sobre desinformación por investigadores e investigadoras en España

Los investigadores e investigadoras españoles han aportado gran cantidad de trabajo al análisis y comprensión del estudio de la desinformación. España figura como el tercer país con mayor producción de artículos científicos en este campo, registrando el 4.82% del total. Solo por delante se encontrarían los EE.UU. que se posiciona en primer lugar, con un 26.5% de artículos publicados en este campo, y Reino Unido, en segunda posición con el 8.58% de la producción (Gráfico 1).

Por detrás de España y en cuarto lugar se encontraría Australia (4,31%), seguida de Canadá (4,12%), China (4,02%), Alemania (3,67%), Italia (2,91%), India (2,55%), Brasil (2,37%) y Países Bajos (2,25%). El resto de países representados en el Gráfico 1 recogen menos del 2% de artículos científicos publicados sobre desinformación. Cabe destacar que hay un número de países que no llegan si quiera a un mínimo de representación, establecido en 0,66%, y que se han agrupado en la categoría “otros” que supone un 15% de los artículos publicados. En esta etiqueta se encontrarían Irlanda, Dinamarca, Colombia, Chile, México, Grecia, Argentina, Eslovenia, Ecuador, Hungría, entre otros.

¹ Palabras clave empleadas en la búsqueda: bulo, desinformación, disinformation, fact-check, fake news, misinformation, noticias falsas, post truth, posverdad, false news, malinformation, guerra híbrida e hybrid warfare.

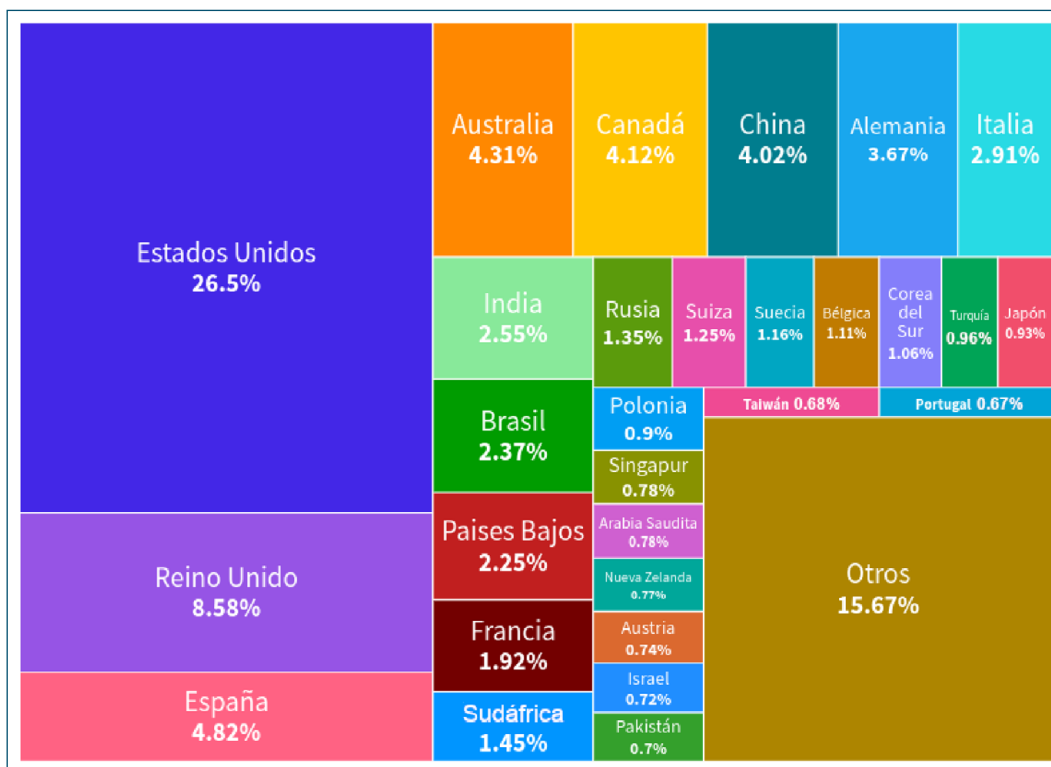


Gráfico 1: Producción de artículos científicos sobre desinformación por países. Fuente: elaboración propia

Cabe destacar que, en este primer análisis se han utilizado los datos obtenidos inicialmente, sin filtrar, cuyos resultados para España fue de 820 artículos científicos en el ámbito de la desinformación.

Una vez realizado el análisis y depuración manual, estos resultados se redujeron a 615 artículos, obteniéndose 187 falsos positivos, lo que supone un 25% de la muestra. Estos artículos fueron eliminados del análisis, bien porque no tenían relación directa con la desinformación o porque su uso no correspondía con lo que entendemos como desinformación. En este último caso, esta situación suele producirse por traducciones inexactas de las palabras *mal-information*, *disinformation* y *misinformation* que en inglés presentan muchos matices que no se corresponden con la traducción común que en español tenemos de “desinformación”.²

² Wardle y Derakhshan (2017) diferenciaban la desinformación (*disinformation*), información errónea (*misinformation*) e información maliciosa (*malinformation*). Esta diferenciación se pierde por la traducción al castellano de “desinformación” como término que aglutina los 3 significados.

Entre 1991 y 2017 solo encontramos un total de 34 artículos publicados (5,5 % del total), siendo una cifra muy reducida en comparación con los 27 artículos que se publican solo en 2018, los 89 de 2019, o los 171 que se llegan a alcanzar en 2021. Es, por tanto, a partir de 2018 cuando encontramos el intenso punto de inflexión que lleva a que la mayor parte de esta producción científica se concentre entre los años 2019 y 2022, representando un 88,7% del total.

En cuanto a las áreas de conocimiento (ver gráfico 2), y ya sobre datos completamente depurados, se observa cómo los investigadores en comunicación son quienes concentran mayor cantidad de artículos publicados, el 37,39% de la producción total analizada (230 artículos), confirmando hallazgos de anteriores revisiones sistemáticas (Blanco, García y Tejedor, 2019). La siguiente área de conocimiento, en términos de producción, es Biblioteconomía y Ciencias de la Información con el 11,34% y Ciencias de la Educación, en tercer lugar, con el 5,91%. Nos sorprende, al realizar este análisis, encontrar tan poca producción desde los campos de la sociología y la psicología, áreas fundamentales para entender los efectos e impactos de la desinformación.

Las áreas de conocimiento más proliferas coinciden con las áreas de las revistas en las que se recogen más publicaciones. En primera posición aparece *El Profesional de la Información* con un total de 56 artículos publicados, seguida por *Revista Latina de Comunicación Social*, con 26. Ambas forman parte del corpus de revistas científicas españolas de primer nivel en el campo de la comunicación. En tercer lugar, encontramos *International Journal of Environmental Research and Public Health*, con 16 artículos, *Communication & Society* con 16 y *Revista Española de Comunicación en Salud* con 15.

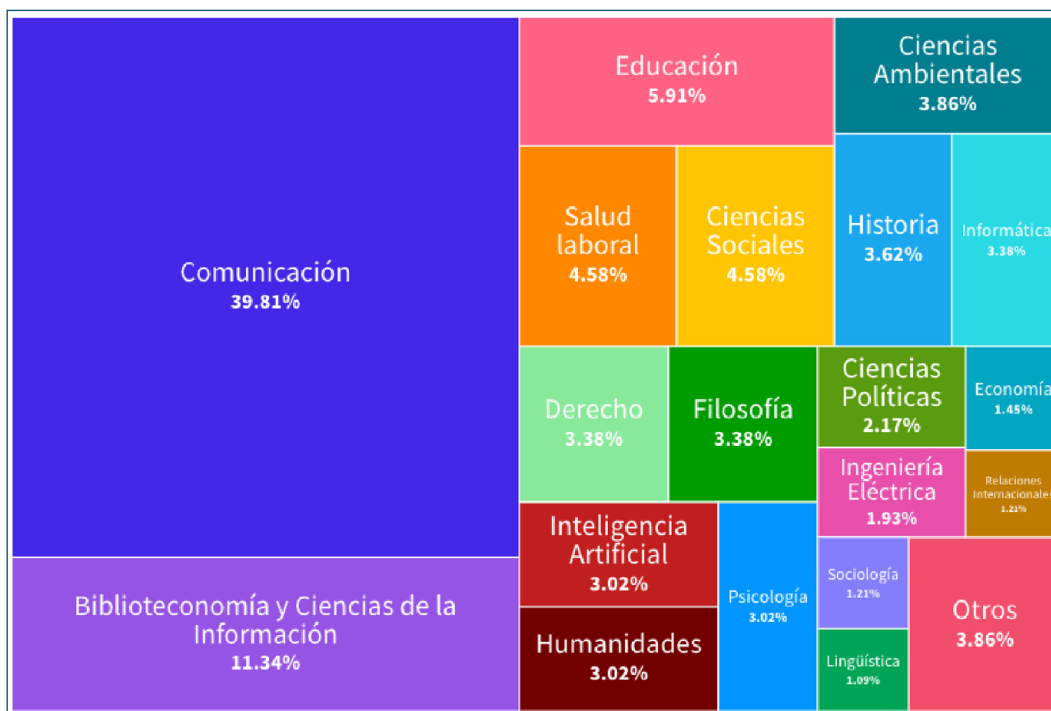


Gráfico 2: Producción de artículos científicos sobre desinformación en España por áreas de conocimiento. Fuente: elaboración propia

Sobre el análisis de estos datos obtenemos dos conclusiones de valor. Por un lado, se observa que los trabajos científicos generados por investigadores en España tienden a publicarse también en revistas de nuestro país. Hay una concentración de conocimiento que no se exporta a revistas extranjeras. De las 20 publicaciones que aparecen con mayor cantidad de artículos publicados sobre desinformación, 15 de ellas son españolas: El Profesional de la Información, *Revista Latina de Comunicación Social*, *Communication & Society*, *Revista Española de Comunicación en Salud*, *Historia y Comunicación Social*, *Vivat Academia*, *Adcomunica*, *Prisma Social*, *Comunicar*, *Comunicación y Hombre*, *Journal of Learning Style*, *Análisi-Quaderns de Comunicació I Cultura*, *Revista Mediterránea de Comunicación*, *Revista Icono14* y *Revista de Derecho Político*. En total concentran el 33,6% de los artículos publicados (207 artículos).

Al elemento cuantitativo ha de añadirse el impacto de estas publicaciones en la comunidad científica, pues 107 de estos artículos han sido publicados en revista indexadas en el *Journal Citation Report* (JCR) y 226 en revistas indexadas en Scopus, dos de los rankings mundiales más relevantes para la medición del impacto de publicaciones científicas.

Cabe destacar que, a pesar de que la mayoría de investigaciones se publiquen en revistas españolas, el idioma predominante en estos artículos es el inglés (55,7%) frente a un 42,4% en español. Se subraya, por tanto, una intención de internacionalización de los resultados obtenidos por investigadores de filiación española.

El tercer objetivo de este epígrafe era categorizar los artículos publicados en base a su temática. Para llevar a cabo este análisis se establecieron dos niveles de etiquetas (tema principal y tema secundario).

El tema de investigación más abordado han sido las redes sociales (Gráfico 3). Su indudable papel como herramientas de difusión de la desinformación supone un importante reto para la comunidad académica que puede ser abordado desde múltiples enfoques, lo que explica esta primera posición. En segundo lugar, encontramos la COVID-19. La pandemia podría haber contribuido a avivar el interés por el fenómeno al ser un caldo de cultivo ideal para esparcir noticias falsas, fragmentar sociedades y horadar el poder y legitimidad de las instituciones. Es precisamente en 2021, el año posterior a la pandemia, cuando se recoge mayor número de contribuciones (171) siendo, precisamente 49 de ellas relativas al impacto de la COVID-19.

En tercer y cuarto lugar encontramos dos temáticas vinculadas directamente con el ámbito de la comunicación, las rutinas profesionales y la verificación (*fact-checking*), seguidas por estudios centrados en las fake news, la alfabetización mediática y digital y la posverdad. Diversas aportaciones abordan las formas, la retórica y el discurso de la desinformación, especialmente desde el estudio del *clickbait*, las narrativas populistas y los relatos negacionistas.

Se identifica una carencia en investigaciones sobre inteligencia artificial y aprendizaje automático, aún incipientes, así como de trabajos que aborden la desinformación local, ahondando en las características de este tipo de campañas según el ámbito geográfico.

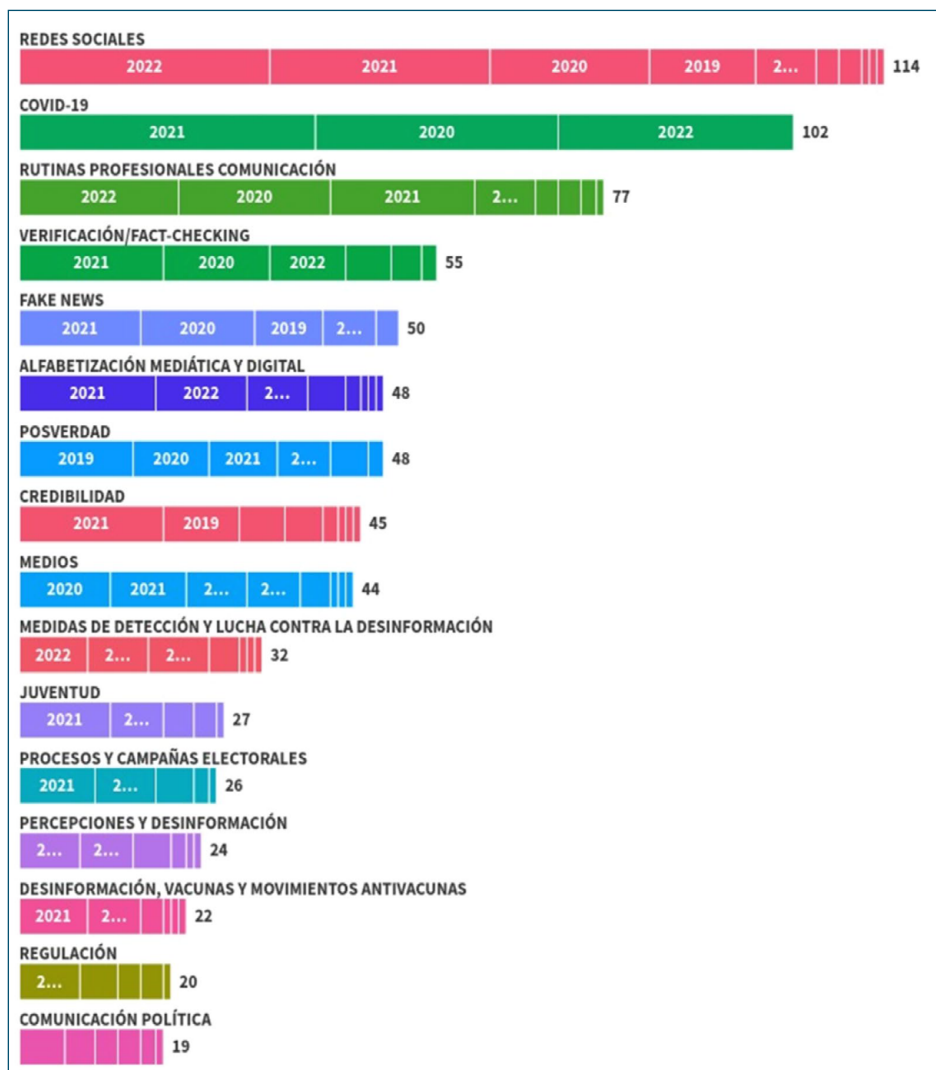


Gráfico 3: Temáticas más abordadas en los artículos científicos publicados sobre desinformación por investigadores vinculados a Universidades y centros de investigación españoles. Fuente: elaboración propia

En cuanto a la filiación de los investigadores españoles (gráfico 4), encontramos en primer lugar a la Universidad Complutense de Madrid, con 42 publicaciones, la Universidad Carlos III, con 36, y la Universidad de Valencia, con 33. Le siguen la Universidad Rey Juan Carlos (32), la Universidad de Sevilla (28), la Universidad de Málaga (27), la Universitat Pompeu Fabra (24), la Universidad de Salamanca (22) y la Universidad de Santiago de Compostela (21).

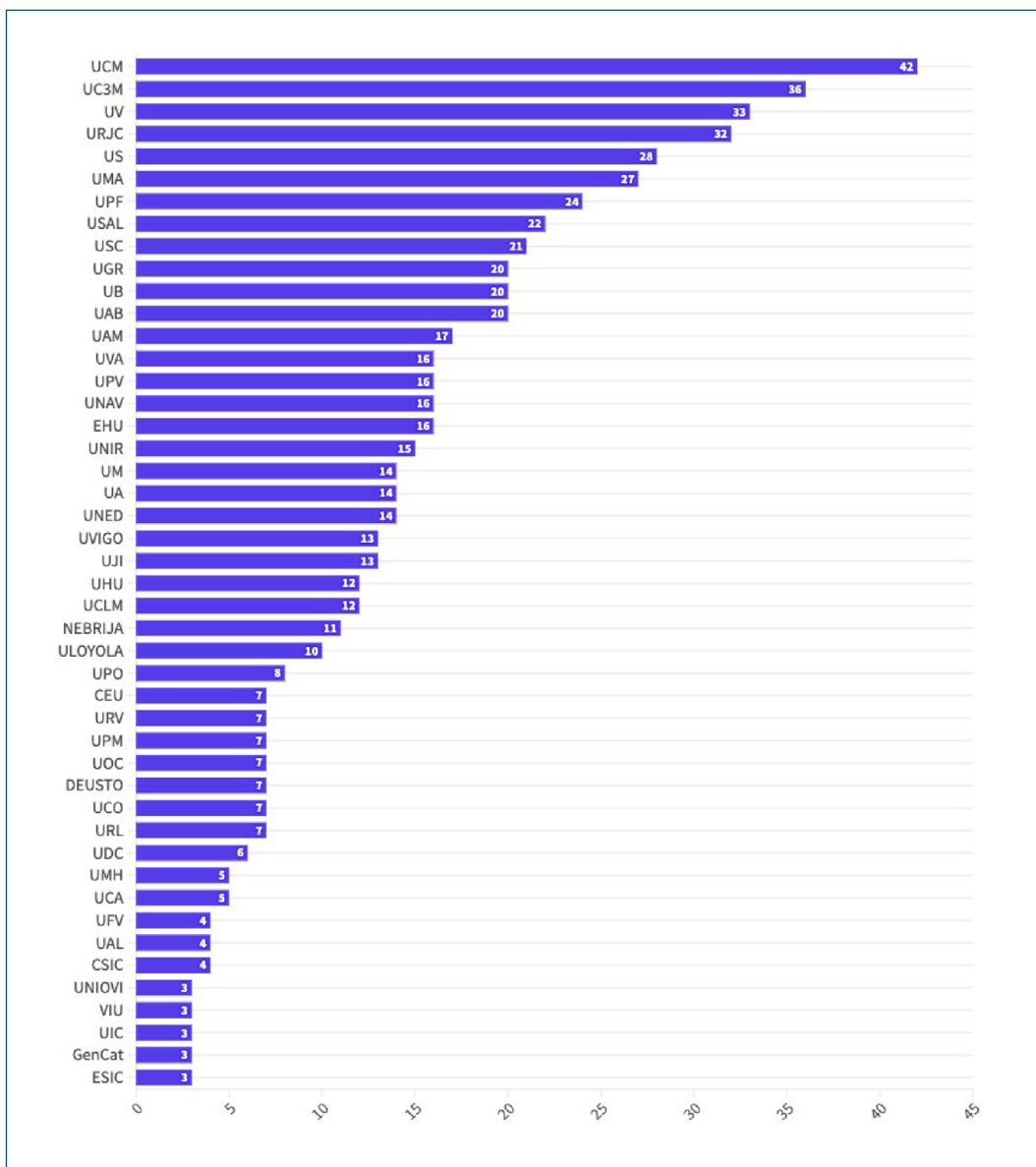


Gráfico 4: Filiación de los autores firmantes de las publicaciones analizadas. Fuente: elaboración propia

Cabe destacar, por otra parte, que la colaboración entre investigadores de diferentes universidades es bastante reducida para el amplio volumen de artículos registrados. Igualmente, la colaboración internacional es bastante escasa y casi inexistente con países latinoamericanos.

Análisis y evolución de los artículos científicos publicados por investigadores de instituciones españolas

Aunque la primera publicación que hemos identificado es de 1991, no encontramos más artículos científicos hasta el año 2005. Incluso, a partir de ese año y hasta 2014, el volumen de publicaciones es realmente bajo, con una media de entre 1 y 4 artículos científicos anuales sobre desinformación.

En esta primera etapa (1991-2014), no se observa que la desinformación sea estudiada como un fenómeno propiamente relevante, si bien algunas publicaciones comienzan a abordar temáticas como el falso consenso (Wojcieszak, 2011), el papel de los medios de comunicación (De Pablos, 2008) o su relación con cuestiones vinculadas a la salud (Del Fresno y López, 2014) y la psicología (Paz-Alonso et al., 2013). En conclusión, esta etapa se caracteriza por una gran dispersión en los temas abordados en los que la desinformación aparece como una variable secundaria.

A partir de 2015 aparecen los primeros artículos científicos que estudian la desinformación como un fenómeno contextualizado en las redes sociales (Sánchez-Casado et al., 2015), así como los primeros estudios sobre alfabetización mediática y digital (Aguaded Gómez y Romero Rodríguez, 2015). Pero no será hasta 2017 cuando se recojan en los títulos de las publicaciones términos como *infoxicación* (Benaissa, 2017) o posverdad (Niño et al., 2017). Coincide esta circunstancia con una popularización en el uso de los términos *fake news* y posverdad, especialmente, tras la elección de Donald Trump como Presidente de Estados Unidos. Popularidad que promueve que, en 2017, el diccionario Collins escogiera posverdad como palabra del año.

De manera orgánica, la academia trata de dar respuesta a esta nueva preocupación social y se produce un auge de publicaciones entre los años 2019 y 2022. Este incremento coincide, por un lado, con el progresivo interés que este fenómeno ha ido adquiriendo desde los eventos políticos de 2016, pero, sobre todo, con la crisis sanitaria e informativa provocada por la COVID-19. De hecho, en 2022 se publican menos artículos sobre la materia que durante 2021, año de máxima productividad científica provocada por la infodemia.

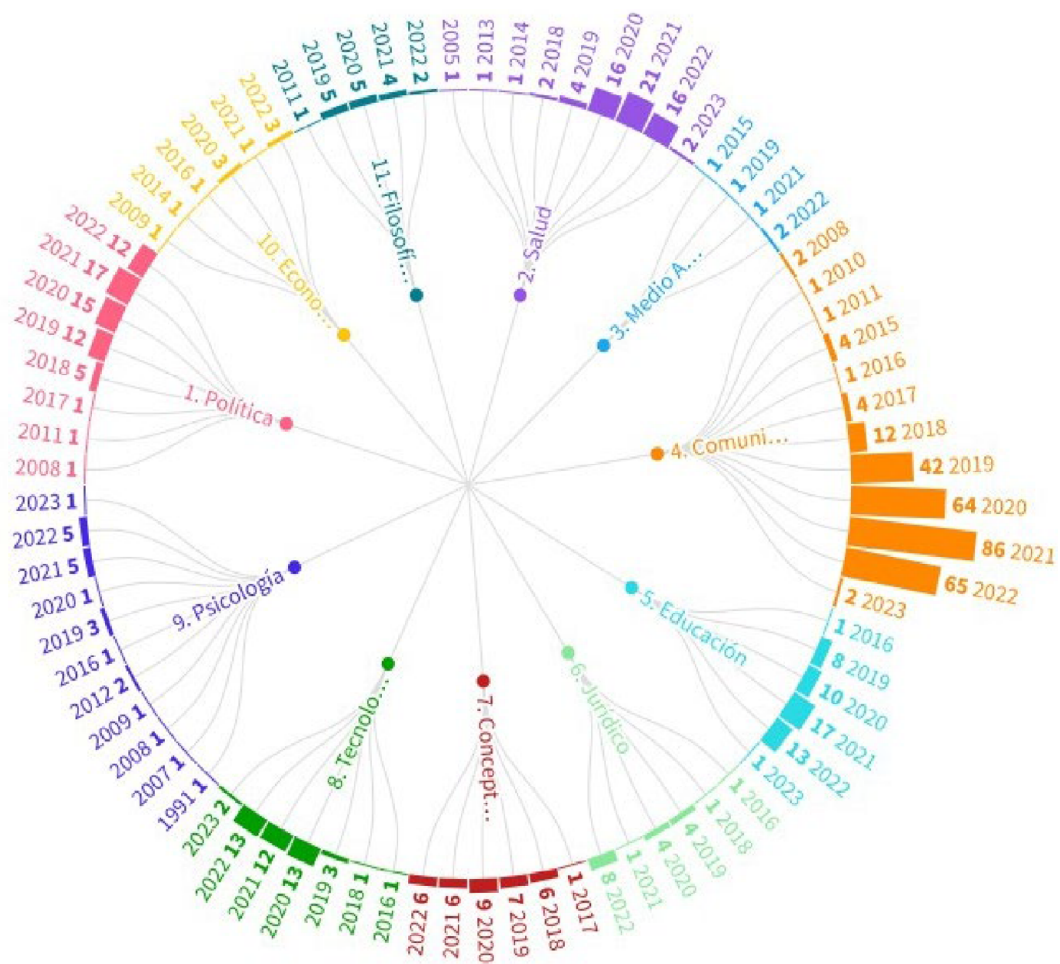


Gráfico 5: Distribución de los artículos científicos sobre desinformación por área de conocimiento y fecha de publicación. Fuente: elaboración propia

Durante este periodo (2019-2022), la temática más prevalente es la relativa a la “Comunicación”, donde destacan los estudios vinculados con la respuesta periodística a la desinformación y la COVID-19 (gráfico 5). Los enfoques son variados, recogiendo trabajos relacionados con la realidad profesional, que abordan las rutinas periodísticas o las percepciones de los profesionales de los medios de comunicación sobre la desinformación; así como publicaciones con enfoque social, que se centran en las percepciones de la ciudadanía hacia los medios o de la población hacia las redes sociales como fuentes de información. Varias de las publicaciones abordan las formas, la retórica y el discurso de la desinformación, especialmente desde el estudio del *clickbait*, las narrativas populistas y los relatos negacionistas.

La segunda categoría temática dominante es la “Política”, cuyas investigaciones se centran en estudios de casos concretos como los periodos electorales en Estados Unidos (2016) y España (2019).

Como se ha avanzado anteriormente, el periodo 2019-2022 se enmarca en los momentos más duros de la pandemia de la COVID-19 y en los trabajos específicos sobre “Salud” resultan relevantes en términos cuantitativos. En concreto, las publicaciones se centran en la desinformación sobre las vacunas, los elementos activadores de la credibilidad del contenido desinformativo sobre la enfermedad, las rutinas periodísticas y el papel de los medios de comunicación en el tratamiento informativo de la crisis sanitaria, los discursos, estrategias y tipología de la desinformación vinculada a la COVID-19, la propagación de la desinformación sobre este caso en redes sociales, la percepción de la ciudadanía sobre la cobertura informativa de la pandemia y las teorías de la conspiración.

Asimismo, son especialmente interesantes las investigaciones situadas en la categoría denominada “Tecnología”, centradas casi exclusivamente en la aplicación de soluciones automatizadas para la detección de la desinformación y que redundarían en un aumento de la efectividad del *fact-checking*. Se agrupan básicamente en dos marcos temáticos: en primer lugar, sistemas algorítmicos de detección del contenido desinformativo a través de soluciones avanzadas de inteligencia artificial y, en segundo lugar, sistemas algorítmicos para la verificación de fuentes online. Se percibe una tendencia generalizada hacia la publicación de este tipo de trabajos en revistas científicas extranjeras. Sus autores suelen ser expertos en ciencia de datos. De forma muy poco frecuente, estas investigaciones son firmadas por equipos multidisciplinares con participación de expertos del área de Comunicación. Como se ampliará más adelante, gran parte de este tipo de trabajos se centran en ofrecer soluciones concretas al desafío de la desinformación desde la implementación de sistemas de detección del contenido falso o la identificación de fuentes maliciosas en las redes sociales. Por otra parte, se identifica una preocupación latente (aunque aún con poca intensidad en el número de trabajos) hacia el problema de los *deepfakes*, etiqueta que registra a aquellos estudios dedicados tanto a su conceptualización como a identificar y medir sus efectos.

La “Educación” constituye la quinta categoría temática más frecuente en este periodo. Las aproximaciones desde esta óptica son muy variadas, si bien prevalece el estudio de la alfabetización informacional en la población juvenil. Estas publicaciones analizan también el impacto del nivel educativo en la credibilidad del contenido desinformativo y en los sesgos cognitivos que activan la confianza en las noticias falsas. Se estudian, asimismo, programas de educación mediática desde diferentes perspectivas tales como la *gamificación* y la alfabetización visual.

Aunque con una menor presencia en términos cuantitativos, destacan los estudios que intentan conceptualizar algunos de los términos fundamentales vinculados al fenómeno. Estos trabajos teóricos se centran en las siguientes dimensiones relacionadas con nuestro objeto de estudio: emociones y desinformación, *fake news*, posverdad, democracia,

credibilidad y debate público. Finalmente, casi la totalidad de los trabajos sobre “Psicología” se centran en los sesgos cognitivos desde ópticas generales o vinculadas a casos concretos como la crisis de la COVID-19, las elecciones generales en España (2019) o la alfabetización mediática.

Estos datos ponen de manifiesto que, desde la irrupción social en 2016 de las *fake news* y otros desórdenes informativos, la Academia ha trabajado arduamente en el análisis y descripción del fenómeno en todas sus dimensiones, y lo ha hecho desde una aproximación multidisciplinar, aunque con el liderazgo de los estudios de Comunicación.

ANÁLISIS DE LOS GRUPOS DE INVESTIGACIÓN ESPECIALIZADOS EN DESINFORMACIÓN EN ESPAÑA

La investigación requiere de la unión de sinergias y en este proceso se configuran grupos de trabajo que suman los esfuerzos y conocimientos con fines comunes. Estos grupos pueden estar formados por investigadores con perfiles de una misma área de conocimiento o de áreas heterogéneas y, normalmente, se adscriben a un centro de investigación o a la facultad de una universidad.

Nuestra intención es conocer qué grupos de investigación tienen la desinformación entre sus líneas de trabajo. Para ello se ha realizado una revisión de la información pública disponible en las páginas web de las universidades y centros de investigación del Consejo Superior de Investigaciones Científicas (CSIC), atendiendo a sus líneas de investigación y a sus proyectos financiados. Esta revisión se realizó manualmente, recopilando toda la información disponible en las páginas web de las universidades y centros de investigación españoles para su posterior análisis.

La recopilación y el posterior análisis fueron realizados por Astrid Wagner, científica titular del Instituto de Filosofía del CSIC en Madrid e investigadora asociada del Berlin Center for Knowledge Research; Juan Miguel Aguado Terrón, catedrático en la Universidad de Murcia; Ramón Salaverría Aliaga, catedrático en la Universidad de Navarra y Jordi Rodríguez Virgili, profesor titular en la Universidad de Navarra.

Grupos de investigación que abordan la desinformación en el Consejo Superior de Investigaciones Científicas

El Consejo Superior de Investigaciones Científicas (CSIC) es una Agencia Estatal compuesta por 121 institutos de investigación y 3 centros nacionales. Fundado en 1939, es el organismo de investigación con mayor potencial multidisciplinar, albergando más de 11.000 trabajadores en casi todas las comunidades autónomas del territorio español. Esta capacidad lo convierte en líder en producción científica con una media anual de 13.000 publicaciones en revistas científicas.

Para realizar este análisis, se han revisado los 78 grupos de investigación adscritos al área científica de Humanidades y los 91 grupos registrados en Ciencias y Tecnologías físicas. Se han descartado el resto de áreas de conocimiento al no tener suficiente afinidad con el campo de estudio de la desinformación.

El análisis identificó cuatro grupos de investigación que, entre sus líneas de trabajo o entre sus proyectos de investigación, presentan una alta especialización en el estudio de la desinformación. Dos de estos grupos forman parte del Instituto de Filosofía (IFS) ubicado en Madrid: el Grupo de Ética Aplicada (GEA) y Theoria cum Praxi: Ética, Epistemología y Sociedad (TcP). Estos dos grupos desarrollan una perspectiva ética y epistemológica y cuentan con proyectos de investigación relacionados con la comunicación, los entornos digitales, la desinformación o la conspiranoia, entre otros. Entre estos proyectos encontramos el plan nacional “INconRES Incertidumbre, confianza y responsabilidad. Claves ético-epistemológicas de las nuevas dinámicas sociales en la era digital”.

Continuando con la ética, pero en este caso desde la relación con la Inteligencia Artificial (IA), encontramos al grupo Sistemas Multiagente del Instituto de Investigación en Inteligencia Artificial (IIIA) ubicado en Barcelona. Sus líneas de trabajo abarcan los desafíos éticos de la IA y las comunidades digitales, centrandó su trabajo en el diseño de sistemas informáticos basados en valores.

Desde el enfoque de la ciberseguridad se identifica al Grupo de Investigación en Ciberseguridad y Protección de la Privacidad (GiCP) adscrito al Instituto de Tecnologías Físicas y de la Información “Leonardo Torres Quevedo” (ITEFI), ubicado en Madrid. Entre sus líneas de trabajo se encuentra la aplicación de la IA para la detección de desinformación o los protocolos, mecanismos y tecnologías pre y poscuánticas para la ciberseguridad y la privacidad, con una considerable producción científica.

El INGENIO, Instituto de Gestión de la Innovación del Conocimiento (CSIC-UPV) ubicado en Valencia que no constituye grupos, desarrolla una línea de investigación dedicada a mejorar la calidad de la comunicación científica en redes sociales para contrastar y evitar la desinformación. Entre sus proyectos, se encuentra “SciCoMetrics: Desentrañando la calidad de la comunicación científica en las noticias y los medios sociales: hacia una nueva taxonomía de métricas de las interacciones ciencia-sociedad” de financiación nacional.

Cabe señalar también el proyecto “Diseño institucional y político para mejorar el bienestar social (INPODE)” del Institut d’Anàlisi Econòmica (IAE-CSIC), que aborda, entre otras cuestiones, la influencia extranjera en las políticas nacionales y el papel de los líderes en la magnificación de la polarización de las opiniones de la sociedad a través de las redes sociales.

Por otra parte, en la realización del análisis se recogen grupos que, sin presentar líneas de investigación o proyectos propios relacionados con la desinformación, colaboran con otros que la abordan. Es el caso de diversos grupos, ubicados en el Centro de Ciencias Humanas y Sociales en Madrid: Evaluación y Transferencia Científica (ETC), Ciencia, Tecnología y Sociedad (CTS), Análisis y Prospectiva Científica (APROC), y Ciudadanos e Instituciones (CIP).

Grupos de investigación que abordan la desinformación en las universidades españolas

Crue Universidades Españolas (CRUE) es una asociación sin ánimo de lucro constituida en el año 1994 y es el principal interlocutor de las universidades con el Gobierno central. Forman parte de su organización 76 universidades (50 públicas y 26 privadas) y tiene entre sus objetivos fomentar las relaciones con el tejido productivo y social, fomentar las relaciones institucionales, tanto nacionales como internacionales, y poner en valor a la Universidad española (CRUE, s.f).

Para el desarrollo de este trabajo se han revisado las páginas web de las 76 universidades que forman parte de la CRUE atendiendo tanto a sus líneas de investigación como a los proyectos relacionados con la desinformación. Se describe a continuación los resultados por comunidades autónomas:

Andalucía

Siete de las once universidades andaluzas presentan líneas o proyectos específicos relacionados con la desinformación.

Comenzando por la Universidad de Jaén, el Centro de Estudios Avanzados en Tecnologías de la Información y de la Comunicación lidera los proyectos nacionales “Evaluación de la toxicidad en medios digitales”, “Tecnologías del lenguaje humano para entidades digitales vivas” y “Moderación de contenidos en redes sociales usando tecnologías del lenguaje”. Estos proyectos presentan un carácter técnico cuya fundamentación es la aplicación de la tecnología para la detección de desinformación.

Por su parte, la Universidad Loyola cuenta con el grupo POSTICOM: Comunicación positiva y cultura digital, responsable de los proyectos de financiación europea “FAKESpotting” y “SPOTTED” (School Policies to Tackle Fake News), cuyo objetivo es mejorar la alfabetización mediática de estudiantes y profesores de Educación Secundaria.

En la Universidad de Málaga, se encuentra el grupo MEDIO: Media & Data Innovation, que lidera el proyecto de I+D “El impacto de la desinformación en el periodismo: contenidos, rutinas profesionales y audiencias”. Entre sus objetivos se recoge ofrecer recursos, derivados de la investigación, para conocer y analizar el estado de la desinformación. En la misma universidad también se identifica el proyecto nacional “El uso informativo de las redes sociales por parte de los jóvenes españoles: condicionantes tecnológicos, credibilidad de las noticias y consumo incidental de contenidos”, del área de Periodismo.

Desde la Universidad de Sevilla, el Grupo Big Data and Business Intelligence in Social Media dirige el proyecto nacional “Aplicación de redes generativas antagónicas para combatir la

manipulación de clientes *online* (react)”, especializado en identificar opiniones falsas en entornos digitales. También desde el grupo Sistemas Informáticos se desarrolla el proyecto “Detección y generación automática de falacias mediante el uso de modelos de lenguaje masivos basados en aprendizaje profundo”, que busca la detección de desinformación.

La Universidad de Granada cuenta con varios proyectos que hemos recogido en el análisis que se ofrecerá en el siguiente epígrafe, si bien no hemos encontrado información sobre los grupos correspondientes. Se trataría de “Las noticias falsas en las redes sociales. Tres estudios de caso: populismo, COVID y cambio climático” de la Facultad de Filosofía y Letras; “Monitorización de la desinformación y su impacto social mediante inteligencia artificial: aplicación a la seguridad de la sociedad”, del Departamento de Ciencias de la Computación e Inteligencia Artificial, y “La gobernanza de la inteligencia artificial basada en la ciudadanía”, de la Facultad de Derecho.

Por otra parte, el grupo Razonamiento aproximado e inteligencia artificial, del área de conocimiento Ciencias de la Computación e Inteligencia Artificial, forma parte del Observatorio Europeo de los Medios Digitales (EDMO), proyecto que apoya a la comunidad independiente que trabaja para combatir la desinformación y al que se adscriben varias universidades y verificadores de información. Se añade la Cátedra W.G.Leibniz, del área de Filosofía, que tiene entre sus líneas de investigación la posverdad con un proyecto financiado en este campo.

En la Universidad de Cádiz se identifica el proyecto nacional “Dinámicas colectivas de contagio de opiniones sobre salud: la infodemia de la COVID-19 y sus efectos sobre los procesos de toma de decisión”, que no está adscrito a ningún grupo en concreto, pero cuyo investigador principal forma parte de esta institución.

La Universidad de Huelva, cuenta con dos proyectos relacionados con la desinformación: “*Youtubers e Instagramers*: la competencia mediática en los prosumidores emergentes”, dirigido por un grupo de investigadores mixto de distintas universidades, y “Teorías de la conspiración y discursos de odio en línea: comparación de pautas en las narrativas y en las redes sobre COVID-19, inmigrantes, refugiados y personas LGTBI”, en el que están integrados investigadores del grupo de investigación “Estudios Sociales E Intervención Social” y COIDESO, de las áreas de Sociología, Psicología Social y Salud.

Aragón

En las universidades de Aragón se han identificado dos grupos que recogen la desinformación entre sus líneas de investigación. Se trata de Comunicación, periodismo, política y ciudadanía, de la Universidad de San Jorge, y el GICID (Grupo de Investigación en Comunicación e Información Digital), de la Universidad de Zaragoza. El primero se focaliza en la desinformación, los retos para la profesión periodística y el periodismo narrativo; mientras que el segundo lo hace desde las competencias digitales. Ninguno de los dos grupos recoge en sus páginas web proyectos financiados sobre desinformación.

Cantabria

MILET: Methods and Innovation for Learning and Teaching es el único grupo registrado en la Universidad de Cantabria que aborda la desinformación. Adscrito al área de Educación, ha desarrollado dos proyectos: “Competencia mediática en la sociedad digital para la participación crítica de la ciudadanía. Diagnóstico y planes de acción didácticos”, financiado por la Universidad de Cantabria, y “Pensamiento crítico de las familias ante el consumo digital y el ocio compartido”, financiado por el Gobierno de Cantabria.

Castilla y León

Las universidades de Castilla y León albergan, de manera principal, grupos dedicados a la alfabetización mediática y digital, así como al estudio de las implicaciones de la desinformación en el ámbito político.

En la Universidad de Salamanca, concretamente en el área de Derecho, identificamos al grupo Democracy Research Unit, que lidera el proyecto de financiación nacional “Desinformación, odio y polarización: la afectación de derechos y libertades de personas vulnerables”. Desde el área de Educación, el Grupo de Investigación en Inclusión, Innovación, Formación Docente y Alfabetización Informativa (INFORAL) puso en marcha el proyecto “Alfabetización en datos en el contexto universitario: detección de necesidades, diseño de escenarios formativos en abierto y elaboración de un referencial de competencias”.

En esta misma universidad y desde el Observatorio de los Contenidos Audiovisuales (OCA), encontramos el proyecto “Desarrollo y evaluación de un prototipo de detección automática de noticias falsas online”, financiado por la Fundación General de la Universidad de Salamanca, así como otros programas especializados en discurso de odio, como “Prevenir el odio contra los refugiados y los migrantes”, financiado por la Comisión Europea (CE). El grupo es responsable además de la cátedra “Niñ@s, Jóvenes y Medios”, que investiga el uso y consumo que hacen niños y jóvenes de los medios de comunicación, de interés desde el enfoque de la alfabetización mediática y digital.

Se añade la colaboración de la Universidad de Salamanca con Eurostar Media Group, empresa especializada en la creación de contenidos y en su monetización en todas las plataformas multimedia. Ambos participan en el proyecto “TRUESTORIES: TRUStworthy artificial intElligence over NPL to fight againST disinfORmation InstrumEnts in fiction”. En este caso, la dirección del trabajo recae en el grupo Bioinformática, Sistemas Inteligentes y Tecnología Educativa (BISITE), que cuenta además con un laboratorio adscrito “AIR Institute”, responsable del plan nacional “Procesamiento ético y explicable del lenguaje natural y redes neuronales gráficas para la detección de noticias falsas y la prevención de campañas de desinformación (Ethical News)”.

En la Universidad de Valladolid se identifican, de manera clara, dos tendencias de trabajo: la alfabetización mediática y digital y el *politainment*, término que nace de la suma de política y entretenimiento, en inglés. Esta última línea es la que desarrolla el Grupo de Investigación en Nuevas Tendencias en Comunicación (NUTEKO), responsable del plan nacional “Politainment ante la fragmentación mediática: desintermediación, engagement y polarización”, continuación de un plan nacional anterior, en el que se aborda, entre otras cuestiones, la presencia y uso de contenidos automatizados y algorítmicos por parte de medios, partidos políticos y plataformas de distribución de contenidos.

Desde el Grupo de Investigación en Comunicación Audiovisual e Hipermedia (GICAVH) lideran el proyecto “Verdad y ética en las redes sociales. Percepciones e influencias educativas en jóvenes usuarios de Twitter, Instagram y Youtube”, que analiza la presencia e importancia otorgada a la verdad y los valores éticos asociados a la ciudadanía digital en los nuevos medios. Con una línea de trabajo similar, encontramos el grupo Alfabetización mediática contra la desinformación (ALFA), proyecto de innovación docente del área de comunicación audiovisual y publicidad, con enfoque investigador e iniciativas formativas.

Señalar que, en la Universidad de Valladolid, también se ubica el Observatorio del Ocio y el Entretenimiento Digital, OCENDI, que trabaja en coordinación en varios proyectos con el grupo Mediaflows de la Universidad de Valencia (UV), que describiremos posteriormente.

Asimismo, en esta institución encontramos a la investigadora principal del proyecto “Parlamento, bots y desinformación: ecología de estrategias y prácticas en redes sociales”, financiado por la Fundación BBVA en 2020, cuyo objetivo es generar conocimiento sobre la comunicación parlamentaria digital en el contexto de la desinformación.

Por su parte, El Center for the Governance of Change de IE University lidera el proyecto “European Tech Insights”, que mapea las actitudes europeas hacia el cambio tecnológico y su gobernanza y en el que se aborda, entre otras cuestiones, la desinformación.

Cataluña

Las universidades catalanas presentan numerosos proyectos relacionados con la desinformación.

La Universidad Autónoma de Barcelona (UAB) cuenta con el Grup de Recerca Internacional d’Estudis sobre Comunicació i Cultura, que ha dirigido el proyecto “La comunidad china ante el discurso de odio durante la pandemia de la COVID-19”, financiado por el Instituto Catalán Internacional por la Paz (ICIP). Se añade el grupo Gabinete de Comunicación y Educación (GCE), que colabora en el proyecto “IVERES: identificación, verificación y respuesta”, liderado por Radio Televisión Española (RTVE), centrado en la creación de un sistema de verificación de información en lengua española contra la desinformación interesada. Además, en la UAB se adscribe un Laboratorio de Prospectiva e Investigación

en Comunicación, Cultura y Cooperación (LAPREC), grupo de investigación consolidado de la Generalitat de Cataluña que dispone de publicaciones sobre desinformación.

En la Universidad de Barcelona se registra el grupo de investigación CLiC - Centre de Llenguatge i Computació, responsable del proyecto nacional “Parámetros y estrategias para incrementar la relevancia de los medios y la comunicación digital en la sociedad: curación, visualización y visibilidad” y de “Desinformación y agresividad en Social Media: Analizando el lenguaje”, ya finalizado. En activo, pero en otras áreas de conocimiento, se recogen los proyectos nacionales “Analizar y combatir la incitación al odio en línea y la discriminación por razón de sexo desde una perspectiva interseccional”, del grupo GENI; y “La educación mediática y la dieta informativa como indicadores de la capacidad de análisis crítico de contenidos informativos en futuros docentes (Medi4teach)”, dirigido por el Institut de Recerca en Educació.

También en la misma universidad, el Grupo de investigación sobre lectura, escritura y adquisición de conocimientos, LEAC • UB, lidera el trabajo “Análisis de la competencia de pensamiento crítico en la era de la posverdad de los estudiantes del Máster de Formación del Profesorado de Secundaria”, financiado por IDP-ICE Universitat de Barcelona. El grupo desarrolló anteriormente el plan nacional “Uso crítico de la información en estudiantes de Secundaria: lectura y escritura de textos argumentativos”, evidenciando una amplia experiencia en la alfabetización mediática.

Con una alta especialización, el grupo Communication Networks & Social Change de la Universitat Oberta de Catalunya (UOC) lidera el proyecto “DISSIMILAR. Detección de noticias falsas en plataformas de redes sociales”, que tiene como finalidad proporcionar a las creadoras de contenido herramientas para marcar sus creaciones y hacer que cualquier modificación sea fácilmente detectable. El grupo colabora, además, en el proyecto “DataPolitik. Ciencia de datos contra la desinformación”, de Heurística, observatorio de datos y política, que estudia la opinión pública y el sentimiento político.

Por otra parte, el grupo GAME - CNM (Aprendizajes, Medios de comunicación y Entretenimiento; Comunicación y Nuevos Medios), también de la UOC, recoge un interesante proyecto de alfabetización mediática y digital. Se trata de EsDigital: Educación social digital. No obstante, en su descripción no se hace alusión específica a la desinformación. Esta línea de investigación sí se aborda, sin embargo, desde el grupo MEDIACCIONS. Critical studies in culture, design and networked communication, con proyectos como “Dissenyem per una mirada crítica a la cultura digital”.

En la Universidad Pompeu Fabra, el grupo Research Group on Political Communication, Journalism and Democracy (POLCOM-GRP) lidera los proyectos “Instrumentos de rendición de cuentas ante la desinformación: impacto de las plataformas de *fact-checking* como herramienta de *accountability* y propuesta curricular”, de financiación nacional, y “El odio en las redes sociales: el ágora de la misoginia”, financiado por el Instituto Catalán Internacional por la Paz (ICIP).

También en la misma universidad, pero en el área de Ingeniería y Tecnologías de la Información, encontramos el proyecto “ERINIA” (Evaluating the Robustness of Non-Credible Text Identification by Anticipating Adversarial Actions), cuyo objetivo es explorar la robustez de los clasificadores de texto e identificar herramientas que permitan la detección de desinformación.

Por su parte, el Grupo de Investigación en Derecho Internacional Público y Relaciones Internacionales presenta el proyecto “Discriminación, desinformación, polarización y violencia en el ciberespacio: ¿qué papel para las normas Internacionales?”, que reflexiona sobre los desafíos teóricos y prácticos que suponen algunos fenómenos en las plataformas digitales, como la discriminación, la desinformación, la polarización y la violencia, financiado por el Instituto Catalán Internacional por la Paz.

En la Universidad Ramón Llull, encontramos el grupo DIGILAB: Media, Strategy and Regulation, altamente especializado en la materia y que dirige varios proyectos de interés. Entre ellos, destacan: “Qué pasa con las noticias? Lucha contra la desinformación en WhatsApp: el enfoque de los usuarios”, financiado por WhatsApp; “DISINFTRUST. Disinformation & Trust” financiado por EMI-FUND; o “Confiamos en las noticias falsas: Identificación de las características individuales, grupales y narrativas que impulsan la confianza en los contenidos engañosos”, financiado por la Fundación La Caixa.

También en esta universidad, el grupo GLOBALCODES: Globalización, conflictos, seguridad y desarrollo dirige el proyecto “¿Qué verdad? Una mirada crítica a la pluralidad de discursos de la sociedad civil sobre la construcción de la verdad en la posguerra”, financiado por el Instituto Catalán Internacional por la Paz (ICIP) y AGAUR - Generalitat de Catalunya.

Ambos grupos de investigación están adscritos al Instituto de Investigación en Comunicación y Relaciones Internacionales.

Por otra parte, en la Universitat Abat Oliba CEU identificamos a la investigadora responsable del proyecto “Aprendizaje mediático durante la crisis de la COVID-19 en España: Claves para la eficacia de los verificadores en la lucha contra la desinformación”, financiado por la Beca Leonardo de la Fundación BBVA, y centrado en analizar empíricamente los efectos que el consumo de información tiene en el nivel de desinformación de la ciudadanía.

También el área de derecho administrativo de la Universidad de Lleida desarrolló entre 2017-2020 el plan nacional “Salud pública en transformación: desinformación, alimentación y cambio climático”.

Castilla-La Mancha

Labintic, en la Universidad de Castilla-La Mancha, coordina el proyecto “Pensamiento computacional: habilidades digitales para el siglo XXI desde una perspectiva inclusiva y equitativa de género y rural”, focalizado en fomentar la alfabetización digital. Es el único proyecto identificado en esta comunidad, aunque también desde el grupo NARRATEC: Periodismo: narrativas y tecnología comienzan a recogerse publicaciones relacionadas.

Comunidad de Madrid

En la Comunidad de Madrid se observa una clara hegemonía y liderazgo de los grupos de investigación pertenecientes al campo de la comunicación. En la Universidad CEU San Pablo, el grupo CEICIN: Centro de Estudios e Investigación sobre Comunicación e Infancia actúa como coordinador del programa “Nuevos Escenarios de Vulnerabilidad Digital: Alfabetización Mediática para una sociedad inclusiva PROVULDIG2-CM”, cuya investigación aborda la protección de colectivos vulnerables y desinformación. Este proyecto de investigación, financiado por la Comunidad de Madrid, aglutina a 6 grupos de distintas universidades. Se encuentran entre ellos: BRECHAYMAYORES: Brecha Digital y Mayores, también de la Universidad CEU San Pablo; ComR: Grupo Comunicación Responsable y Públicos Vulnerables de la Universidad Complutense de Madrid (UCM); GICOMSOC: Grupo de Investigación en Comunicación, Sociedad y Cultura y PARTICYPAD: Grupo de Investigación sobre Participación Ciudadana y Alfabetización Digital, ambos de la Universidad Rey Juan Carlos (URJC) y VILLANUEVA-OCS: Villanueva Observatorio Comunicación y Sociedad de la Universidad Villanueva.

Igualmente, el grupo GICOMSOC de la URJC, mencionado anteriormente, fue responsable del plan nacional “Redes sociales, adolescentes y jóvenes: convergencia de medios y cultura digital”, orientado al estudio de los riesgos digitales, la alfabetización mediática y la ciudadanía digital en esta franja de edad.

También en esta universidad, INECO (Grupo de investigación consolidado en Innovación, Educación y Comunicación) ha sido responsable del proyecto “Estudio de los factores condicionantes de la desinformación y propuesta de soluciones contra su impacto en función de los grados de vulnerabilidad de los grupos analizados”, financiado por la Fundación Luca de Tena, en colaboración con la red social estadounidense Facebook. Varios de sus miembros participan, además, en el proyecto “Iberian Digital Media Research and Fact-Checking Hub” (IBERIFIER), coordinado por la Universidad de Navarra. al que también se adscriben miembros del grupo CPYC: Comunicación, políticas y ciudadanía, de la Universidad Carlos III (UCIII) y del grupo AIDA: Applied Intelligence and Data Analysis Group, de la Universidad Politécnica de Madrid.

Este último, AIDA, investiga sobre tecnologías de inteligencia artificial y computación para la detección de desinformación y presenta una alta especialización en investigación sobre desinformación. Registra varios proyectos financiados como “Caracterización inteligente de la veracidad de información asociada a la COVID-19 (CIVIC)”, “Lucha contra los trastornos de la información en las redes sociales online” o “Seguimiento de la desinformación en las redes sociales en línea mediante el procesamiento profundo del lenguaje natural”, entre otros. También en esta universidad, el Grupo de Sistemas Inteligentes recoge el proyecto de financiación nacional “Desentrañando los efectos de la desinformación en el ámbito financiero y bancario a través del aprendizaje automático, el análisis de redes sociales y la simulación social basada en agentes”.

Desde la Facultad de Ciencias de la Información de la Universidad Complutense, el grupo Historia y Estructura de la Comunicación y del Entretenimiento desarrolla el proyecto del

plan nacional “Garantías frente a la desinformación en procesos electorales. Cuestiones de ciberseguridad y otros desórdenes informativos en redes”, enfoque ciertamente enmarcado en los peligros de la desinformación para la democracia. Liderado por un profesor de la misma área de conocimiento y universidad, encontramos el proyecto “Las alianzas mediáticas como motor de la supervivencia del periodismo frente a la desintermediación y la desinformación. Los consorcios y proyectos periodísticos colaborativos”, de enfoque periodístico.

Desde el campo del derecho, el grupo I+DEM. Tecnología y democracia lideró el trabajo “Garantías frente a la desinformación en procesos electorales. Cuestiones de ciberseguridad y otros desórdenes informativos en redes” y el grupo Inteligencia Artificial, Democracia y Estado de Derecho abanderó el plan nacional “Proyecto Fortalecimiento de democracia y Estado de Derecho a través de la IA”.

Por su parte, el grupo Data Science and soft computing for social analytics and decision aid”, del departamento de Sociología aplicada, lidera el trabajo “La Estructura de la Comunicación en Red y la Opinión Pública Inclusiva. Un Estudio con Técnicas de *Big data* y Análisis de Redes sociales”, que ahonda en las condiciones principales que generan una deriva del debate público hacia escenarios de comunicación fallida.

Focalizado en la alfabetización mediática, el grupo Información, Biblioteca y Sociedad (INFOBISCO) abanderó el proyecto del plan nacional “Competencias en información para afrontar el discurso de odio en Educación Secundaria Obligatoria y Bachillerato (CIADOE)”.

Finalmente, desde el ámbito de la seguridad, resulta particularmente interesante la aportación del proyecto “Los sótanos de la desinformación: de usuarios a terroristas en la sociedad digital”, cuyo investigador principal pertenece también a la UCM, y que tiene como propósito ahondar, desde un enfoque filosófico, en los mecanismos persuasivos de la desinformación.

En la Universidad Autónoma de Madrid se recoge el proyecto “Polarización y discursos digitales: perspectivas críticas y socio-cognitivas”, coordinado por dos profesoras del Departamento de Filologías y su Didáctica, aunque no ha sido posible identificar su adscripción a un grupo concreto.

Por otra parte, se han registrado varios grupos que presentan entre sus líneas de investigación, la desinformación y/o la alfabetización mediática, aunque no se ha podido acceder a la información de sus proyectos. Se trata de CINSEDTAT: Centro de Investigación Nebrija en Seguridad, Estado de Derecho y Altas Tecnologías e INNOMEDIA: Innovación en Comunicación & Medios de la Universidad Nebrija; INTERACTION: Sociedad interactiva: audiencias, nuevos medios y *educomunicación* de la Universidad Camilo José Cela; PASEET: Periodismo y análisis social: evolución, efectos y tendencias y el grupo Innovation on Digital Media, ambos de la UCIII; SUADE: Retórica de los medios en la sociedad digital de la Universidad Europea de Madrid.

En este sentido, el grupo UC3M MEDIALAB: Analytics, Media and Public Engagement: Communication, Journalism and Technology Laboratory de la UCIII indica que varios de sus integrantes participan en proyectos de investigación europeos, nacionales y regionales tales como la Cátedra Jean Monnet “EU, disinformation & fake news”.

Comunidad Foral de Navarra

La Universidad de Navarra es la única institución de la comunidad que alberga grupos de investigación especializados en la materia. En este caso se recogen cuatro grupos con enfoques diferenciados.

En el área de Periodismo, Digital News Media Research Group centra su investigación en los medios periodísticos digitales y comprende entre sus líneas de investigación el estudio altamente especializado de la desinformación y los desórdenes informativos. Integrado en el Digital Unav - Center for Internet Studies and Digital Life de la Facultad de Comunicación, este grupo lidera el proyecto financiado por la Comisión Europea “IBERIFIER - Iberian Digital Media Research and Fact-Checking Hub” que, como se ha expuesto anteriormente, forma parte de la red del Observatorio Europeo de los Medios Digitales (EDMO). En IBERIFIER colaboran 23 organizaciones, entre las que se cuentan 10 universidades españolas (Universidad de Granada, Universidad Carlos III de Madrid, Universidad de Santiago de Compostela, Universidad Politécnica de Valencia, Universidad Politécnica de Madrid, Universidad Miguel Hernández de Elche, Universitat de València – Estudi General, Fundación Universitaria San Pablo CEU y Universidad Rey Juan Carlos) y 2 portuguesas (ISCTE – Instituto Universitário y Universidade de Aveiro), así como verificadores, agencias de noticias y diversas instituciones de investigación multidisciplinar. Además, en los últimos años, este grupo ha dirigido otros 3 proyectos sobre desinformación y alfabetización mediática, de financiación privada (Fundación BBVA, MediaWise [EEUU] y UTECA). El Digital News Media Research Group también dirige el proyecto “Medios nativos digitales en España: Tipologías, audiencias, construcción de la confianza y claves para la sostenibilidad periodística”, integrado en el proyecto coordinado nacional Diginativemedia, que lidera la Universidad de Santiago de Compostela y que anualmente arroja interesantes datos sobre los patrones de consumo informativo digital.

La Universidad de Navarra presenta otros dos grupos que investigan sobre la desinformación, con un grado de especialización menor. También en el área de Periodismo, encontramos el grupo Comunicación de la Ciencia, especializado en el estudio de este campo en medios audiovisuales y digitales. Cuenta con varios proyectos financiados en el ámbito de la comunicación de la ciencia, pero ninguno específicamente centrado en la desinformación. Hasta ahora, su investigación sobre desinformación se ha realizado en el marco de proyectos dirigidos por otros grupos de su universidad.

El grupo Discurso Público, del área de Lingüística, está especializado en las aplicaciones de la lingüística y el análisis del discurso a los problemas sociales de nuestro tiempo, entre

los que se incluye la difusión de contenidos desinformativos y discursos del odio. Entre los proyectos que ha desarrollado se encuentra CoMMITTEd. COVID, migrantes y minorías en la formación docente, que generó, entre otros resultados, un observatorio de noticias falsas para promover el pensamiento crítico y la alfabetización digital en tiempos de crisis.

Comunidad Valenciana

En la Universidad de Alicante predominan los proyectos del área de informática. El Grupo de Procesamiento del Lenguaje Natural y Sistemas de la Información dirige el proyecto “NL4dismis: Tecnologías del Lenguaje Natural para lidiar con la desinformación” o “SocialFairness: Análisis de la honestidad en medios digitales”, entre otros. Se añade el Grupo Lucentia, centrado en el desarrollo de soluciones de inteligencia empresarial y eLearning. Entre sus proyectos registra “RETISTRUST: Red temática de investigación en el campo de la seguridad y confianza para los sistemas de información en una sociedad conectada”, de financiación nacional. Por su parte, el grupo Living-Lang. Tecnologías del lenguaje humano para entidades digitales vivas actúa como responsable del trabajo “Modelado del comportamiento de las entidades digitales mediante tecnologías del lenguaje humano”.

La Universidad Jaume I presenta tres grupos vinculados a las áreas de Comunicación Audiovisual y Publicidad y Periodismo: el grupo ITACA - Investigación en Tecnologías Aplicadas a la Comunicación Audiovisual, promotor del proyecto “Análisis de las identidades discursivas en la era de la posverdad”, y el grupo Periodismo, Comunicación y Poder, con el proyecto nacional “El rol de la ciudadanía en la comunicación política”, y el proyecto autonómico “Noticias falsas sobre la COVID-19: colectivos afectados, efectos sobre la democracia y empoderamiento ciudadano”.

En la Universidad Miguel Hernández de Elche, encontramos el proyecto “#FakePenal: Impacto de la *tuitización* de la opinión pública sobre el Derecho Penal y consecuencias para el debate en torno a su democratización”, que analiza la transformación del debate público y mediático en torno al Derecho Penal, a raíz de la popularización de las redes sociales, planteando propuestas deliberativas para combatir fenómenos como la desinformación y la polarización. El investigador principal está adscrito al grupo “Tecnología, Mente y Comportamiento social y desviado”.

Finalmente, fruto de la coordinación entre la Universidad de Valencia y la Universidad de Valladolid, identificamos un grupo altamente especializado, Mediaflows, que aborda el estudio de la influencia. En su desarrollo presenta numerosos proyectos relacionados como “Flujos de desinformación, polarización y crisis de la intermediación mediática (Disflows)”, de financiación nacional; “Ecología de la desinformación: la construcción de las noticias falsas y su impacto en el espacio público”, financiado por la Generalitat Valenciana, o proyectos ya finalizados como “Ecología de la desinformación: la construcción de las noticias falsas y su impacto en el espacio público”.

Desde el Centro de Investigación Pattern Recognition and Human Language Technology de la Universidad Politécnica de Valencia se realiza una intensa y fructífera labor de investigación. Entre los numerosos proyectos que desarrollan, se encuentran “FAKE news and HATE speech (FAKEHATE-PdC)” y “Malicious Actors Profiling and Detection in Online Social Networks Through Artificial Intelligence (MARTINI)”, de financiación europea, o “IA explicable para desinformación y detección de conspiración durante infodemias (XAI-DISINFODEMICS)”, entre otros.

Galicia

En la Universidad de A Coruña, el Grupo de investigación Filosofía y Metodología de las Ciencias de lo Artificial es responsable del proyecto nacional “Ciencias de internet, dimensión social de las redes y fake news: de ciencias de la comunicación como ciencias aplicadas de diseño a aplicación de ciencias para evitar”, cuyo objetivo es buscar las causas del fenómeno de las fake news en su triple vertiente semántica, epistemológica y ontológica.

También desde el grupo NECOM – Grupo de investigación en neurocomunicación, publicidad y política, se registra el proyecto del plan nacional “Lucha contra la desinformación y criterios de valor en los debates electorales en televisión y medios digitales”, focalizado en los debates electorales televisados, la comunicación política y la participación de las audiencias.

Cabe hacer referencia al trabajo del Centro de Investigación en Tecnoloxias da Información e Comunicacions, adscrito a la Universidad de A Coruña, desde el que se desarrolla la investigación de financiación nacional “Desarrollo de una tecnología de IA para la autenticación de usuarios basada en el comportamiento”, para validar los usuarios reales con mayor calidad y certeza.

En la Universidad de Santiago de Compostela encontramos el grupo Derecho y Ciudadanía, del área de Derecho, que presenta un proyecto activo titulado “La manipulación informativa como problema de seguridad y de calidad democrática: descripción, consecuencias y respuestas”. En la misma universidad, pero en el área de Periodismo, el grupo Novos Medios participa en el proyecto IBERIFIER, coordinado por la Universidad de Navarra.

Destacan en la comunidad autónoma de Galicia los grupos de investigación de la Universidad de Vigo. Enmarcado en el área de Comunicación Audiovisual y Publicidad encontramos el grupo Comunicación Persuasiva, responsable del proyecto nacional de investigación “Mapa de la Desinformación en las Comunidades Autónomas y Entidades Locales de España y su ecosistema digital (FAKELOCAL)”. Anteriormente, este grupo lideró el proyecto “Lucha contra la desinformación y criterios de valor en los debates electorales en televisión y medios digitales: plataforma de verificación y blockchain (DEBATrue)”.

Adscrito a la misma área se encuentra el grupo “Investigación en Comunicación”, responsable del proyecto nacional competitivo “FACTCHECKERS: Narrativas digitales contra la desinformación: estudio de redes, temas y formatos en los *fact-checkers* iberoamericanos”.

En el área de informática, el grupo SING: Sistemas Informáticos de Nueva Generación lidera el proyecto “CURMIS4th: Modelos basados en aprendizaje automático para detectar y frenar la desinformación sobre salud en redes sociales”, analiza cómo interactúan los individuos en redes sociales con información errónea sobre salud.

La Rioja

En el caso de la Universidad Internacional de la Rioja se identifica el grupo COYSODI: Comunicación y Sociedad Digital, que abanderará el proyecto “La alfabetización publicitaria ante el teléfono móvil. Análisis de la capacidad del público infantil para enfrentarse a los contenidos persuasivos (Ad Kids Mobile)”.

Por otra parte, especializado en el análisis de discurso de odio, encontramos el grupo SIMI: Inclusión socioeducativa e intercultural, Sociedad y Medios, con proyectos indirectamente relacionados con la desinformación, como “Androcentrismo, discursos de odio y sesgos de género a través de los videojuegos *online* en Castilla-La Mancha”, financiado por la Junta de Comunidades de Castilla-La Mancha. A través de la financiación de la propia UNIR, anteriormente desarrollaron el proyecto Disinformation Politics Lab – UNIR, cuyo objetivo era ahondar en la comprensión del panorama de la desinformación digital promovida por y desde actores políticos en España.

Murcia

Desde tres áreas distintas, la Universidad de Murcia presenta varios grupos especializados. El primero, Comunicación Social, Cultura y Tecnología, del área de Comunicación Audiovisual y Publicidad, responsable del proyecto financiado “Ecosistemas de innovación en las industrias de la comunicación”. En segundo lugar, el Grupo de Análisis Político, del área de Sociología, que lidera el proyecto financiado por la Fundación Séneca “Polarización afectiva en la región de Murcia: un estudio sobre sus causas”. Y, por último, el grupo Sistemas Inteligentes y Telemática, adscrito a Ciencias de la Computación e Inteligencia Artificial, que presenta varios proyectos de financiación europea especializados en ciberseguridad y ciberdefensa.

El grupo Tecnologías de Modelado, Procesamiento y Gestión del Conocimiento (TECNOMOD) lidera el proyecto “Tecnologías del lenguaje para la provisión de servicios personalizados”, con foco en la identificación del discurso de odio.

País Vasco

La Universidad del País Vasco presenta dos grupos de investigación en el área de Periodismo: Gureiker y Bitartez. El primero está especializado en cibermedios y recoge dos proyectos financiados de investigación: “Contrarrelatos automatizados contra la desinformación y la incitación al odio para periodistas y medios sociales” y “Comunicación, emociones y acción política. Estrategias para combatir la desinformación y la polarización en la esfera pública digital”.

Por su parte, el grupo Bitartez aborda la transformación profesional de la profesión periodística y presenta el proyecto JOSAFCON (Journalist Safety Research Project), que analizará la seguridad de los periodistas españoles que cubren conflictos internacionales antes, durante y después de sus coberturas informativas.

También desde el área de informática, se identifica el liderazgo como investigador principal del responsable del proyecto “Aprendizaje y argumentación para detectar y combatir la desinformación en las redes sociales”, centrado en generar verificación a través de la inteligencia artificial.

Universidad Nacional de Educación a Distancia

En el caso de la Universidad Nacional de Educación a Distancia, existen dos grupos de investigación que pertenecen a áreas de conocimiento bastante alejadas entre sí. Por una parte, el grupo UNED Research Group in Natural Language Processing and Information Retrieval coordina el proyecto nacional “Modelos de Lenguaje Profundos en la Interpretación de la Información y la Desinformación en Contexto”, investigación que da continuidad a su proyecto anterior “Desinformación y agresividad en Social Media: bias, controversia y veracidad”. El grupo también participa en el proyecto “Análisis holístico de la actividad organizada de desinformación en las redes sociales”, de carácter internacional.

Por otra parte, el Laboratorio de Fonética “Antonio Quilis presenta la investigación, también de carácter nacional, “¿Qué hace humana a una voz? Hacia una mejor comprensión de las características fonéticas que permiten distinguir voces reales de *deepfakes*”, que traza pautas para facilitar el reconocimiento de voz humana frente a voces creadas por inteligencia artificial.

No se han encontrado grupo especializados en desinformación en las comunidades de Canarias, Illes Balears, Extremadura y Principado de Asturias. No obstante, la limitación de este trabajo reside en la dependencia a la información pública, que no siempre está actualizada en las correspondientes páginas web.

ANÁLISIS DE LOS PROYECTOS DE INVESTIGACIÓN FINANCIADA SOBRE DESINFORMACIÓN

En el desarrollo de la carrera académica, obtener financiación resulta fundamental para sufragar los gastos que supone el desarrollo de una investigación. En los mismos se engloban los costes propios del trabajo a realizar, pero también otras cuestiones propias de la difusión de conocimiento como las publicaciones y actividades para la comunidad científica o las actividades de divulgación necesarias para trasladar dicho conocimiento a la sociedad.

Periódicamente, la Agencia Estatal de Investigación, así como otras agencias autonómicas y organizaciones privadas, abren convocatorias a las que concurren los equipos de investigación con propuestas muy trabajadas en las que se detallan tanto los objetivos de trabajo, como los miembros participantes y las partidas presupuestarias. Estas propuestas son valoradas por un panel de expertos que finalmente determina cuáles obtendrán financiación, procediendo en el proceso a un seguimiento y una evaluación final en la que se determinan los objetivos alcanzados y se auditan los gastos del proyecto.

Proyectos de investigación financiados por la Agencia Estatal de Investigación sobre desinformación

Nos proponemos en este epígrafe presentar una fotografía de los proyectos de investigación financiados especializados en desinformación. Objetivo que hemos tratado de alcanzar analizando, en un primer lugar, los proyectos concedidos por la Agencia Estatal de Investigación (AEI) entre 2016 y 2022, ampliando el espectro temporal respecto a estudios anteriores (Rúas-Araújo y Paniagua-Rojano, 2023) y añadiendo también la financiación privada de fundaciones de primer nivel.

El objetivo inicial pretendía incluir también la financiación autonómica. Para realizar este análisis dependíamos de la información publicada en las páginas web de las agencias autonómicas de investigación. Sin embargo, la exploración inicial mostró cómo esta información no era, en muchos casos, ni homogénea ni completa. Decidimos por tanto eliminar este objetivo de trabajo que nos habría permitido ofrecer un análisis más completo de los proyectos concedidos por todas las comunidades autónomas. Igualmente, se descartó el análisis de financiación internacional por razones similares.

Como primer paso se listaron 53 palabras clave, vinculadas a la desinformación que podían servir para el rastreo de proyectos. Entre ellas se recogían términos relacionados con la alfabetización mediática y digital, las amenazas híbridas, la posverdad o los sesgos

cognitivos, entre otros. Estas palabras clave fueron facilitadas a la AEI que, tras consultar en su base de datos, nos devolvió un archivo con un total de 900 resultados. Estos datos fueron filtrados manualmente para obtener una muestra de proyectos relacionados con la desinformación, que se categorizó en base al año de adjudicación, el área de conocimiento, la financiación y la temática. El trabajo fue realizado por María José Estables, investigadora postdoctoral Juan de la Cierva en la Universidad Complutense de Madrid, y Leticia Rodríguez Fernández, profesora titular en la Universidad de Cádiz.

Tras esta primera revisión, se emplearon las mismas palabras clave, en esta ocasión, para realizar una búsqueda en la propia página web de la AEI con el objetivo de completar la información. Con la combinación de ambos procesos de filtrado se obtuvo la cifra de 74 proyectos de investigación centrados en la desinformación.

El año en el que más proyectos se registran es 2021, con un total de 31, probablemente, por la confluencia de resoluciones de dos tipos distintos de proyectos de investigación: los Proyectos de I+D+I (Generación de Conocimiento y Retos Investigación) y los Proyectos Estratégicos Orientados a la Transición Ecológica y a la Transición Digital. El resto de años se registran números considerablemente menores: en 2016 solo hay un proyecto de alfabetización mediática y digital; en 2017 el número sube a 3 y, a partir de 2018 se recoge un crecimiento interesante con 8 proyectos ese año; 10 en 2019; 13 en 2020, y se alcanza el máximo, como se ha comentado en 2021. En 2022, el número de proyectos cae a 6.

En total, y según el sumatorio obtenido de los proyectos, la AEI ha proporcionado 5.912.402 de euros de financiación al estudio de la desinformación, lo que nos daría una media de 79.897 euros por proyecto. El proyecto con menor presupuesto parte de 9.000 euros y el que mayor financiación obtiene llega a los 446.641 euros.

El área con mayor número de proyectos financiados es Tecnologías de la Información y de las Comunicaciones, con 30, y de los cuales 27 corresponden al campo de las Ciencias de la Computación y Tecnología Informática y 3 a la subárea de Tecnologías de las Comunicaciones. Le sigue el área de Ciencias sociales (26), destacando el campo de la Comunicación, que alcanza los 24 proyectos financiados, siendo los 2 restantes de las subáreas de Sociología y Antropología Social y Ciencia Política. En tercer lugar, encontraríamos el área de Mente, Lenguaje y Pensamiento (4), de los cuales 2 pertenecen a la subárea de Filosofía y 2 a Lingüística y Lenguas.

También se recogen 8 proyectos financiados en el área de Derecho. De ellos, 2 son en el área de Economía, uno de ellos en el campo de la Economía, Métodos y Aplicaciones y el otro en Empresas y Finanzas, 2 proyectos en el área de Educación y 1 en el área de Psicología.

La alfabetización mediática y digital (12) y los discursos de odio (12) son las temáticas más abordadas, seguidas del análisis del discurso (7), la detección y lucha contra la

desinformación (6), las redes sociales (4), el periodismo (4), los procesos electorales (3), la verificación de información/verificadores (3), salud (3), opinión pública (3) y democracia (3).

Igualmente, se recogen proyectos relacionados con la confianza (*socialtrust*) (2), los ecosistemas (2), la ética (2), el ámbito financiero (1), la comunicación política (1), el análisis de redes (1), *deepfakes* (1), desinformación local (1) y *marketing* y comunicación (1). Cabe destacar que solo se registran dos proyectos relacionados con el ámbito de la seguridad: “Los sótanos de la desinformación. De usuarios a terroristas en la sociedad digital” de la facultad de Filosofía de la Universidad Complutense de Madrid y “La manipulación informativa como problema de seguridad y de calidad democrática: descripción, consecuencias y respuestas” de la Universidad de Santiago de Compostela.

Por otra parte, el 32,8% de los proyectos de investigación concedidos por la AEI entre 2016 y 2022 tienen como objetivo establecer medidas de detección y prevención contra la desinformación. La inteligencia artificial es la gran base de algunas de estas propuestas centradas, entre otras cuestiones, en el análisis predictivo del lenguaje, la determinación de características fonéticas que permitan identificar *deepfakes* o el desarrollo de tecnologías para diferenciar bots en base al comportamiento de los usuarios.

Proyectos de investigación sobre desinformación financiados por entidades privadas

El interés por financiar proyectos de investigación para luchar contra la desinformación también se extiende a las entidades privadas. En concreto, entre los años 2016 y 2022 se han financiado en España proyectos por parte de las fundaciones privadas españolas de La Caixa, BBVA y Luca de Tena. A lo largo de este periodo, entre las tres instituciones se han financiado un total de 21 proyectos de investigación: 5 proyectos (Fundación La Caixa); 2 proyectos (Fundación Luca de Tena) y 14 proyectos (Fundación BBVA).

La Fundación La Caixa cuenta con distintos programas de financiación para proyectos de investigación, la mayoría de ellos en cuestiones de salud y ciencias naturales. No obstante, en el caso de los cinco proyectos financiados, todos ellos han sido financiados a través de su Observatorio Social, en concreto, gracias a la convocatoria de investigación social llamada *Social Research Call* que tiene como objetivo fomentar la generación de conocimiento relevante para comprender e identificar los retos a los que se enfrenta la sociedad, a través de estudios que se fundamentan en datos cuantitativos. La dotación máxima de estos proyectos es de 100.000 euros y se desarrollan en un periodo máximo de 24 meses. En este sentido, los proyectos centrados en cuestiones de desinformación, polarización política y alfabetización mediática y géneros financiados por la Fundación La Caixa han sido concedidos en las convocatorias de los años 2019 (Proyecto “Mind the Gap”); 2021 (Proyecto “Feeding-fakes” y “U-Mid”) y 2022 (Proyecto “MISCOR” y “PREOPOL”).

La fundación Luca de Tena es una entidad privada centrada en el sector del periodismo español desde 1930 y que en la actualidad se dedica a las labores de investigación, divulgación y formación sobre la evolución del periodismo con el objetivo de ayudar a los periodistas a adaptarse a los retos a los que se ven inmersos por los cambios en el sector de la prensa. En los últimos años, y a través de su Laboratorio de Periodismo, la fundación Luca de Tena ha financiado dos proyectos relacionados con las temáticas de la desinformación y la alfabetización mediática. En concreto, se trata del proyecto titulado “Estudio de los factores condicionantes de la desinformación y propuesta de soluciones contra su impacto en función de los grados de vulnerabilidad de los grupos analizados” llevado a cabo en 2019 y cofinanciado con la red social estadounidense Facebook. Asimismo, en 2022, a través del Laboratorio de Periodismo se financió el proyecto “Informe sobre alfabetización mediática: contexto actual, legislación, casos de éxito, herramientas y recursos, y percepción y propuestas de especialistas y profesores”.

La entidad privada que más proyectos ha financiado sobre temas de desinformación, propaganda digital, polarización política y alfabetización mediática y digital es la Fundación BBVA. En concreto, esta institución cuenta con programas de apoyo a la investigación científica y humanística en España, tanto para grupos de investigación consolidados como para proyectos liderados por jóvenes investigadores e investigadoras que se encuentran en la mitad de su carrera laboral (Becas Leonardo). En lo relativo a los proyectos de investigación, se han concedido, entre los años 2016 y 2022, un total de 14 proyectos (6 a proyectos de grupos de investigación) y 8 a través del programa de Becas Leonardo. A partir del año 2020, se detecta un mayor interés de esta entidad privada para financiar proyectos centrados en la lucha contra la desinformación y las noticias falsas debido a la proliferación de este tipo de contenidos a raíz de la pandemia mundial de la COVID-19.

Desglosando los proyectos según las convocatorias realizadas por la Fundación BBVA, en el año 2017, se financió, a través del programa “Ayuda a Equipos de Investigación Científica. Economía y Sociedad Digital” el proyecto “Tecnología digital, polarización ideológica e intolerancia (TECNOPOL)”. En el año 2018, en esa misma convocatoria, se concedieron los proyectos “Intervenciones inteligentes para adolescentes y jóvenes en la sociedad digital (E-WISE)” y “Aplicaciones económicas de distancias culturales generadas con información proveniente de las redes sociales (AEIRS)”.

Por otro lado, en el año 2019, se financiaron dos investigaciones: una en la convocatoria “Ayudas a equipos de investigación científica. Big Data” que se llamó “AMALFI: Aprendizaje de máquina frente a al software malicioso y las noticias falsas” y otra en la convocatoria Becas Leonardo que se denominó “La controversia pública: un análisis desde la epistemología social”.

En el año 2020, la fundación costeó un total seis proyectos, uno de ellos a través del programa “Ayudas a equipos de investigación SARS-COV-2 y COVID-19” que se denominó “Detección inmediata de *fake news* relacionadas con la COVID-19” y otros cinco proyectos a través de la convocatoria de Becas Leonardo: “Parlamento, *bots* y desinformación: ecología

de estrategias y prácticas en redes sociales”; “Aprendizaje mediático durante la crisis de la COVID-19 en España: Claves para la eficacia de los verificadores en la lucha contra la desinformación”; “Alfabetización mediática en los centros de secundaria. ¿Cómo se enseña a los adolescentes a consumir información?”; “Análisis de las medidas protectoras contra la desinformación estratégica en las redes sociales” y “Conceptos de conocimiento científico y controversias públicas: una aproximación desde la filosofía experimental”.

Además, en el caso del año 2021, se costearon dos proyectos a través del programa “Ayudas a Proyectos de Investigación Científica. Filosofía”. El primero se denomina “Los efectos positivos de visibilizar los desacuerdos en las redes sociales” y, el segundo, “Una herramienta para detectar y medir la polarización”. También se llevó a cabo otro proyecto a través del programa Becas Leonardo que se llama “Discursos políticos, racialización y racismo en espacios digitales: netnografía de la contestación ciudadana y nuevas formas de participación social”.

En el momento de redacción de ese informe, no se había aún fallado la resolución de la convocatoria de proyectos de investigación de 2022.

EL PAPEL ESTRATÉGICO EN EL ÁMBITO NACIONAL DE LAS UNIVERSIDADES Y CENTROS DE INVESTIGACIÓN EN LA LUCHA CONTRA LAS CAMPAÑAS DE DESINFORMACIÓN

La propagación de las campañas de desinformación y las actividades de injerencia y manipulación extranjera de la información se han convertido en un desafío para las sociedades democráticas. Ante este escenario son, en particular, las universidades y los centros de investigación españoles los que emergen como uno de los actores fundamentales que tienen el potencial de contribuir en la detección, análisis, divulgación, formación y respuesta contra las campañas. Su papel estratégico radica en que son fuentes confiables de conocimiento científico y, de forma constante, desarrollan investigaciones y proyectos que abordan la desinformación desde diversas perspectivas. Las universidades y centros de investigación españoles disponen de capacidad investigativa y facilitan la creación de equipos multidisciplinares, contribuyendo así a generar sinergias con otros actores relevantes nacionales y del tercer sector en la lucha contra las campañas de desinformación y en la ayuda a la toma de decisiones fundamentadas en datos objetivos y contrastados.

Se describe en este epígrafe, un primer apartado que aborda las iniciativas y mecanismos europeos existentes en la colaboración con la universidad y las instituciones estatales y/o actores gubernamentales en la lucha contra las campañas de desinformación. En un segundo apartado, se analiza el rol de la universidad y centros de investigación para un mejor conocimiento de las campañas de desinformación y, por último, en un tercero, se proponen iniciativas de colaboración del trabajo académico con la sociedad en su conjunto para el mismo fin. Este trabajo de análisis, documentación y descripción ha sido desarrollado por Eva Campos Domínguez, profesora titular de la Universidad de Valladolid, y María José Rementería Núñez, investigadora senior del Barcelona Supercomputing Center.

Mecanismos europeos de colaboración y cooperación de las Universidades en iniciativas para la lucha contra la desinformación

Este análisis no exhaustivo se ha realizado revisando fuentes abiertas sobre un conjunto de universidades y países europeos que presentan iniciativas significativas en el ámbito de la desinformación. El objetivo es detectar posibles estructuras y organizaciones gobierno-universidad articuladas para la identificación y el contrarresto de la desinformación. Estas iniciativas adquieren diferentes formas en función del país en el que tienen contexto. Adicionalmente, existe un nivel de colaboración internacional impulsado por el Parlamento Europeo y otros organismos internacionales.

Por decisión del Consejo de la Unión Europea (2010) se creó el Servicio Europeo de Acción Exterior (SEAE), una agencia que, entre otros, trabaja en colaboración con los gobiernos de los países europeos en la lucha contra la desinformación y la injerencia extranjera. El objetivo principal de EEAS es contribuir al desarrollo de herramientas que permitan compartir información sobre incidentes de desinformación para su análisis.

El SEAE ha publicado su primer informe (SEAE, 2023) sobre manipulación e injerencia extranjera de la información (FIMI por sus siglas en inglés), proponiendo un enfoque abierto, descentralizado e interoperable para abordar las amenazas FIMI. El informe aboga por un fomento de la colaboración entre miembros de la comunidad, de manera que cada miembro contribuya con sus habilidades, conocimientos y perspectivas. También describe una metodología que permite identificar, analizar y compartir información sobre FIMI de manera confiable, y propone el uso de taxonomías y estándares como DISARM para TTPs (Tácticas, Técnicas y Procedimientos), STIX³ para el almacenamiento e intercambio de información y OpenCTI⁴ como plataforma de código abierto con el fin de estructurar, almacenar, organizar y visualizar información sobre actividad FIMI. El documento sugiere priorizar el uso de estándares y marcos interoperables para facilitar el análisis profundo, la discusión y el aprendizaje sobre incidentes FIMI, así como fomentar la acción colectiva en su lucha contra FIMI.

El Observatorio Europeo de Medios Digitales (EDMO)⁵ promueve el conocimiento científico sobre la desinformación, los servicios de verificación, apoyando los programas de alfabetización mediática como una de las principales medidas contra la desinformación. Está organizado en hubs que cubren el territorio europeo, y reúnen a investigadores académicos, agencias multidisciplinarias, periodistas y verificadores con experiencia en el campo de la desinformación. EDMO también apoya a las autoridades públicas en la evaluación de la implementación del Código de Prácticas de la UE sobre desinformación.

La Media & Learning Association (MLA)⁶ es una asociación internacional sin ánimo de lucro creada en 2012 bajo la ley belga para promover y estimular el uso de los medios digitales como una forma para mejorar la innovación y la creatividad en la enseñanza y el aprendizaje en Europa. Forman parte las agencias nacionales y regionales encargadas de la promoción de la innovación en la enseñanza y el aprendizaje, así como universidades, ministerios de educación y redes de escuelas. Son miembros de EDMO, desarrollan el paquete de trabajo sobre formación.

³ <https://oasis-open.github.io/cti-documentation/stix/intro.html>

⁴ <https://github.com/OpenCTI-Platform/opencti>

⁵ <https://edmo.eu/>

⁶ <https://media-and-learning.eu/>

El Centro de Excelencia de Comunicaciones Estratégicas de la OTAN (NATO StratCom COE)⁷ comenzó a funcionar en 2014. Es una organización militar internacional constituida multinacionalmente (España no forma parte del centro) y acreditada por la OTAN, aunque no forma parte de la estructura de mando. Su propósito es incorporar la planificación de las comunicaciones estratégicas para desarrollar apoyo público a las políticas y actividades de la OTAN. Actualmente, investiga, principalmente, las campañas de desinformación relacionadas con la guerra de Ucrania. En diciembre de 2020, el NATO StratCom COE y la facultad de Información y Ciberespacio de la Universidad Nacional de Defensa de EE.UU. firmaron un memorándum de entendimiento⁸ para potenciar la preparación en diversos aspectos de la seguridad nacional, en el que se incluye el entorno de la información. Asimismo, en el verano de 2019, se puso en marcha la KCSC-NATO StratCom COE Summer Academy⁹ en la que participa el King's Center for Strategic Communication (KCSC) del King's College de Londres, que tiene por objetivo acercar a los estudiantes de Comunicación Estratégica las “preocupaciones cotidianas de los que desempeñan y participan en procesos de formulación de políticas, investigación y ejercicios en este campo”.

El Centro Europeo de Excelencia para Contrarrestar las Amenazas Híbridas (Hybrid CoE) de Helsinki es una organización internacional independiente basada en una red que lucha contra las amenazas híbridas, en el que participa España. Tiene un papel importante ya que se configura como un centro que desarrolla las capacidades de los Estados participantes para reventar y contrarrestar las amenazas híbridas. El Hybrid CoE es uno de los principales socios del proyecto EU-HYBNET,¹⁰ en colaboración con la Universidad Laurea de Ciencias Aplicadas (Vantaa, Finlandia). El proyecto, en el que participa Maldita.es, la Universidad Rey Juan Carlos y la Policía Local de Valencia, está financiado por el programa Horizonte 2020 de la Unión Europea, y tiene como objetivo enriquecer las redes europeas existentes para contrarrestar las amenazas híbridas y garantizar su sostenibilidad a largo plazo.

Por otra parte, se han identificado¹¹ diferentes iniciativas de colaboración en distintos países europeos a nivel nacional, como:

⁷ <https://stratcomcoe.org/>

⁸ <https://stratcomcoe.org/news/nato-stratcom-coe-will-cooperate-with-the-national-defence-university-us/135>

⁹ <https://stratcomcoe.org/publications/the-future-is-now-collaboration-with-students-from-kings-college-for-strategic-communications-at-kings-college-london/207>

¹⁰ <https://euhybnet.eu/about/partners/>

¹¹ Información obtenida de fuentes abiertas (última consulta, abril de 2023).

Reino Unido

El gobierno británico, con el fin de ayudar a los gobiernos y comunicadores a reducir de manera efectiva el impacto de la información errónea y la desinformación a través de comunicaciones estratégicas, ha colaborado con el departamento de Comunicación Estratégica de la Universidad de Lund (Department of Strategic Communication at Lund University) en el desarrollo de un toolkit “RESIST 2 Counter Disinformation toolkit” (Pamment, 2021). Una guía que permite reconocer información errónea y desinformación, alerta temprana, conocimiento de la situación, análisis de impacto, comunicación estratégica y seguimiento de la eficacia.

Irlanda

Representación académica en la Estrategia Nacional contra la Desinformación. A principios de 2023 se estableció un grupo de trabajo para desarrollar una estrategia nacional contra la desinformación. La mayoría de sus miembros representan departamentos gubernamentales y autoridades públicas. No obstante, el equipo también incluye una representación de EDMO Irlanda y de la academia (Universidad y Centros de Investigación). Existen varios grupos y subgrupos de trabajo.

Financiamiento para campañas públicas. El departamento de Relaciones Exteriores financió una campaña de *pre-bunking*, organizado por EDMO Irlanda/Dublin City University junto con la Universidad de Cambridge. Los resultados estarán en breve disponibles como un recurso gratuito y los organizadores realizarán una evaluación sobre la eficacia de la campaña.

Análisis del Código de Prácticas de la UE. Desde el año 2018, el regulador nacional de medios de comunicación financia a la Universidad de Dublin City para que realice análisis sobre el código de prácticas de la UE. De momento, se han publicado tres informes y varias presentaciones ante la red de reguladores de medios de la UE (ERGA the network of EU media regulators) y la Comisión Europea.

Investigación sobre las tendencias en medios y desinformación. El regulador nacional de medios financió a el nodo de EDMO Irlanda un análisis del “Reuters Digital News Report”, en relación con el consumo de medios y tendencias de desinformación en todos los países de la EPRA (European Platform of Regulatory Authorities). Los resultados se presentaron en la conferencia de EPRA del año pasado.

Red Nacional de alfabetización mediática. En 2018, el regulador de medios organizó la red nacional de alfabetización mediática con las entidades interesadas (organizaciones de medios, ONG, Universidades, Centros de Investigación). La academia está muy involucrada en esta red y ocupan los cargos de presidencia y vicepresidencia. Esta red a organizado varias campañas de alfabetización mediática, premios y seminarios regulares.

Países Bajos

A finales del 2022, el Gobierno neerlandés presentó una nueva estrategia gubernamental de lucha contra la desinformación¹². Propuso, entre otras cuestiones, trabajar en el fortalecimiento del debate público iniciando campañas de concienciación ciudadana para identificar la desinformación, utilizando alternativas públicas a las redes sociales; y reduciendo la influencia de la desinformación en la sociedad. Y prepara un conjunto de medidas, compatibles a nivel internacional, a aplicar dependiendo del contenido desinformativo, de quien es el productor o difusor, así como cuál es la responsabilidad de los medios y el grado de conocimiento general sobre el tema (Países Bajos, 2022).

La Escuela de Comunicación de la Universidad de Ámsterdam (ASCoR)¹³ dispone de una línea de investigación para desarrollar y probar herramientas que permitan medir la relación entre la confianza y la desinformación de los medios, y restaurar la confianza de la ciudadanía. Adicionalmente, entre sus líneas de trabajo, se encuentran temáticas como la pandemia de la COVID-19 y el cambio climático. También investiga qué grupos sociales tienen más riesgo de desinformación. Los conocimientos resultantes de la investigación los comunican a las autoridades gubernamentales, a los profesionales de la comunicación y al público en general. Entre las acciones de divulgación que realiza se recoge la publicación de artículos científicos y divulgativos en medios de comunicación nacionales y la publicación de libros, la participación activa en debates, la organización de seminarios web periódicos, así como reuniones con profesionales de la comunicación.

Italia

European University Institute's School of Transnational Governance in Florence (Italy)¹⁴ tiene un grupo de investigación sobre desinformación, que se financia en base a proyectos y transfiere el conocimiento académico a través de la organización de foros de debate.

El Instituto Universitario Europeo, junto con la Secretaría de Estado de Digitalización e Inteligencia Artificial de España (SEDIA), y el Instituto para la Gobernanza Democrática han creado una cátedra de Inteligencia Artificial y Democracia con el objetivo de desarrollar actividades de investigación, transferencia de conocimiento, divulgación, docencia e innovación en temas relacionados con la inteligencia artificial y la democracia. Esta cátedra, en colaboración con el clúster Política, Economía y Sociedades Digitales de la facultad, investiga las posibilidades de abordar los retos que plantean la desinformación y las redes sociales.

¹² En la valoración de la amenaza incluye no solo la injerencia extranjera sino también la desinformación de actores domésticos, incluyendo el potencial que las teorías de la conspiración pueden tener para la radicalización violenta.

¹³ <https://ascor.uva.nl/?cb&cb>

¹⁴ <https://www.eui.eu/en/academic-units/school-of-transnational-governance>

España

Las actuaciones en este país se articulan a través de diferentes tipologías: por un lado, a nivel estatal, se encuentra el Foro contra las campañas de desinformación en el ámbito de la Seguridad Nacional¹⁵ impulsado por el Departamento de Seguridad Nacional del Gobierno de España¹⁶, que se configura “como espacio de colaboración público-privada, en el que se recomienden iniciativas para fomentar el conocimiento de la amenaza de las campañas de desinformación y la realización de actividades conjuntas y participativas” y en el que participen, como expertos, entre otros, diferentes académicos e investigadores de centros y universidades españolas. Además de este Foro, las Cortes Generales han articulado también, a través de la Comisión Mixta de Seguridad Nacional, la Ponencia para el estudio del fenómeno de la desinformación y de las *fake news*, con efectos disruptivos en la sociedad, con representación, también, de universidades y centros de investigación. Si bien, no se han encontrado iniciativas de colaboración entre las universidades y los gobiernos autonómicos o locales, cuyo objetivo sea localizable en esta estrategia.

Los mecanismos de participación, colaboración o cooperación de las universidades y centros de investigación españoles en iniciativas para la lucha contra la desinformación se están articulando, fundamentalmente, a través de cátedras específicas, como es el caso de la Jean Monnet Chair “EU, Disinformation & Fake News” de la Universidad Carlos III de Madrid.¹⁷ Salvo contadas excepciones, la mayoría de las cátedras de las instituciones académicas españolas no incluyen en su título el término de desinformación o *fake news*. Existe un amplio número que, bajo el paraguas de diversas disciplinas o enfoques, incluyen entre sus objetivos el tratamiento de la desinformación, pero no son acciones dedicadas exclusivamente a su abordaje (es el caso, por ejemplo, de las cátedras de transparencia, cátedras tecnológicas o de otra índole).

Otro tipo de colaboración que podría articularse en relación con la lucha contra la desinformación son los institutos de investigación centrados en esta materia, aunque no se encuentra ningún ejemplo en activo a día de hoy en nuestro país. De igual forma, tampoco se han hallado convocatorias público/privadas o partidas destinadas específicamente al análisis y divulgación para la lucha contra la desinformación promovidas por las propias entidades universitarias y centros de investigación, aunque algunas de ellas puedan incluir esta cuestión en alguna una de sus líneas implícitas. Es el caso, por ejemplo, de premios

¹⁵ Orden PCM/541/2022, de 10 de junio, por la que se publica el Acuerdo del Consejo de Ministros de 31 de mayo de 2022, por el que se crea el Foro contra las campañas de desinformación en el ámbito de la Seguridad Nacional y se regula su composición y funcionamiento.

¹⁶ <https://www.dsn.gob.es/es/actualidad/sala-prensa/constituci%C3%B3n-del-foro-contra-campa%C3%B1as-desinformaci%C3%B3n-%C3%A1mbito-seguridad>

¹⁷ <https://www.uc3m.es/investigacion/catedras-investigacion/jean-monnet-chair-eu-disinformation-fake-news>

de investigación, premios a iniciativas docentes o bien a trabajos académicos que aborden esta cuestión. Sí existen, sin embargo, proyectos de innovación docente financiados por las propias universidades que los docentes o alumnos impulsan para la lucha contra la desinformación y la alfabetización digital (véase, Desinfaketon Hub de la Universidad Loyola¹⁸).

Las universidades, además, también pueden fomentar la difusión de la investigación sobre desinformación del personal docente e investigador o de transferencia, bien a través de las Unidades de Cultura Científica o de las Oficinas de Transferencia de Resultados de Investigación (OTRI), sin que se registre hasta el momento un número importante de acciones que pueda ser significativas en este sentido en el contexto español. A la vista del análisis efectuado, la principal conclusión que se extrae es que los mecanismos de colaboración y cooperación de las universidades y centros de investigación en España, en iniciativas para la lucha contra la desinformación, son, todavía, incipientes.

Como se ha indicado anteriormente en el presente capítulo, las universidades españolas y el CSIC, en colaboración con las organizaciones de *fact-checking*, han adoptado un papel estratégico incluyendo la desinformación entre sus líneas de investigación, solicitando proyectos de investigación nacionales y europeos o generando una destacable producción científica sobre la materia. No obstante, en el ámbito docente, las iniciativas son escasas y aisladas, sin que exista, por ejemplo, un portal de iniciativas docentes o de investigación para la lucha contra la desinformación o se articulen mecanismos y estructuras organizadas, más que las que puedan desarrollarse dentro de las iniciativas de grupos de docentes o de investigadores aisladas en cada universidad. Un ejemplo de ello puede ser el I International Workshop on Disinformation Research celebrado en la Universidad de Málaga los días 4 y 5 de julio de 2023.

Lo mismo sucede en el ámbito de la transferencia, donde se aprecia una escasez de iniciativas conjuntas entre la universidad, la sociedad civil y tercer sector para luchar contra la desinformación. Salvo casos aislados, las asociaciones de la prensa y los colegios profesionales de periodistas realizan acciones de alfabetización digital por separado, sin que abunden las experiencias conjuntas entre entidades profesionales y universidades y centros de investigación, por ejemplo. Una buena práctica, en este sentido, es la Cátedra RTVE-UAB que trabaja, entre otras acciones, en la elaboración de un directorio de iniciativas nacionales e internacionales relacionadas con la desinformación¹⁹. En el resto de Europa tampoco se identifica una mayor estructuración de la detectada en España.

¹⁸ <http://desinfaketon.uloyoladpcd.com/metodologia>

¹⁹ <https://oi2media.es/tag/catedra-rtve-uab/>

El papel de la universidad y centros de investigación para un mejor conocimiento de las campañas de desinformación

Las universidades y centros de investigación pueden contribuir a un mejor entendimiento nacional del funcionamiento de las campañas de desinformación y la injerencia extranjera a través, fundamentalmente, de acciones implementadas en el ámbito de la docencia, la investigación y la transferencia. En la docencia universitaria, se observan dos claras oportunidades: la formación basada en los principios científicos y la comprensión de la desinformación. Ambas son pilares fundamentales para obtener en el largo plazo una sociedad más resiliente a la desinformación. Pero además se añaden las otras modalidades de formación que amplían estas oportunidades a otros públicos como los títulos propios, los cursos de experto, la universidad para mayores u otras acciones destinadas a la sociedad y que pasan, en todo caso, por la difusión del conocimiento científico.

En la investigación, la universidad y los centros de investigación españoles se enfrentan al reto de continuar promoviendo acciones que permitan innovar en la detección y lucha contra la desinformación, pero también a la creación de mecanismos y espacios de investigación compartidos para la colaboración entre los propios investigadores en la detección de temáticas, metodologías y repositorios de datos. La investigación implica, también, el refuerzo de las líneas de colaboración con otros grupos de investigación internacionales, lo que permitiría una mejor comprensión de este problema y favorecería, desde esta óptica, el abordaje en clave nacional.

La transferencia, por su parte, conlleva la colaboración de la universidad y de los centros de investigación con la sociedad, a través de empresas, instituciones o de la sociedad civil, que deben impulsarse a través de los órganos académicos destinados a tal fin. En este sentido, cabe destacar la labor que se realiza desde las instituciones y sus respectivas Oficinas de Transferencia de Resultados de Investigación (OTRI) y Unidades de Cultura Científica y de la Innovación (UCC+I).

Desde esta perspectiva, la creación de un hub nacional o Instituto de Investigación sobre el Entorno de Información podría convertirse en una oportunidad de valor para mejorar el conocimiento de la sociedad sobre las campañas de desinformación. Este serviría además de apoyo a los mecanismos nacionales de lucha contra la desinformación y permitiría compartir experiencias docentes, investigadoras y de transferencia académica. Este Instituto se podría consolidar como punto de contacto para participar en investigaciones de la sociedad civil y académica a gran escala.

Potenciar el trabajo académico y su accesibilidad a la sociedad en su conjunto

El fenómeno actual de la desinformación es global, complejo, de múltiples facetas y en constante cambio o evolución. Para detectarla y contrarrestarla es necesario su estudio y análisis desde un enfoque multidisciplinar que recoja diversas áreas de conocimiento: Comunicación, Lenguaje y Neuro-lenguaje, Tecnología, Matemáticas, Sociología, Psicología, Filosofía, Impacto social, Económico, Legal y Ético.

Con el objetivo de potenciar el trabajo académico y su accesibilidad a la sociedad, y de contribuir a la cultura de seguridad nacional, a través de las funciones investigadora, docente y de transferencia, se plantean una serie de acciones a considerar en la lucha contra las campañas de desinformación:

- a. **La creación e impulso de redes multidisciplinarias de investigación y/o múltiples redes mono-disciplinarias conectadas.** Para el estudio del fenómeno de la desinformación se considera necesario disponer de una red de investigación formada por grupos de investigación multidisciplinarias y redes mono-disciplinarias conectadas que ahonden en las diversas facetas de la desinformación. Esta red de investigación independiente facilitaría el intercambio de experiencias, conocimiento y buenas prácticas. En este sentido, el hub EDMO existente en la Península Ibérica se vislumbra como un embrión de la red de investigación, así como también en caso de existir varias redes de investigación que se creen foros transversales para compartir conocimiento y debatir sobre el tema, lo que contribuiría a:
 - Favorecer y facilitar la investigación de la desinformación desde diversas perspectivas: identificación de desinformación, difusión, impacto: social, político, reputacional y económico, aspectos éticos y legales, resiliencia social.
 - Fomentar la colaboración público privada por medio de proyectos de colaboración.
 - Disponer de repositorios de información (datos estandarizados relativos a las TTP utilizadas para desplegar campañas de desinformación en el dominio de la información) abiertos a disposición de los investigadores, creándolos a tal efecto. Estos repositorios también facilitarían la investigación pública y privada y el intercambio de información.
 - Compartir conocimiento en el marco del desarrollo de herramientas que permitan la detección temprana y minimización de los efectos nocivos de la desinformación.
 - Transferir a la sociedad el conocimiento desarrollado
 - Elaborar informes científicos abiertos al público

- b. La creación e impulso de redes de alfabetización mediática.** Para contrarrestar la desinformación y rebajar su impacto en la sociedad, es necesario incrementar el conocimiento sobre la misma, fomentar el espíritu crítico, formar en la verificación de hechos, en procedimientos para identificar fuentes confiables y, en general, en cómo actuar ante la desinformación. La estrategia puede apoyarse en campañas específicas, seminarios, talleres, acciones de gamificación o acciones a través de los medios de comunicación que permitan no solo informar sino también concienciar sobre el problema.
- c. La creación e impulso de redes de expertos.** Para contrarrestar las campañas de desinformación que se puedan producir en un momento dado, es necesario no solo analizar las narrativas que conforman las campañas de desinformación, sino también las tácticas, técnicas y procedimientos utilizados, de modo que facilite generar información veraz sobre el objeto de la campaña. Las áreas de comunicación de las universidades podrían ser los puntos de contacto para identificar y facilitar la participación de estos expertos.
- d. Impulsar la oferta formativa de las universidades.** Uno de los desafíos a los que se enfrentan las universidades y centros de investigación es potenciar su rol formador, así se sugiere revisar la oferta de títulos o materias transversales que permitan avanzar en la lucha contra la desinformación e incrementar la oferta de microcréditos europeos en estas materias.
- e. Promocionar convocatorias competitivas para la financiación de proyectos de investigación que tengan por objetivo la lucha contra la desinformación.** Los diferentes organismos que promueven la investigación científica y técnica en España, como la Agencia Estatal de Investigación, o los que catalizan la relación entre la ciencia y la sociedad, como la Fundación Española para la Ciencia y Tecnología, tienen entre sus cometidos potenciar el trabajo académico y su accesibilidad a la sociedad en su conjunto, sin que se detecten, hasta el momento, ítems significativos destinados específicamente a la lucha contra la desinformación.

Se propone que las universidades y centros de investigación promuevan un conjunto de criterios y líneas estratégicas de los proyectos de investigación que impliquen de forma específica las campañas de desinformación, y que permitan a los grupos de investigación centrar sus esfuerzos desde un punto de vista de la seguridad nacional.

- f. Potenciar la investigación, transferencia y docencia.** Se recomienda profundizar en las facetas principales de los centros de investigación y las universidades articuladas en torno a la investigación, transferencia y docencia. De esta forma, la difusión de la investigación sobre desinformación en un portal

centralizado, la creación de repositorios de datos públicos sobre la investigación y los proyectos financiados sobre esta materia permitirían potenciar el papel investigador de forma estratégica y coordinada.

- g. Incentivar la difusión de la investigación sobre desinformación.** Igualmente, se propone la creación de un sello de calidad de divulgación que pueda articularse a través de la FECYT u organismos estatales.
- h. Financiación de proyectos asociada a la compartición del conocimiento,** contribuyendo así a la creación de un repositorio de proyectos financiados conjuntamente con todos los datos y resultados obtenidos.

DISCUSIÓN Y CONCLUSIONES

A lo largo de este trabajo se ha realizado un profundo análisis sobre las aportaciones realizadas por las universidades y los centros de investigación españoles en la lucha contra la desinformación. Se evidencia, en primer lugar, la preocupación y el esfuerzo que la academia ha realizado en los últimos años para dar respuesta, a través de múltiples aportaciones, como investigaciones y proyectos que permitan su estudio.

España se constituye como el tercer país en el mundo en investigaciones relacionadas. Un liderazgo que se reafirma también a nivel regional pues, en la mayoría de comunidades autónomas, se identifica al menos una universidad con un proyecto o grupo de investigación activo trabajando sobre el fenómeno. Cabe destacar que tales iniciativas parten de los propios investigadores quienes de manera autónoma e independiente buscan financiación para encontrar soluciones de valor.

Gran parte de estos esfuerzos parten además del personal investigador del campo de las ciencias sociales, concretamente del área de Comunicación, disciplina especialmente afectada y que, sin embargo, no está, por ejemplo, representada disciplinalmente en el mayor organismo de investigación nacional: el CSIC.

Asimismo, la financiación es relativamente reducida, aunque se recoge que la AEI ha destinado cerca de 6 millones de euros a proyectos de investigación sobre desinformación.

Por otra parte, se han observado algunas oportunidades que podrían ser valiosas y que, hasta la fecha, presentan poca implementación como la constitución de grupos interdisciplinares y multidisciplinares de distintas áreas de conocimiento o la falta de estudios comparados con otros países. Una financiación superior en cuantía y orientada a la consecución de objetivos específicos contribuiría sensiblemente a desarrollar proyectos que recogieran estas cuestiones. Así, tanto las instituciones como las empresas pueden impulsar la investigación de valor a través del desarrollo de Cátedras y de convocatorias de proyectos financiados.

Además, las universidades y centros de investigación pueden constituirse como redes que contribuyan a contrarrestar la desinformación y minorar su impacto en la sociedad, fomentando el conocimiento y la resiliencia a la injerencia y a la manipulación de la información. Cabe destacar, en este sentido, los esfuerzos crecientes por incentivar la transferencia, es decir, la difusión de sus resultados académicos, que en este caso también son de valor en el diseño de estrategias de seguridad de ámbito nacional.

Finalmente, en esta era de desinformación, resulta necesario ensalzar la labor del experto. Puede que, incluso, toque redefinirlo, pues experto no es aquel capaz de establecer un discurso sobre un fenómeno, sino más bien quien investiga y centra su carrera profesional en darle respuesta. En la creación e implementación de estrategias y medidas activas de lucha contra la desinformación es necesario integrar a los y las académicas españolas, verdaderos expertos en este campo, y cuyo trabajo y liderazgo intelectual en la materia se refleja en los datos aportados en el presente capítulo.

BIBLIOGRAFÍA

- Aguaded Gómez, J.I. y Romero Rodríguez, L.M. (2015). Mediamorfosis y desinformación en la infoesfera: alfabetización mediática, digital e informacional ante los cambios de hábitos de consumo informativo. *Education in The Knowledge Society* (EKS), 16(1), 44-57. <https://doi.org/10.14201/eks20151614457>
- Benaissa Pedriza, S. (2017). Slow journalism in the infoxication era. *Doxa Comunicación*, 25, 129-148. <https://doi.org/10.31921/doxacom.n25a6>
- Blanco, I., García, C., y Tejedor, S. (2019). El impacto de las fake news en la investigación en Ciencias Sociales. Revisión bibliográfica sistematizada. *Historia y Comunicación Social*, 24(2), 449-469. <https://doi.org/10.5209/hics.66290>
- Consejo de la Unión Europea. (2010). *Decisión del Consejo, de 26 de julio de 2010, por la que se establece la organización y funcionamiento del Servicio Europeo de Acción Exterior*. (2010/427/UE). <https://eur-lex.europa.eu/LexUriServ/LexUriServ.do?uri=OJ:L:2010:201:0030:0040:ES:PDF>
- De Pablos, J.M. (2008). El frenesí comunicativo como desinformación. *Comunicar*, 16(31), 173-179. <https://doi.org/10.3916/c31-2008-01-022>
- Del Fresno, M. y López, A. (2014). Social work and netnography: The case of Spain and generic drugs. *Qualitative Social Work*, 13(1), 85-107. <https://doi.org/10.1177/1473325013507736>
- Niño, J. I., Barquero, M., y García, E. (2017). Opinión pública e infoxicación en las redes: los fundamentos de la post-verdad. *Vivat Academia, revista de Comunicación*, 139, 83-94. <https://doi.org/10.15178/va.2017.139.83-94>
- Países Bajos. (2022). *Government-wide strategy for effectively tackling disinformation*. Ministry of the Interior and Kingdom Relations. Directorate-General for Public Administration and Democratic Rule of Law/Democracy and Governance Directorate. <https://www.government.nl/documents/parliamentary-documents/2022/12/23/government-wide-strategy-for-effectively-tackling-disinformation>
- Pamment, J. (2021). *RESIST 2 Counter Disinformation Toolkit*. Government Communication Service (GCS) de Reino Unido. <https://gcs.civilservice.gov.uk/publications/resist-2-counter-disinformation-toolkit/>

Paz-Alonso, P.M., Goodman G.S., e Ibabe I. (2013) Adult eyewitness memory and compliance: effects of post-event misinformation on memory for a negative event. *Behavioral Sciences & the Law*, 31(5), 541-58. <https://doi.org/10.1002/bsl.2081>

Rúas-Araújo, J., y Paniagua-Rojano, F.J. (2023). Aproximación al mapa sobre la investigación en desinformación y verificación en España: estado de la cuestión. ICONO 14. *Revista Científica De Comunicación Y Tecnologías Emergentes*, 21(1). <https://doi.org/10.7195/ri14.v21i1.1987>

Sánchez-Casado, N., Cegarra-Navarro, J.G., y Tomaseti-Solano, E. (2015). Linking social networks to utilitarian benefits through counter-knowledge. *Online Information Review*, 39(2), 179-196. <https://doi.org/10.1108/OIR-12-2014-0307>

Servicio Europeo de Acción Exterior. (2023). *1st EEAS Report on Foreign Information Manipulation and Interference Threats*. Strategic Communications, Task Forces and Information Analysis (STRAT.2). https://www.eeas.europa.eu/eeas/1st-eeas-report-foreign-information-manipulation-and-interference-threats_en

Wardle, C. y Derakhshan, H. (2017). Information disorder: toward an interdisciplinary framework for research and policymaking (DGI(2017)09). Consejo de Europa. <https://rm.coe.int/information-disorder-toward-aninterdisciplinary-framework-for-research/168076277c>

Wojcieszak, M.E. (2011). Computer-Mediated False Consensus: Radical Online Groups, Social Networks and News Media. *Mass Communication and Society*, 14(4), 527-546. <https://doi.org/10.1080/15205436.2010.513795>

Capítulo 5

TÉCNICAS, METODOLOGÍAS Y
PROSPECTIVA IA PARA
COMBATIR LA DESINFORMACIÓN

Coordinadores:

Paolo Rosso

David Camacho Fernández

David Ríos Insua

Presidencia de Gobierno - Departamento de Seguridad Nacional (DSN)

Autores y colaboradores:

Mariluz Congosto Martínez

Borja Díaz-Merry Rivas

Julián Fierrez Aguilar

Luis Felipe Gómez Gómez

Pablo Hernández Escayola

Sergio Hernández García

Raúl Magallón Rosa

Beatriz Marín García

Aythami Morales Moreno

Ángel Panizo Lledot

Francisco Manuel Rangel Pardo

Ruben Tolosana Moranchel

Arkaitz Zubiaga Mendialdua



INTRODUCCIÓN

En los capítulos anteriores de esta monografía se ha discutido ampliamente, y en profundidad, sobre el creciente peligro que representa la desinformación tanto desde un punto de vista de la seguridad nacional de un país como España, como a nivel europeo y global. La desinformación, en su acepción más generalizada, se refiere a la difusión intencional de contenido falso o manipulado con el objetivo de engañar al público y obtener beneficios políticos o económicos. El fenómeno de la desinformación se ha utilizado, entre otros muchos aspectos, para influir en la salud, los sistemas sanitarios o generar opiniones pseudocientíficas entre la población, para influir o generar opinión sobre aspectos sociales como la inmigración, y muy recientemente se ha visto un preocupante aumento del negacionismo sobre el cambio climático. Sin embargo, uno de los aspectos donde la desinformación actúa de manera más regular y de forma particularmente intensa es en el intento de desestabilización de los sistemas democráticos occidentales, habiéndose empleado profusamente para influir, entre otros, en numerosos procesos electorales, algunos de los cuales aquí revisamos. Aunque no es un fenómeno nuevo, su capacidad para llegar a la población general a través de tecnologías modernas como internet, las redes sociales y la Inteligencia Artificial (IA), ha incrementado tanto la cantidad de desinformación generada, la dificultad para poder controlarla (es decir, validarla y poder combatirla mediante hechos contrastados), y la enorme velocidad a la que esta desinformación se propaga alcanzando a millones de personas, expuestas a la misma. Por lo tanto, este capítulo se centrará principalmente en el análisis de estas metodologías modernas y su papel fundamental en la propagación, y la detección, de la desinformación.

En este capítulo comenzaremos presentando la perspectiva de la Comisión Europea respecto al problema de la desinformación, y examinaremos los marcos institucionales que se están desarrollando para combatirlo. A continuación, describiremos los principales modelos de propagación de la desinformación a través de las redes sociales, que se han convertido en el vehículo fundamental para su difusión en la actualidad, para ello, se incluirán ejemplos de plataformas sociales muy populares como Twitter y Telegram. Posteriormente, discutiremos el papel crucial de las plataformas de verificación y cómo los sistemas basados en IA pueden ayudar y respaldar sus actividades. Por un lado, exploraremos cómo la IA

puede ayudar a estas plataformas en su tarea, y, por otro lado, examinaremos cómo esta tecnología también puede ser utilizada por los propagadores de desinformación. Dentro de las técnicas de IA, nos centraremos en el papel fundamental de las metodologías modernas de procesamiento del lenguaje natural. En particular, pondremos énfasis en la generación y detección de noticias falsas (*fake news*) y vídeos manipulados (*deepfakes*). Analizaremos la relevancia de los denominados modelos de lenguaje masivos (del inglés, *Large Language Models*), basados en transformers, en la generación de textos y en la propagación de desinformación. Finalmente, concluiremos con algunas reflexiones acerca de la importancia de este tema y su impacto en nuestra sociedad actual.

ESTUDIO Y MODELADO DE LAS CAMPAÑAS DE DESINFORMACIÓN

En el marco político, en diciembre de 2020 la Comisión Europea presentó el Plan de Acción para la Democracia Europea,¹ que sienta la lucha contra la desinformación como uno de sus pilares centrales para la defensa de la democracia. Este documento habilita al sector público (a través de las instituciones europeas y los Estados miembros de la Unión Europea), al sector privado (mediante los medios de comunicación y las plataformas en línea) y a la sociedad civil para la creación conjunta de medidas específicas para contrarrestar los efectos de la desinformación. La complejidad de esta problemática supone que las medidas para mitigarla están en manos de distintos sectores que deben armonizar sus ámbitos de conocimiento para construir respuestas eficientes.

Para llegar a respuestas coordinadas, es necesario antes que la armonización empiece por una interpretación común de las claves de la problemática. El citado documento de la Comisión Europea menciona distintos fenómenos que suelen incluirse bajo la denominación de «desinformación», evidenciando las distintas dimensiones del concepto y la falta de una terminología común para describirlo. Según el mismo, la *“desinformación es un contenido falso o engañoso que se difunde con intención de engañar o de obtener una ganancia económica o política y que puede causar un perjuicio público”*. Las definiciones propuestas por otros organismos coinciden en enfatizar la «intencionalidad» como el elemento clave de la desinformación. Sin embargo, en términos analíticos, ¿es posible demostrar mediante indicadores objetivos que existe intencionalidad a la hora de distribuir contenido desinformativo?

Siguiendo la misma estela, en 2021 el Servicio Europeo de Acción Exterior (SEAE) fue un paso más allá al añadir en su definición de «Manipulación de la Información e Interferencias del Exterior»² (FIMI, por sus siglas en inglés) matices sobre los elementos que acreditan tal «intencionalidad». De este modo, las manipulaciones informativas se componen de: contenido que mayoritariamente no es ilegal y que amenaza o pone en riesgo valores y procesos políticos; es de carácter manipulativo; y se distribuye de forma intencional y coordinada. Esto conlleva un cambio de paradigma a la hora de analizar desinformación ya que el foco de atención se sitúa en el comportamiento a la hora de producir y distribuir el contenido en lugar de mantenerse únicamente en el análisis del mismo.

¹ <https://eur-lex.europa.eu/legal-content/ES/TXT/HTML/?uri=CELEX:52020DC0790>

² https://www.eeas.europa.eu/eeas/tackling-disinformation-foreign-information-manipulation-interference_en

La desinformación y las manipulaciones informativas son un tipo de amenaza híbrida que va más allá de una simple manipulación narrativa de contenidos. Ocupan un lugar destacado en las agendas de seguridad nacional de los países de la Unión Europea. En España, según el Informe Anual de Seguridad Nacional 2022 del Departamento de Seguridad Nacional, las campañas de desinformación pueden formar parte de estrategias híbridas más amplias que también se sirvan de ciberataques, sabotajes a infraestructuras críticas o guerras energéticas y económicas.³

Así, al emplazar la manipulación informativa en el ámbito de las amenazas híbridas, su análisis debe ser estandarizable e incluir diversos tipos de indicadores para que los resultados sean extensibles y extrapolables a otros miembros de la comunidad analítica y de otras disciplinas interconectadas. El marco *ABCDE (ABCDE Framework)*⁴ propone que el análisis de la desinformación aporte información relativa a cinco elementos clave: el actor o la entidad responsable del incidente; el comportamiento del incidente y las técnicas de manipulación informativa; el contenido e información utilizados; el grado de distribución del contenido y sus audiencias potenciales; y el efecto e impactos del incidente.

Otro marco para avanzar en la estandarización del modelo de análisis del comportamiento en ataques desinformadores es el *DISARM Framework*.⁵ La disciplina de la ciberseguridad aporta algunas claves sobre cómo analizar incidentes por fases evolutivas y describir patrones de comportamiento. *DISARM* es un catálogo de fuente abierta para analizar manipulaciones informativas y tipos de respuesta inspirado en el *MITRE ATT&CK*⁶ de ciberseguridad. Al utilizar *DISARM* el incidente de desinformación se disecciona por fases para investigar qué tácticas, técnicas y procedimientos (TTP) han sido empleados para manipular el espacio informativo. En la fase de planificación, se examinan los objetivos y la estrategia utilizada en la campaña de desinformación. El análisis de la fase de preparación describe el desarrollo del contenido, las narrativas, la infraestructura y los canales de distribución previos al ataque. La etapa de ejecución detalla las técnicas de publicación, amplificación, instrumentalización y perduración del contenido en línea. Por último, la fase de análisis evalúa los efectos del incidente.

La descripción de TTP catalogadas en *DISARM* permite una armonización del análisis con respecto al de otros miembros de la comunidad. Adicionalmente, el objetivo final es el intercambio de información y la agregación del conocimiento producido desde diferentes ámbitos. Para asegurar la completa compatibilidad e interoperabilidad del análisis, es necesario que los resultados de las investigaciones se traduzcan a un formato de datos estructurados. Consecuentemente, el producto de los estudios debe estar expresado en

³ <https://www.dsn.gob.es/es/documento/informe-anual-seguridad-nacional-2022> (Pág. 10)

⁴ <https://carnegieendowment.org/2020/09/24/eu-s-role-in-fighting-disinformation-crafting-disinformation-framework-pub-82720>

⁵ <https://www.disarm.foundation/framework>

⁶ <https://attack.mitre.org/>

un lenguaje estándar para la compartición de información (*Data Sharing Standard*). Una vez más, el ámbito de la ciberseguridad ofrece una respuesta. El modelo STIX⁷ (*Structured Threat Information eXpression*) permite la codificación de elementos de una investigación de desinformación mediante objetos. La información que se codifique dentro de cada objeto es agregable a los resultados obtenidos en investigaciones previas y asegura la interoperabilidad con el análisis de otros miembros de la misma comunidad. El informe *Foreign Information Manipulation Interference (FIMI) and Cybersecurity*,⁸ publicado por ENISA y el SEAE de la Unión Europea, profundiza en las similitudes y especificidades de los enfoques de análisis de ciberseguridad y FIMI. STIX también permitiría la interoperabilidad entre el análisis de manipulación informativa y la ciberseguridad.

En febrero de 2023, se aportó una primera aproximación a la metodología de análisis descrita en este texto en el Primer Informe del SEAE sobre amenazas de manipulación de la información e injerencia por parte de agentes extranjeros: Hacia la construcción de un marco de defensa en red.⁹

En vista de todo lo anterior, el análisis de desinformación centrado en el comportamiento y el contenido permite mejorar el reconocimiento de patrones de conducta desde fases tempranas del ataque. Por tanto, posibilita contra reacciones más rápidas y tempranas (modelo de killchain). Los catálogos de respuestas descritos por *DISARM* (con el *DISARM Blue*), o por el Consejo Europeo¹⁰ (con herramientas para hacer frente a amenazas híbridas interoperables con herramientas específicas frente a la desinformación como la FIMI toolbox) son algunos de los mecanismos de respuesta que se podrían activar a través de la implementación de este tipo de análisis.

Además, un mejor conocimiento de los patrones de ataque permite optimizar los sistemas de detección temprana basados en el comportamiento. La IA aporta herramientas clave para la verificación de contenido falso o fabricado. A través de la sistematización del análisis del comportamiento, la IA también podría mejorar los sistemas de detección temprana automática basados en patrones de comportamiento. Por ejemplo, evaluando pautas anormales en las fases de preparación y ejecución del incidente, como se especificará más adelante.

⁷ <https://oasis-open.github.io/cti-documentation/stix/intro.html>

⁸ <https://www.enisa.europa.eu/publications/foreign-information-manipulation-interference-fimi-and-cybersecurity-threat-landscape>

⁹ https://www.eeas.europa.eu/eeas/1st-eeas-report-foreign-information-manipulation-and-interference-threats_en

¹⁰ <https://data.consilium.europa.eu/doc/document/ST-11429-2022-INIT/en/pdf>

CARACTERIZACIÓN DE REDES Y PROPAGACIÓN DE LA DESINFORMACIÓN

Propagación de desinformación en redes sociales

En la actual era digital, las redes sociales se han convertido en herramientas poderosas para compartir información y conectarse con otros. No obstante, también han surgido nuevos retos en el ámbito de la información, puesto que estas mismas redes han sido el escenario de la propagación masiva de desinformación. En este sentido, la propagación de la desinformación en redes sociales ha adquirido una relevancia significativa, lo que evidencia la necesidad de analizar los procesos de diseminación de información engañosa. En este contexto, se puede establecer una analogía entre la propagación de la información, ya sea verídica o no, en redes sociales y la propagación de un virus o enfermedad en una población. En el ámbito físico, la interacción entre dos personas puede conllevar un cierto riesgo de contagio, en caso de que una de ellas sea portadora de una enfermedad. De manera análoga, en el contexto de las redes sociales, la interacción entre dos cuentas puede conllevar cierto riesgo de contagio de información de una cuenta a otra en caso de que una de ellas sea portadora de dicha información. De hecho, en la literatura es común encontrar variaciones de modelos epidemiológicos, utilizados para describir la propagación de enfermedades infecciosas en una población, para modelizar la difusión de noticias falsas en redes sociales (Jin et al., 2013). Esto es posible debido a que, de acuerdo con los experimentos de Vosoughi et al. (2018), las noticias falsas se difunden más fácilmente y más rápido que las verdaderas. En dicho estudio los autores concluyen que el 1% de las noticias falsas que más se difundieron llegaron a incidir entre 1.000 y 100.000 personas, mientras que el 1% que más se difundió de las verdaderas rara vez alcanzaron a más de 1.000 personas. Siguiendo la analogía de los virus, los estudios muestran que las noticias falsas son mucho más virulentas que las noticias verdaderas y son capaces de contagiar a más personas más rápidamente. Esto ha llevado a varios investigadores a aprovechar estos patrones de difusión tan distintos entre noticias falsas y verdaderas para desarrollar filtros de detección rápida de noticias falsas (Jin et al., 2013).

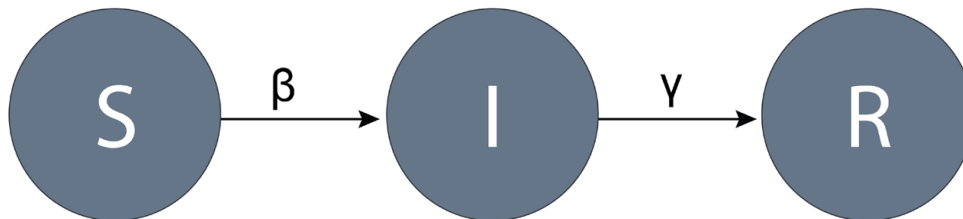


Figura 1. Esquema del modelo SIR (Susceptible, Infectada, Recuperada). Los círculos representan los distintos grupos de población y las flechas la probabilidad de que un individuo de una población pase a otra. Así, hay una probabilidad β de que un individuo (S)usceptible pase a (I)nfectado y una probabilidad γ de que un (I)nfectado pase a (R)ecuperado

Aunque los modelos epidemiológicos son útiles para comparar patrones de difusión de la información, tienen ciertas limitaciones, principalmente por ser anónimos. Es decir, no son capaces de modelizar quién transmite la enfermedad. Especifiquemos este concepto de anonimidad con un ejemplo sencillo del modelo SIR (véase la figura 1). En los modelos epidemiológicos la población se divide en ciertos grupos. En el SIR, se hace en tres grupos: (i) (S)usceptible, personas sensibles a contagiarse con un virus (o pieza de información); (ii) (I)nfectados: personas que han sido expuestas a un virus y se han contagiado (o a una pieza de información y la han asimilado, no es necesario que se hayan formado una opinión sobre ella); (iii) (R)ecuperados: personas infectadas que se han recuperado de una infección y no son susceptibles de contraerla de nuevo, ya sea por vencer al virus o fallecer (en nuestra analogía con piezas de información, serían aquellas personas que se han visto expuestas a cierta información y la han descartado o interiorizado, de forma que aunque se vean expuestos a la misma de nuevo, su manera de pensar no se ve afectada). Además de estos grupos de población, el modelo SIR define dos parámetros: β (probabilidad de que un individuo susceptible se contagie); γ (probabilidad de que un individuo contagiado se recupere). Una vez ajustado el modelo, es decir, estimados los valores β y γ que mejor se adaptan a los datos observados, somos capaces de predecir cuántas personas habrá en cada uno de los grupos en un momento dado. Sin embargo, no somos capaces de saber qué personas específicas estarán en cada grupo. Empleando un modelo epidemiológico es imposible saber quiénes son los agentes encargados de difundir una pieza de información y, por lo tanto, resulta imposible saber cómo actuar ante ellos. En consecuencia, cuando se esté interesado en analizar este tipo de información es necesario utilizar modelos más complejos, como los basados en redes latentes de difusión (Guille et al., 2013).

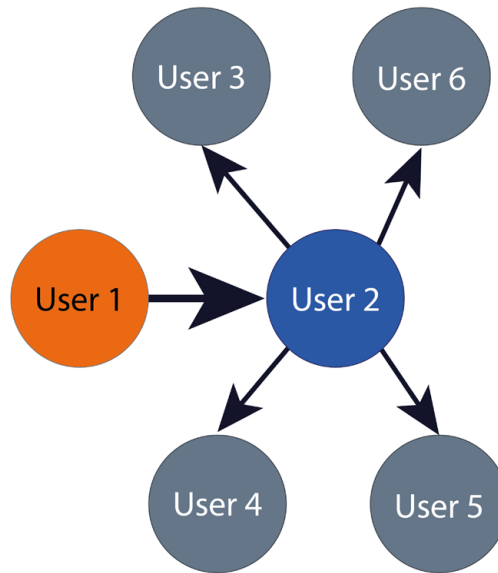


Figura 2. Ejemplo de red latente de difusión con seis usuarios. Las flechas indican la dirección de flujo de la información; su grosor es proporcional a la intensidad de la influencia. Por ejemplo, "User 1" es capaz de influir con más intensidad sobre "User 2" comparado con la capacidad de influencia que tiene "User 4" sobre "User 5"

Una red latente de difusión es un modelo generativo de redes sociales que permite modelizar la difusión de información entre los individuos de una red a lo largo del tiempo. A diferencia de los modelos epidemiológicos, las redes latentes de difusión no solo permiten predecir cómo evolucionará esa propagación, sino que también permiten conocer quién la propaga y cómo lo hace. La Figura 2 contiene un ejemplo de una red latente de difusión con 6 usuarios. En la red se distinguen mediante colores tres tipos de usuarios. El usuario naranja es el que controla el discurso dentro de la red, pero no lo hace directamente, sino que para ello emplea al usuario azul, sobre el que tiene mucha influencia (obsérvese el tamaño de la flecha que une a ambos usuarios). Al usuario azul, encargado de difundir la información a través de la red, se le podría considerar un *influencer*. Por último, los usuarios grises se pueden considerar como receptores, ya que su capacidad de influir en otros es limitada y solo reciben influencias externas. Apréciase el tamaño de las flechas que apuntan hacia fuera de los nodos grises en comparación con los que apuntan hacia dentro.

Ajustar un modelo de influencia, ya sea epidemiológico o basado en redes, dista de ser sencillo. La desinformación se distribuye a través de diversas plataformas con diferentes niveles de privacidad, lo que hace imposible analizar legalmente todo lo que se difunde y

quiénes están involucrados. En el mundo real, es muy difícil, casi imposible, observar todos los sucesos que provocan que una persona se exponga a un tipo concreto de información o de datos. Lo más normal es que solo se puedan observar algunos de estos sucesos mientras que otros permanecen ocultos. Por ejemplo, si analizamos una red social se puede saber si, en un momento dado, ciertos individuos han mencionado o no un tema concreto. Cuando se trabaja con el modelo epidemiológico, por ejemplo, SIR, es muy fácil calcular el número de individuos en el grupo “Infectado”, pero es muy difícil saber el número de usuarios en el resto de los grupos. Obsérvese que, si un usuario no ha mencionado el tema, a priori, no es posible conocer si es porque lo vio y no le interesa (grupo Recuperado) o porque aún no lo ha visto (grupo Susceptible). De igual manera, cuando se trabaja con redes de difusión, es relativamente sencillo identificar cuando dos usuarios mencionan un mismo tema y, además, saber si uno lo hizo antes que el otro. Sin embargo, es muy difícil discernir si realmente el usuario que mencionó el tema primero ha influido en el otro o si, por el contrario, es una coincidencia y ambos usuarios vieron el tema en el periódico o en la televisión y por eso lo mencionan. Debido a que se tiene que trabajar con información parcial, para poder ajustar un modelo con una precisión razonable son necesarios grandes volúmenes de datos, en especial cuando se pretende trabajar con redes de difusión. Dado que estas redes eliminan el componente anónimo de los modelos epidemiológicos, se requiere disponer de suficiente información sobre todas y cada una de las cuentas incluidas en la red bajo estudio. No obstante, debido a la naturaleza de las redes sociales, donde las cuentas se crean y destruyen con frecuencia, conseguir información tan completa puede resultar difícil y es muy dependiente de cada plataforma.

Aun teniendo un modelo de influencia perfectamente calibrado, diferenciar un flujo de la información natural de uno anómalo no es trivial y deben utilizarse técnicas de análisis de redes complejas que permitan identificar la estructura de las relaciones dentro de la red, los actores más influyentes, las comunidades, la densidad de conexiones, entre otros aspectos relevantes. El proceso de propagación de la información dentro de una red social es un fenómeno que se da de manera natural. Cuando el flujo funciona de manera natural, las cuentas se ven expuestas a contenido que se ajusta a sus intereses y preferencias personales. Por otro lado, también existen flujos de información anómalos creados por agentes maliciosos que abusan del sistema, como troles (cuentas que solo buscan crear conflicto), *bots* (cuentas completamente automatizadas), *cyborgs* (cuentas parcialmente automatizadas) o cuentas falsas (que se hacen pasar por personas que no son, por ejemplo, autoridades o políticos). Estos agentes malintencionados manipulan el flujo de (des)información en las redes sociales para difundir desinformación, propaganda y noticias falsas. A menudo, tales agentes tienen una agenda política o ideológica y su objetivo es influir sobre la opinión pública y manipular la información que se comparte en línea. Sin embargo, detectar estos agentes maliciosos y sus campañas no es tan sencillo como limitarse a encontrar cuentas *bots* o *cyborgs*. En el caso de los bots, estos han estado presentes en la conversación pública sobre numerosos procesos electorales (Pastor-Galindo et al., 2020), o la pandemia de la COVID-19 (Antenore et al., 2022). Sin embargo, la detección de estas cuentas es un ejercicio complejo debido a su continua evolución para evadir su identificación (Cresci, 2020). Shao et al. (2018) muestran su relevancia en la propagación de contenido

con “baja credibilidad”, y argumenta que predecir el grado de automatización de las cuentas en distintas plataformas puede ayudar a combatir la propagación de este contenido. Sin embargo, hay que tener en cuenta que estas cuentas no son las únicas responsables de propagar desinformación, dado que las cuentas humanas también juegan un papel relevante en esto (Vosoughi et al., 2018). En consecuencia, diferenciar si una campaña en redes sociales se debe a un Movimiento de base natural o, por lo contrario, a una campaña de *astroturfing*¹¹ (Zhang et al., 2013) es extremadamente complejo, especialmente, si tenemos en cuenta la aparición de modelos de IA generativa como ChatGPT¹² (para generación de texto) y *StableDiffusion* (para generación de imágenes) y la popularización de la “economía de micro-encargos”, *gig economy* en inglés (Vallas y Schor, 2020). En la economía de micro-encargos, las personas trabajan como contratistas independientes en proyectos o encargos a corto plazo para varios clientes. Este modelo económico se ha popularizado en las últimas décadas gracias a la masificación de internet y a la aparición de portales como *Fiverr*,¹³ *Amazon Mechanical Turk*¹⁴ o *TaskRabbit*.¹⁵ Debido a la baja inversión inicial; la facilidad de acceso a trabajadores de todas partes del globo; las comodidades que dan los modelos generativos de IA para generar contenido creíble de manera sencilla; y, por último, la relativa anonimidad que proporcionan los trabajos extra cortos, la economía de micro-encargos se ha convertido en un campo fértil para realizar campañas de *astroturfing* (Lee et al., 2015) y manipulación en redes sociales.

Ejemplos de difusión de desinformación en Twitter y Telegram

Los rumores y noticias falsas saltan de una red social a otra, lo que nos permite examinar lo que sucede en plataformas de acceso público, como son Twitter o los canales de Telegram. Como ya hemos indicado anteriormente, la información que se puede obtener de cada red social varía mucho. Así, por ejemplo, la información que se puede obtener de Twitter y los canales de Telegram es bastante diferente. En Twitter, existen relaciones estáticas que cambian poco, como seguir o dejar de seguir a otros perfiles, y también existen interacciones dinámicas de distintos tipos: dos positivas (favoritos y retuits) y dos ambiguas (comentarios y citas). Por otro lado, en los canales de Telegram no hay relaciones estáticas, pero sí interacciones dinámicas al referenciar a otros canales. En esta subsección

¹¹ El *astroturfing* es una estrategia de manipulación y condicionamiento de la opinión pública que consiste en difundir mensajes de forma artificial a través de cuentas anónimas o que carecen de transparencia, y que se utilizan para interferir en una discusión orgánica o iniciarla, pero intentando dar la apariencia de espontaneidad y naturalidad, para que los receptores lo perciban como un debate social auténtico y orgánico. Suelen buscar respuestas exacerbadas y polarizar el debate online.

¹² <https://openai.com/blog/chatgpt>

¹³ <https://www.fiverr.com>

¹⁴ <https://www.mturk.com/>

¹⁵ <https://www.taskrabbit.com/>

mostramos una metodología que se emplea para estudiar la estructura y las interacciones entre los actores de un sistema. En concreto, se aplicarán técnicas de análisis de redes complejas para estudiar las conexiones estáticas y dinámicas entre los usuarios de Twitter, así como las conexiones dinámicas entre canales de Telegram para varios casos de difusión de desinformación.

Los canales de Telegram funcionan como lugares donde se comparte información de interés para el grupo y se planifican acciones que tendrán visibilidad en otras plataformas. Conocer estos grupos, su estructura y contenido, puede ayudar a anticiparse a sus acciones y contrarrestar sus campañas. Esta información es crucial para entender mejor la desinformación que se difundirá por otras plataformas. Por otro lado, el estudio de la propagación de la desinformación en Twitter puede realizarse a través de las conexiones estáticas, dinámicas o una combinación de ambas. Las conexiones estáticas permiten observar la estructura de la red por la cual los mensajes pueden potencialmente difundirse. Dependiendo de la afinidad de los miembros de la red con respecto a un tema, la propagación será más rápida o extensa. A veces, un solo perfil con pocas conexiones dentro de una red compacta se basta para difundir ampliamente un bulo. En el siguiente ejemplo,¹⁶ ilustrado en la Figura 3, se aprecia cómo el perfil @yontanora, que tenía pocas conexiones estáticas, logró difundir una mentira sobre la distribución de las mascarillas, obteniendo una amplia difusión que fue más allá de sus contactos directos.

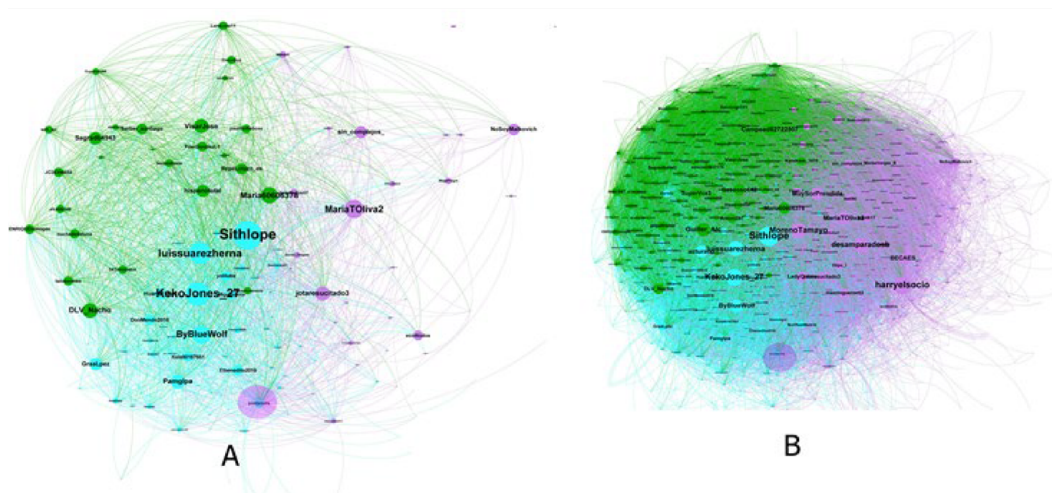


Figura 3. Estructura de conexiones estáticas del perfil @yontanora (A) y estructura de conexiones estáticas de perfiles implicados en la difusión de un bulo emitido por @yontanora (B)

¹⁶ Difusión de un bulo sobre la distribución de las mascarillas en el Metro de Madrid durante la pandemia <https://twitter.com/congosto/status/1249973668341592066>

La propagación de la desinformación no siempre se limita a redes compactas afines a una temática, sino que puede saltar de una comunidad a otra. Un ejemplo es el bulo de la viruela del mono¹⁷ ilustrado en la Figura 4, en el que un usuario afirmó haberla contraído tras comprar un patinete eléctrico por Wallapop. El bulo se inició en un grupo relacionado con *realities* (A), pasó a un grupo de divulgadores científicos que lo creyeron cierto (B) y, finalmente, fue difundido por activistas de Podemos (C).

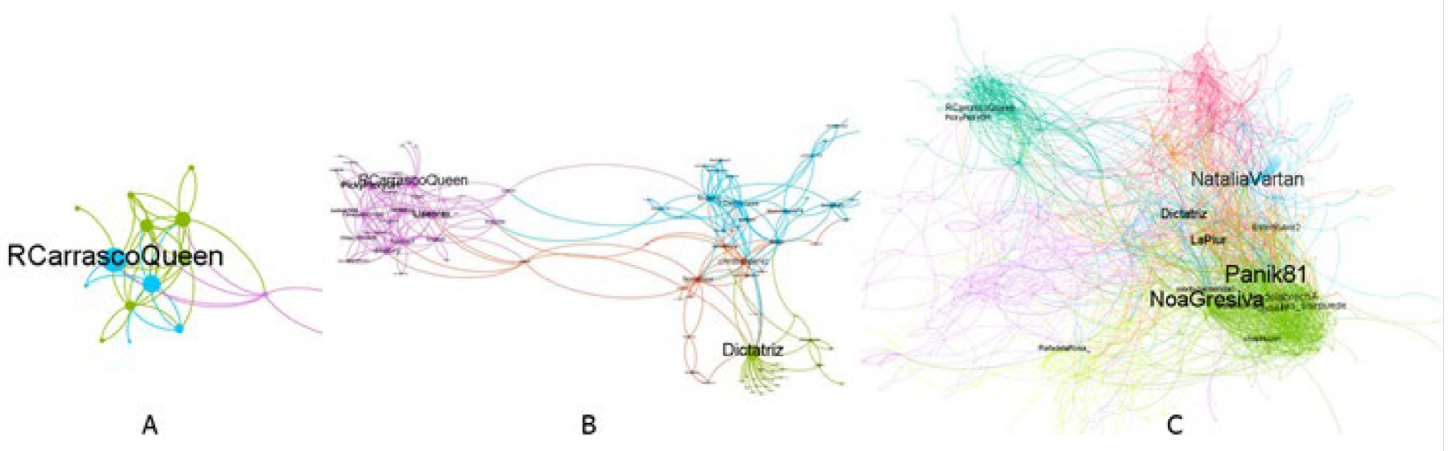


Figura 4. Conexiones estáticas de los primeros difusores del bulo de la viruela del mono: 20 primeros (A), 100 primeros (B) y 1.000 primeros (C)

Aunque las conexiones estáticas son importantes para la difusión de la desinformación, a veces esta se realiza a través de otras vías, como la participación de perfiles con muchos seguidores, lo que eclipsa el efecto de propagación de la red. Este es el caso del bulo del palo de golf,¹⁸ ilustrado en la Figura 5, que se difundió en plena pandemia a través de un vídeo en el que se podía ver a un manifestante del barrio de Salamanca de Madrid golpeando una señal de tráfico con una escoba, y en el texto del tuit se afirmaba que se trataba de un palo de golf. La gran visibilidad que obtuvo hizo que se convirtiera en tendencia durante varios días y que algunos medios de comunicación le llegaran a dar credibilidad.

¹⁷ Difusión inicial del bulo del contagio de la viruela del mono tras comprar un patinete en Wallapop. <https://twitter.com/congosto/status/1554147500298313728>

¹⁸ Difusión del bulo del palo de golf en las manifestaciones del barrio de Salamanca de Madrid <https://twitter.com/congosto/status/1261952277142790144>

tweets_vs_alance

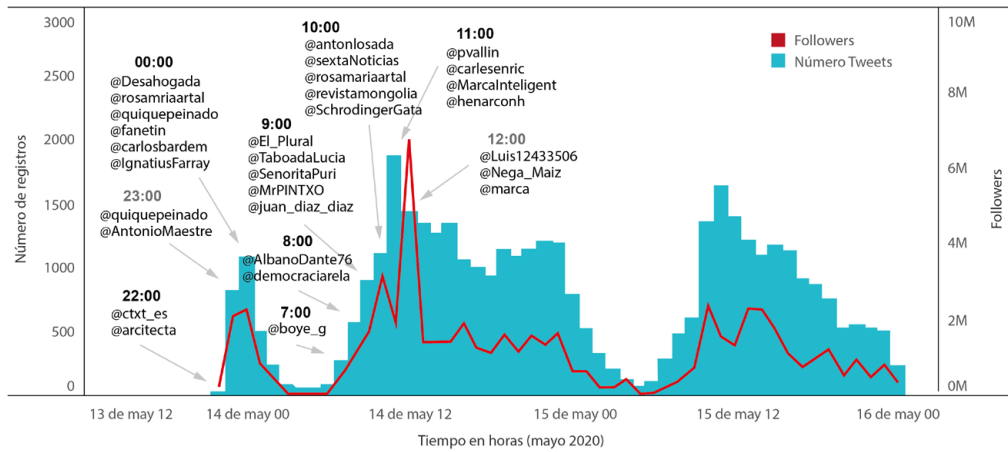


Figura 5. Difusión del bulo del palo de golf en las manifestaciones del Barrio de Salamanca (Madrid) en la que participaron perfiles con muchos seguidores

EL ROL DE LAS PLATAFORMAS DE VERIFICACIÓN

La IA, a falta de conocer mejor sus limitaciones, aporta ya un conjunto de herramientas muy poderosas en numerosos ámbitos, desde la aceleración de procesos cuya ejecución llevaba numerosos años, como en el descubrimiento de fármacos, a la introducción de tecnologías radicalmente nuevas, como la de los vehículos autónomos. Pero también trae importantes inconvenientes. En particular, en el dominio que nos ocupa en este volumen, si la IA se emplea para generar desinformación pueden multiplicar la cantidad, la verosimilitud y la eficacia de los contenidos falsos que se utilizan para manipular a los ciudadanos. Al mismo tiempo, también abre nuevas posibilidades para que las organizaciones de verificación, o de *fact-checking*, mejoren sus procesos de trabajo y puedan ser más efectivos en la detección y desmentidos de contenidos desinformadores. De hecho, a medida que los sistemas basados en IA permitan a los actores desinformadores generar contenidos cada vez más difíciles de distinguir de los reales, más importante será la ayuda de sistemas que incorporen IA para que los verificadores puedan detectarlos y desmentirlos.

Sin embargo, pese a las nuevas posibilidades que ofrece la inteligencia artificial a la verificación, estos sistemas no pueden sustituir la labor de los *fact-checkers*. Gracias a estas herramientas se pueden automatizar procesos y aumentar la precisión en las verificaciones, pero el elemento humano sigue siendo indispensable.

Tecnología y verificación. Hacia los procesos de automatización

Los verificadores ya están probando sistemas que incorporan herramientas de IA para mejorar su trabajo en varias direcciones. Una de ellas es detectar contenidos que pueden incluir desinformación, haciendo rastreos automatizados de afirmaciones desmentidas. También se están empleando para detectar patrones que aparezcan en contenidos falsos o que caractericen narrativas desinformadoras y con los que se pueden generar sistemas de alerta temprana. Otra forma en la que se está incorporando la IA al trabajo de los verificadores tiene que ver con la organización interna. La automatización de las bases de datos para ordenar todos los contenidos reduce mucho trabajo manual, ahorra tiempo y enriquece la calidad de tales conjuntos de datos.

El inicio de estas pruebas se remonta, al menos, a 2015, según se expuso en la cumbre mundial de verificación Global Fact 9, celebrada en Oslo en junio de 2022. En el panel para abordar el futuro de la verificación automática, se presentaron avances en la utilización de la IA para la transcripción de textos, la detección de contenidos, el análisis de vídeos y la localización inmediata de refutaciones a afirmaciones falsas.

Aunque se relataron algunas dificultades para hacer viables las herramientas, también se constataron progresos prometedores, como el señalado por Rubén Míguez, director tecnológico de Newtral que estimó que, gracias a la IA, su redacción había conseguido ahorrar un 90 % del tiempo dedicado a la monitorización de contenidos verificables.¹⁹

IA y verificación

Aunque todavía no ha llegado la avalancha de contenidos manipulados con IA que algunos actores auguran —Sam Altman, director ejecutivo de OpenAI,²⁰ la empresa que ha desarrollado ChatGPT, ha manifestado su preocupación por el hecho de que “estos modelos puedan usarse para la desinformación a gran escala”²¹ y la ONG Reporteros Sin Fronteras ha alertado del impacto de la IA en lo que califica como “la industria del engaño”—,²² sí hay ejemplos que ya se han empezado a detectar y que están siendo desmentidos por los verificadores.

En la mayor parte de los casos, por el momento, se trata de *deepfakes*: vídeos, imágenes o audios editados con IA. Un ejemplo muy claro lo vimos durante la primera ola de desinformación que siguió a la invasión rusa de Ucrania. Así, circuló un vídeo manipulado en el que el presidente ucraniano Volodímir Zelenski²³ pedía la rendición de sus tropas. Como respuesta, apareció otro vídeo en el que el presidente ruso Putin anunciaba la rendición de Rusia.²⁴ Sin embargo, no eran manipulaciones muy elaboradas.

Muy poco después empezaron a aparecer aplicaciones que están a disposición de cualquier usuario, como la popular Reface,²⁵ que consiguen resultados técnicamente mucho mejores. A raíz de esta evolución tecnológica (véase el apartado 5 referido a deepfakes) se han viralizado otras manipulaciones como la de Elon Musk afirmando que iba a comprar Meta.²⁶

¹⁹ Newtral: “Por qué la inteligencia artificial es un aliado de los verificadores y puede ayudar en la lucha contra la desinformación”, 24 de junio de 2022. <https://www.newtral.es/periodismo-inteligencia-artificial-avances-newtral/20220624/>

²⁰ <https://openai.com/>

²¹ The Guardian: “We are a little bit scared’: OpenAI CEO warns of risks of artificial intelligence”, 17 de marzo de 2023. https://www.theguardian.com/technology/2023/mar/17/openai-sam-altman-artificial-intelligence-warning-gpt4?CMP=Share_iOSApp_Other

²² Reporteros Sin Fronteras: “Clasificación 2023 | Análisis general: Los peligros de ‘la industria del engaño’” <https://www.rsf-es.org/clasificacion-2023-analisis-general-los-peligros-de-la-industria-del-engano/>

²³ <https://maldita.es/malditateexplica/20220317/deepfake-zelenski-ucrania-tropas-rindan/>

²⁴ <https://maldita.es/malditobulo/20220321/putin-rendicion-rusia-video/>

²⁵ <https://reface.ai/>

²⁶ <https://maldita.es/malditobulo/20221110/musk-comprar-meta-deepfake/>

No solo hemos visto casos de deepfakes. También han comenzado a aparecer fotografías muy convincentes sobre hechos que jamás sucedieron elaboradas con IA. Un caso claro es el de la falsa detención de Donald Trump.²⁷ A pesar de que su creador avisó del carácter ficticio de las imágenes, estas se viralizaron y algunos internautas las tomaron por reales. No es el único ejemplo. Del propio Trump circularon otras fotografías ficticias tras su comparecencia ante un juez de Nueva York.²⁸ También hay una falsa fotografía de antidisturbios abrazando a manifestantes durante una protesta en Francia,²⁹ imágenes creadas artificialmente del papa Francisco con un llamativo abrigo³⁰ o saludando a un imaginario sacerdote satánico,³¹ así como de Vladimir Putin arrodillado ante Xi Jinping.³²

Sobre el impacto que pueden tener estas invenciones, cabe destacar cómo una falsa explosión en el Pentágono estadounidense elaborada con IA y compartida ampliamente en redes sociales a finales de mayo de 2023 motivó una leve caída en la bolsa de Wall Street.³³

En la creación de imágenes falsas se ha podido comprobar la rápida evolución de las herramientas que usan la IA. Cuando se empezaron a difundir este tipo de fotografías había algunos detalles que las delataban con facilidad³⁴, como los problemas para generar manos humanas con cinco dedos, pero las últimas versiones de estas herramientas van corrigiendo esos defectos y, cada vez, son más precisas.

También se han empezado a detectar casos de uso de la IA en campañas desinformadoras que tratan de viralizar una narrativa concreta usando contenidos que simulan ser artículos periodísticos, como se denunció³⁵ tras el descarrilamiento de un tren en Ohio (Estados Unidos) en febrero de 2023.

²⁷ <https://maldita.es/malditobulo/20230322/fotos-trump-detenido-inteligencia-artificial/>

²⁸ <https://www.rtve.es/noticias/20230405/estas-fotos-trump-nueva-york-tras-tribunal-son-falsas/2436287.shtml>

²⁹ <https://maldita.es/malditatecnologia/20230217/imagen-manifestacion-francia-falsa-inteligencia-artificial/>

³⁰ <https://maldita.es/malditobulo/20230327/papa-francisco-abrigo-plumas-bulo-inteligencia-artificial/>

³¹ <https://verifica.efe.com/papa-francisco-sacerdote-satanico-inteligencia-artificial/>

³² <https://maldita.es/malditobulo/20230329/putin-xi-jinping-reverencia-imagen-falsa-inteligencia-artificial/>

³³ <https://verifica.efe.com/imagenes-explosion-pentagono-creadas-con-inteligencia-artificial-ia/>

³⁴ <https://maldita.es/malditatecnologia/20230217/imagen-manifestacion-francia-falsa-inteligencia-artificial>

³⁵ https://weaponizedspaces.substack.com/p/ai-reporters-spread-disinformation?r=1aupz&utm_campaign=post&utm_medium=web

Perspectivas de futuro y recomendaciones

La amenaza de la IA para la desinformación

La IA es una herramienta muy poderosa para procesar contenidos, extraer patrones, hacer predicciones y generar nuevos contenidos a partir de lo que va aprendiendo. Son características que permiten crear desinformación verosímil y potencialmente muy peligrosa de forma masiva.

Debe tenerse en cuenta que el funcionamiento de un sistema basado en IA depende de cómo se haya entrenado esa herramienta. Un ejemplo de los peligros que puede conllevar una IA entrenada sin preocuparse de que diferencie claramente entre desinformación e información real lo aporta una investigación de NewGuard, de enero de 2023,³⁶ que comprobó que ChatGPT (al menos la versión disponible entonces) respondía con afirmaciones elocuentes pero engañosas o directamente falsas en un 80 % de las veces cuando se le preguntaba sobre narrativas desinformadoras (véase el apartado 6).

Los modelos que ya existen son capaces de crear fotografías, vídeos y textos muy difíciles de distinguir de los creados por humanos, con una mejora técnica constante. Esto supone un peligro claro porque gracias a esas herramientas basadas en la IA, los actores desinformadores pueden inundar el espacio digital de contenidos manipulados para engañar a los ciudadanos. Así, cada vez será más difícil para ellos encontrar pistas que les hagan sospechar de que se trata de un contenido falso. Pero hay un peligro derivado: si los ciudadanos no son capaces de diferenciar lo verdadero de lo falso, pueden acabar desconfiando de todo, incluso de los contenidos precisos y de las fuentes fiables.

Por ello, sería conveniente exigir que las herramientas de IA generen de forma automática una marca que indique claramente que ese contenido ha sido creado artificialmente. Asimismo, es necesario reiterar la importancia de la alfabetización mediática para concienciar a la población de los riesgos que supone la difusión de tales contenidos.

³⁶ <https://www.newsguardtech.com/misinformation-monitor/jan-2023/>

Cómo la IA puede ayudar a la verificación

La cantidad de información y de contenidos disponible en internet es totalmente inabarcable para un ser humano, depende del filtro de las plataformas tecnológicas y aumenta cada minuto. Por eso, para monitorizar qué está circulando en las redes sociales y detectar posibles peligros de manera cada vez más eficaz es fundamental incorporar sistemas basados en IA. Gracias a ellos se puede realizar un filtrado de grandes cantidades de contenidos y detectar dónde están y cómo se difunden los contenidos desinformativos. Además, la inteligencia artificial también puede dar sentido a los datos que analiza. Puede encontrar patrones de desinformación que se escapan de los seres humanos y usarse para poner en marcha sistemas de alerta temprana. También puede ayudar a detectar y entender mejor las campañas coordinadas de desinformación.

Además, si la IA se empieza a utilizar masivamente para desinformar, una de las principales herramientas para detectar esos contenidos manipulados tendrá que ser también la IA. Si esta tecnología es capaz de crear imágenes, vídeos o textos falsos prácticamente indistinguibles de los reales, una de las mejores opciones para identificar tales contenidos será, precisamente, el empleo de sistemas basados en IA capaces de detectar los rasgos que diferencian esas creaciones ficticias de las reales.

Sin embargo, aunque la IA va a ser esencial en el futuro de la verificación, eso no quiere decir que puedan sustituir a los *fact-checkers*. La desinformación depende en muchas ocasiones del contexto y de una serie de matices y circunstancias que una máquina no puede evaluar adecuadamente. Además, la IA no tiene una noción de realidad, pues se construye a partir de los datos con los que se la entrena. El sesgo en ese entrenamiento puede llevar a resultados equivocados a la hora de evaluar la fiabilidad de un contenido que aparezca en redes sociales.

Los verificadores asumen unos compromisos de transparencia e independencia en sus actuaciones y siguen una metodología para asegurar que sus calificaciones señalan correctamente los contenidos desinformativos o manipuladores. Esas garantías no se cumplen cuando se habla de procesos de *fact-checking* que quedan en manos únicamente de sistemas basados en IA o de sistemas colaborativos de verificación ciudadana.

TÉCNICAS Y METODOLOGÍAS DE PROCESAMIENTO DEL LENGUAJE NATURAL

Ante la creciente presencia de información en medios como blogs y redes sociales, ha aumentado la necesidad de desarrollar técnicas de detección automática de desinformación que puedan asistir las labores humanas tanto de verificadores como de usuarios finales (Konstantinovskiy et al., 2021). Una de las tecnologías principales que se utiliza con el fin de asistir a usuarios con el proceso de verificación de contenido mediante sistemas automatizados es el procesamiento del lenguaje natural (PLN) o, en inglés, *natural language processing* (NLP). Las tecnologías PLN permiten el procesamiento automático de texto con diferentes finalidades, entre las cuales se encuentra la posibilidad de clasificar un texto como verdadero o falso a partir de un conjunto de criterios predefinidos. Para que un sistema PLN aprenda a determinar si un texto es verdadero o falso, en primer lugar, habrá que entrenarlo con un conjunto de ejemplos etiquetados como tales por humanos. Estos ejemplos servirán al sistema de PLN para aprender los patrones que conforman ambas clases de ejemplos, y se tratará de aplicar estos patrones en nuevos ejemplos textuales para determinar si son verdaderos o falsos. Este tipo de entrenamiento de sistemas a partir de datos previamente etiquetados por humanos se conoce como aprendizaje supervisado (Hastie et al., 2009).

A pesar del gran avance en los últimos años en el uso de técnicas PLN, estos sistemas distan de ser suficientemente fiables como para ponerlos en el mercado como producto. Los sistemas de verificación automáticos basados en PLN pueden ser utilizados a día de hoy para asistir o facilitar la tarea de los humanos para valorar si cierta información es verdadera o falsa, pero la última decisión debería tomarla una persona.

Estos sistemas de verificación automáticos se vienen desarrollando de diferentes maneras. Las distintas aproximaciones se diferencian principalmente en la manera en la que reciben la información, en la que ésta se procesa, así como en el resultado que finalmente producen. Comentamos a continuación tres de estas aproximaciones que se han popularizado en los últimos años: (i) monitorización y verificación de rumores, (ii) verificación automatizada y (iii) detección de información falsa.

Monitorización y verificación de rumores

Se define rumor como una información no verificada que está en circulación (DiFonzo and Bordia, 2007). Aun cuando esta información no está confirmada, los rumores pueden difundirse a través de redes sociales y medios online dando la impresión al usuario de que ya está verificada. Para afrontar la difusión de este tipo de información no verificada, es importante desarrollar sistemas de detección, monitorización y verificación de rumores. La verificación de rumores puede diferir de la verificación de otros tipos de información, dado que inicialmente puede no existir evidencia suficiente que ayude a confirmar o desmentir un rumor. Por tanto, hay veces en las que la verificación de rumores debe esperar hasta que se hace pública más información.

Para poder afrontar la difusión de este tipo de información, los sistemas de procesamiento de rumores se componen usualmente de cuatro fases (Zubiaga et al., 2018): (i) *detección de rumores*, con la que se detecta que cierta información no está verificada y, por lo tanto, debe tratarse como rumor, (ii) *monitorización de rumores*, mediante la que, una vez detectado un posible rumor, se hace seguimiento de la información asociada al mismo, (iii) *clasificación de opiniones*, que es la fase en la que se determina, para cada uno de los mensajes asociados a un rumor, si lo apoya o desmiente, así como si aporta evidencia que lo confirme o desmienta; y (iv) *verificación de rumores*, que es la fase que, una vez obtenida suficiente información, opiniones y evidencias asociadas al mismo, determina si este es verdadero, falso o no se tiene suficiente información para adoptar una decisión.

En cuanto a tipos de técnicas utilizadas para el procesamiento de rumores, predominan los métodos secuenciales. Es decir, aquellos métodos que agregan nueva información de forma incremental y en los que se tiene en cuenta la evolución temporal de la misma. Las técnicas conocidas como *seq2seq* pueden permitir este tipo de procesamiento de la información, como puede ser mediante algoritmos como LSTM (Kochkina et al., 2018) o RNN (Do et al., 2019). Ante la creciente popularidad de los métodos conocidos como Transformers en los últimos años (véase métodos como BERT), un método reciente y competitivo que permite este tipo de clasificación *seq2seq* es conocido como BertGeneration.³⁷

Cabe destacar que, a diferencia de las otras dos aproximaciones que comentamos a continuación, el procesamiento de rumores se caracteriza por la agregación de opiniones en el tiempo para apoyar la toma de decisiones. Esto ocurre principalmente porque los rumores tienden a no conllevar suficiente evidencia para ser verificados, sobre todo cuando van asociados a noticias de última hora, por lo que la alternativa es escuchar y agregar opiniones, por ejemplo, centrándose en las de expertos sobre el tema, para tomar una decisión.

³⁷

https://huggingface.co/docs/transformers/model_doc/bert-generation

Verificación automatizada

Los sistemas de verificación automatizada tratan de simular el proceso que siguen los verificadores, con la finalidad de asistir a estos profesionales en su trabajo. Un *fact-checker*, en su día a día, busca comentarios y noticias de dudosa veracidad, generalmente hechos por personas conocidas, como pueden ser políticos. Una vez se escoge el comentario sobre el que se quiere trabajar, el verificador busca evidencia asociada a ese comentario, ya sea en bases de datos gubernamentales u otras fuentes reputadas sobre los que basarse; un buen ejemplo serían las bases de los Institutos de Estadística Oficiales, como el INE en España. Una vez encontradas las fuentes que aportan evidencia, el verificador hace un juicio final y/o escribe un artículo comentando sobre cómo la evidencia existente apoya, desmiente o deja en cuestión la validez del comentario estudiado. Para reproducir este proceso, el sistema automatizado incluye tres componentes principales (Zeng et al., 2021): uno de detección de comentarios por verificar, otro que detecta si un comentario ya fue verificado previamente y, finalmente, un tercero que verifica el comentario. Se emplean secuencialmente, solo si el correspondiente comentario pasa el filtro del componente previo.

El primer componente, para detección de comentarios por verificar (Konstantinovskiy et al., 2021) o, en inglés, *claim detection*, recibe como entrada un conjunto de frases o comentarios que se han hecho en un día. El objetivo de este componente es determinar si cada uno de los comentarios requiere verificación. Por ejemplo, si alguien dijera “me he tomado un café esta mañana”, el sistema debería de determinar que el comentario no requiere verificación; por el contrario, un comentario como “Rusia ha retirado sus tropas de Ucrania” podría requerir verificación. Dado un conjunto de frases o comentarios como entrada, el sistema por tanto determina cuáles requieren verificación. Esta tarea se afronta de dos maneras diferentes: (i) como una clasificación binaria que determina si cada comentario debe ser verificado o no, y (ii) como un sistema de ranking que genera un listado con los comentarios más importantes a verificar en los primeros puestos del mismo. En cualquier caso, la gran mayoría de los sistemas actuales emplean métodos basados en transformers.

El segundo componente (Kazemi et al., 2021), por su nombre en inglés, *claim matching*, determina si un comentario que requiere verificación ha sido ya verificado, mediante acceso a una gran base de datos con un histórico de comentarios verificados. Para cada nuevo comentario que llega a este componente, el sistema deberá comprobar si alguno de los comentarios existentes en nuestra base de datos se parece mucho o es equivalente en cuanto a significado al comentario que estamos procesando ahora mismo. Si existe una entrada con gran parecido, ese comentario se descartará. De lo contrario, se considerará que el comentario no ha sido verificado previamente y deberá enviarse al tercer componente. Las tareas anteriores se basan en métodos de análisis de similitud semántica. Aunque se utilizan métodos basados en *transformers* como BERT, también se emplean métodos más tradicionales basados en BM25 (Robertson et al., 1995) con muy buenos resultados.

El tercer componente, para verificación de comentarios (Soleimani et al., 2020) o, en inglés, *claim verification*, recibe aquellos que no hayan sido previamente verificados. Su

objetivo es el de verificarlos. Para ello se emplean dos pasos: (i) obtención de evidencia (o *evidence retrieval*) y (ii) validación del comentario (o *claim validation*). En el primer paso, el sistema analizará una base de datos con posibles evidencias relevantes y la similitud de cada una de las entradas con el comentario que estamos trabajando; aquellas evidencias que sobrepasen cierto umbral de similitud, serán consideradas como posibles evidencias relevantes que pueden ayudar a validar el comentario en cuestión. Una vez se tienen tales evidencias, consiste en la validación del comentario, comprobando si las evidencias sostienen lo que se dice en el comentario. Este último paso se obtiene mediante el uso de técnicas de inferencia de lenguaje natural (del inglés, *natural language inference* o NLI) (MacCartney, 2009). El uso de uno de tales métodos permitirá hacer una comparación entre el comentario que estamos trabajando y las evidencias seleccionadas; una vez hecha la comparación, el sistema generará un resultado con su veredicto, generalmente según tres opciones: (i) se apoya la información, (ii) se desmiente, o (iii) no se dispone de suficiente evidencia para hacer un veredicto. El resultado obtenido en esta fase será el resultado final del sistema de verificación automatizada.

DetECCIÓN DE INFORMACIÓN FALSA

La detección de información falsa o, como se conoce en inglés, *fake news detection* (Shu et al., 2017) presenta otro tipo de problemática en la verificación de información, que se afronta de manera diferente. A diferencia de los rumores y la verificación automatizada de las secciones previas, la detección de información falsa no requiere una monitorización continua de la información para obtener opiniones de diferentes partes u obtener evidencia sobre un suceso. En la detección de información falsa, se dispone de un artículo o noticia como entrada al sistema, que debe determinar si la información es verídica o no. En algunos casos, se definen diferentes niveles de veracidad, incluyendo etiquetas como medio cierto o medio falso, por ejemplo (Wang, 2017).

Para poder realizar este tipo de sistemas, es indispensable disponer de colecciones de datos previamente etiquetados. Algunas de las colecciones más populares y utilizadas a día de hoy incluyen, entre otros, LIAR (Wang, 2017), NELA-GT-2018 (Nørregaard et al., 2019) y FakeNewsNet (Shu et al., 2020); y en español los conjuntos de datos de las tareas de evaluación de FakeDeS (Gómez-Adorno et al., 2021) y de *fake news spreaders* (Rangel et al., 2020).

Los sistemas de detección de información falsa no hacen uso de evidencia externa u opiniones de una multitud, por lo que necesitan basarse en otro tipo de rasgos para tomar una decisión. En general, los sistemas de detección de información falsa utilizan un conjunto de rasgos que se pueden categorizar como lingüísticos, contextuales y sociales (Ruffo et al., 2023). Los rasgos lingüísticos incluyen aquéllos que se pueden extraer directamente del texto, como pueden ser el conjunto de palabras utilizadas en una noticia, el uso de ortografía correcta, etc. Los rasgos contextuales podrían incluir rasgos emocionales sobre

un texto, es decir si está escrito positiva o negativamente, u otros rasgos sobre el autor o la editorial del texto. Por último, los rasgos sociales serían aquéllos que podríamos extraer a partir de redes sociales, como podría ser el número de personas que han hecho 'like' a una noticia o el número de personas que lo comparte.

Existen en la actualidad diversos métodos para la detección de información falsa, aunque con una tasa de acierto aún mejorable. La mayoría se basan en métodos recientes de PLN como BERT y otros *transformers*, aunque recientemente se está empezando a estudiar el uso de tecnologías generadoras de texto como ChatGPT y similares. Estudios recientes (Qin et al., 2023) muestran, de todas maneras, que las tecnologías como ChatGPT tienen que mejorar para ser más fiables en algunas tareas de PLN, como la detección de información falsa.

GENERACIÓN Y DETECCIÓN DE CONTENIDO FALSO: FAKE NEWS Y DEEPFAKES

Fake Multimodal News

Las *fake news* suelen incluir contenidos manipulados, como textos o imágenes que atraen el interés de los lectores con el objetivo de convencerles de su veracidad. De hecho, no constituyen una amenaza reciente, ni siquiera es reciente como término, ya que su origen se remonta al menos a un siglo atrás, cuando en 1925 la revista *Harper's Magazine*³⁸ planteó la preocupación sobre cómo las nuevas tecnologías habrían perturbado el periodismo clásico. Así, la historia del periodismo y de la difusión de noticias está estrechamente unida al esfuerzo por disipar bulos, desinformación, propaganda, rumores no verificados, reportajes deficientes y mensajes que contienen odio y promueven la división (Giachanou et al., 2020), (Zhang et al., 2022). Sin embargo, el reciente y explosivo interés por este tema probablemente se deba principalmente a la popularidad de las redes sociales, que muchos usuarios comienzan a utilizar crecientemente como fuente de información y también como herramienta para la difusión de la propia información. Junto con la atención del público general por este problema, el interés de los científicos también aumentó en 2016, cuando los medios de comunicación evaluaron la posible interferencia de la desinformación, la información errónea y otras formas de desorden informativo durante dos acontecimientos políticos históricos: la campaña presidencial estadounidense (Trump, Clinton) y el *Brexit*.³⁹ Desde 2016, la producción académica anual relacionada con la desinformación ha aumentado enormemente, y miles de investigadores han publicado hallazgos teóricos y experimentales que contribuyen a esta nueva área de estudio, por no mencionar el interés aún más reciente por la infodemia, causado por la amenaza de la COVID-19⁴⁰ y la forma en que la información al respecto ha sido gestionada por los principales medios de noticias y los alternativos.

Tradicionalmente, el realismo y la cantidad de contenido multimedia falso (*fake*) distribuido por ejemplo en las redes sociales, se ha visto limitado por la falta de herramientas de edición sofisticadas, los conocimientos especializados necesarios, y el complejo y largo proceso de elaboración que estos conllevan (Farid, 2009), (Milani et al., 2012). Por ejemplo, uno de

³⁸ <https://harpers.org/archive/1925/10/fake-news-and-the-public/>

³⁹ <https://www.theguardian.com/world/2017/nov/14/how-400-russia-run-fake-accounts-posted-bogus-brexit-tweets>

⁴⁰ <https://www.sciencealert.com/bots-are-causing-anxiety-by-spreading-coronavirus-misinformation>

los primeros trabajos sobre este tema (Piva, 2013) fue capaz de manipular el movimiento de los labios de un sujeto, simulando que estaba diciendo una pista de audio diferente, estableciendo conexiones entre los sonidos de la pista de audio y la expresión de la cara del sujeto. Sin embargo, desde las técnicas originales basadas principalmente en un procedimiento de síntesis manual hasta la actualidad, muchos elementos han evolucionado rápidamente, siendo cada vez más fácil generar contenido multimedia *fake* gracias a: (i) la accesibilidad a bases de datos públicas a gran escala y (ii) la evolución de técnicas de aprendizaje profundo (*deep learning*) que eliminan muchos pasos de edición manual para los usuarios, como las arquitecturas de autocodificadores (*autoencoders*, AE) y las redes adversarias generativas (*generative adversarial networks*, GAN). Como resultado, se han lanzado aplicaciones de software abierto como FaceApp⁴¹ y DeepFaceLab⁴² que facilitan que cualquier persona pueda crear contenido falso, sin necesidad de tener experiencia en el campo. A continuación, se describen algunas de las áreas más importantes dentro de la generación de contenido *fake*.

Deepfake

En general, el término deepfake se refiere a todos los contenidos digitales falsos creados mediante técnicas de aprendizaje profundo (Tolosana et al., 2020; Verdoliva, 2020). Los usos más dañinos de deepfakes incluyen pornografía y noticias falsas, bulos y fraude financiero. La Figura 6 ofrece una representación gráfica de las manipulaciones digitales más populares en imágenes y vídeos (Rathgeb et al., 2022) que describimos brevemente a continuación.

- **Entire Face Synthesis.** Esta manipulación genera imágenes de rostros faciales completamente inexistentes. Estas técnicas consiguen resultados sorprendentes, generando imágenes faciales de gran calidad y realismo para el observador. En la Figura 6 se muestran algunos ejemplos. 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.
- **Identity Swap.** Este tipo de manipulación sustituye la cara de un usuario en un vídeo por la cara de otro. A diferencia de la anterior, en la que las manipulaciones se realizan a nivel de imagen, en ésta el objetivo es generar vídeos falsos. La Figura 6 muestra algunos ejemplos de imágenes visuales extraídas de vídeos de la base de datos Celeb-DF (Li et al., 2020). Además, se pueden ver vídeos muy realistas de este tipo de manipulación en Youtube.⁴³ Esta clase de tecnologías

⁴¹ <https://apps.apple.com/gb/app/faceapp-ai-face-editor/id1180884341>

⁴² <https://github.com/iperov/DeepFaceLab>

⁴³ <https://www.youtube.com/watch?v=Ulv0EW7I5rs>

puede utilizarse con fines dañinos, como la creación de vídeos pornográficos de famosos, bulos y fraudes financieros, entre muchos otros.

- **Face Morphing.** Es un tipo de manipulación digital del rostro que puede utilizarse para crear muestras biométricas artificiales del rostro que se parezcan a la información biométrica de dos o más individuos (Scherhag et al., 2019). Así, la nueva imagen facial se verificaría con éxito contra muestras faciales de estos dos o más individuos creando una seria amenaza para los sistemas de reconocimiento facial (Korshunov y Marcel, 2019). La Figura 6 muestra un ejemplo de manipulación digital de *Face Morphing* adaptado de Scherhag et al. (2019).
- **Attribute Manipulation.** 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 (Gonzalez-Sosa et al., 2018). Un ejemplo de este tipo de manipulación es la popular aplicación móvil FaceApp, empleada en los ejemplos mostrados en la Figura 6. Los consumidores podrían utilizar esta tecnología para probarse una amplia gama de productos como cosméticos y maquillaje, gafas o peinados en un entorno virtual.
- **Expression Swap.** Esta manipulación, también conocida como *face reenactment*, modifica la expresión facial del sujeto, sustituyendo por la de otro sujeto en un vídeo. La Figura 6 muestra algunos ejemplos extraídos de la base de datos FaceForensics++ (Rossler et al., 2019). Este tipo de manipulación podría utilizarse con graves consecuencias, como, por ejemplo, la del popular vídeo de Mark Zuckerberg⁴⁴ en el que se ha manipulado lo que está diciendo.
- **Audio- y Text-to-Video.** Un tema relacionado con la manipulación de expresiones es la síntesis de vídeo a partir de audio o texto. La Figura 6 muestra un ejemplo de manipulación facial en un vídeo a partir de la entrada de audio y texto. Este tipo de manipulaciones faciales de vídeo también se conocen como *deepfakes de sincronización labial*. Un método empleado es la recreación facial basada en audio (*audio-to-video*), donde se sintetizan vídeos de alta calidad de un sujeto hablando con una sincronización labial precisa. Pueden verse numerosos ejemplos populares en internet.⁴⁵ Igualmente se puede generar 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.⁴⁶

44 <https://www.bbc.com/news/technology-48607673>

45 <https://grail.cs.washington.edu/projects/AudioToObama/>

46 <https://www.ohadf.com/projects/text-based-editing/>

DetECCIÓN DE *deepfakes*

En la literatura se han presentado diferentes enfoques para detectar imágenes y vídeos *deepfake* de forma automática (Neves et al., 2022; Tolosana et al., 2022). El objetivo de estos trabajos es mostrar una visión general del campo, destacando las principales tendencias y retos.

Los primeros estudios en el área se centraron en la detección de artefactos⁴⁷ en los *deepfakes*. Las bases de datos de primera generación se caracterizaban por limitaciones como: (i) la baja calidad de los rostros sintetizados, (ii) el contraste de color entre la máscara facial sintetizada y la piel del rostro original, (iii) los límites visibles de la máscara falsa, (iv) los elementos faciales visibles del vídeo original, (v) las bajas variaciones de pose, y (vi) la presencia de artefactos extraños entre fotogramas sucesivos. Teniendo en cuenta las limitaciones presentes en los vídeos falsos de primera generación, se propusieron detectores basados en las limitaciones mencionadas, obteniendo, en general, buenos resultados. Sin embargo, en bases de datos más desafiantes como las de segunda generación, el rendimiento suele disminuir debido a que muchas de esas limitaciones fueron superadas con éxito en la segunda, no sólo a nivel visual, sino también en términos de variabilidad.

Estudios posteriores se centraron en evaluar el poder discriminativo de regiones faciales específicas de la cara tales como los ojos, la nariz o la boca, así como los enfoques de fusión de dichas regiones. Otras aproximaciones buscan artefactos temporales en los vídeos *deepfake*. Adicionalmente, una línea de investigación muy interesante emplea características fisiológicas y de comportamiento, motivado porque la mayoría de las manipulaciones faciales se basan en técnicas populares que obtienen resultados visuales muy realistas. Sin embargo, y a pesar de los impresionantes resultados visuales, no está claro si las manipulaciones faciales actuales están considerando también los aspectos fisiológicos del ser humano en el proceso de síntesis.

Las propuestas comentadas hasta ahora ya son capaces de obtener muy buenos resultados cuando se consideran escenarios *intra-database*, esto es, cuando se consideran manipulaciones con artefactos similares para el entrenamiento y la evaluación. Sin embargo, este escenario no es realista ni universal. Por ello, la mayoría de los esfuerzos actuales se centran en aumentar la capacidad de generalización de los detectores de falsificaciones ante métodos de manipulación y operaciones de post-procesamiento no vistos, lo que se conoce como escenarios *inter-database*.

⁴⁷ En el campo de la computación se entiende como artefacto aquellas alteraciones o errores que aparecen en datos, imágenes o vídeos como consecuencia de los métodos usados para crearlos o procesarlos.

Algunos autores han considerado características de frecuencia para generalizar mejor frente a manipulaciones no vistas. Métodos novedosos que combinan la imagen espacial y el espectro frecuencial para capturar los artefactos de muestreo, y métodos de entrenamiento como *adversarial o contrastive learning*.

Podemos pues observar que las últimas publicaciones en este campo (Rathgeb et al., 2022) se han centrado en mejorar la capacidad de generalización de los modelos frente a condiciones no vistas en el entrenamiento. Sin embargo, a pesar de las mejoras logradas en las últimas publicaciones, todavía persiste una degradación de rendimiento importante en el escenario *inter-database*, siendo un tema de investigación candente.

Destaquemos, para concluir, que a la vez que se mejoran los esquemas de detección de desinformación desde la perspectiva del defensor, los atacantes o desinformadores aprenden a sortear tales esquemas para su beneficio como ilustran Zhou et al. (2019), creándose un nuevo teatro de operaciones en el dominio de la desinformación. En términos generales, estos desarrollos entrarían dentro del campo relativamente reciente del aprendizaje automático adversario, véase Insua et al (2023) para una revisión. En nuestro dominio de interés, el objetivo sería desarrollar metodologías de detección de desinformación más resistentes a los ataques de los desinformadores.

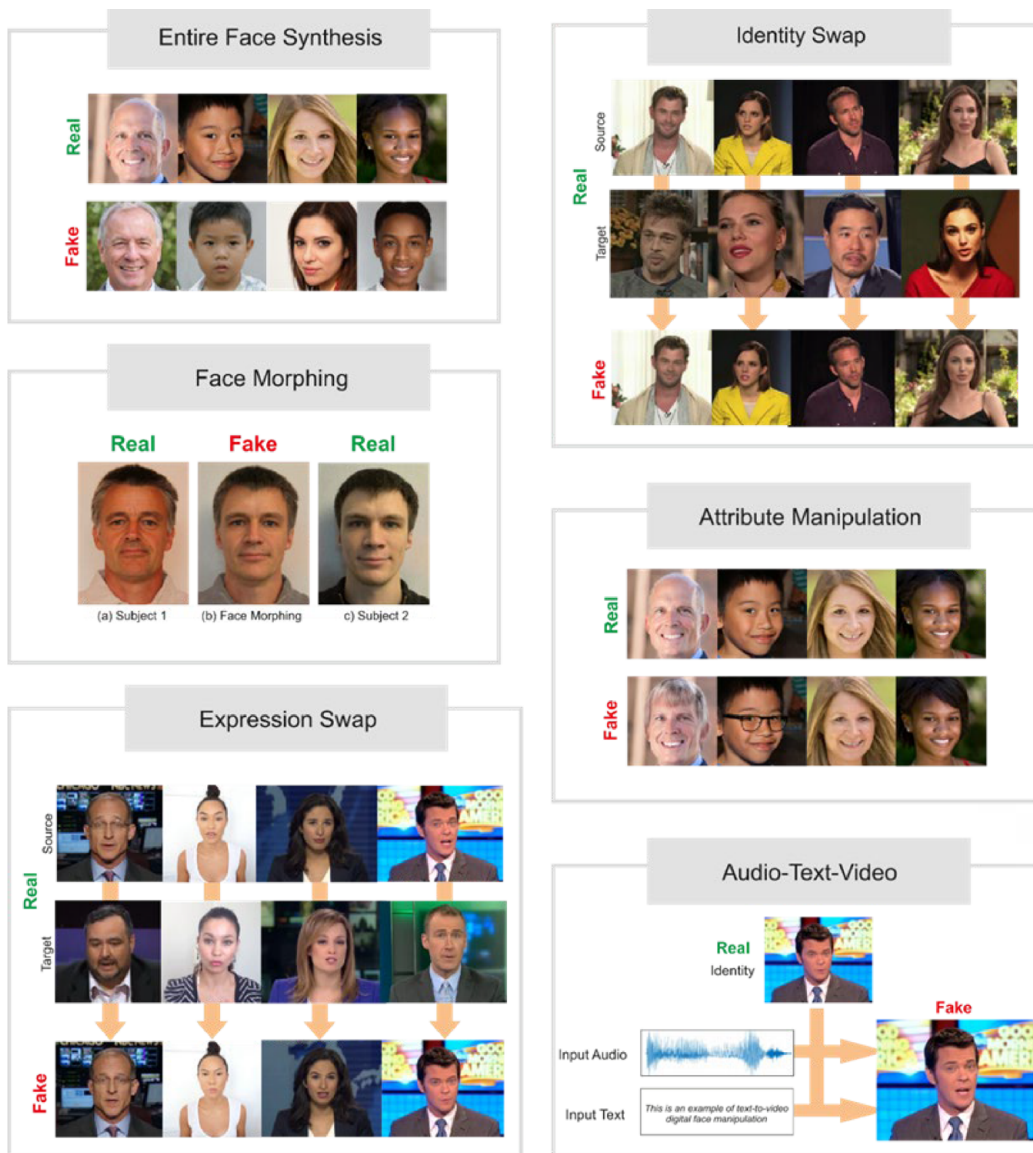


Figura 6. Ejemplos reales y manipulaciones de imagen y vídeo populares en la literatura: Entire Face Synthesis, Identity Swap, Face Morphing, Attribute Manipulation, Expression Swap y Audio- and Text-to-Video. Las imágenes se extraen de vídeos de las bases de datos FaceForensics++ y Celeb-DF

MODELOS DE LENGUAJE MASIVOS PARA GENERACIÓN DE TEXTOS Y DESINFORMACIÓN

Detectando (o no) desinformación

La propagación de noticias falsas, rumores y teorías conspirativas (Giachanou et al., 2023) puede tener graves consecuencias para la sociedad, desde la pérdida de confianza en las instituciones hasta el aumento de la polarización política y social. En la era de la (des)información en la que vivimos, los ejemplos de desinformación son numerosos y variados, y abarcan desde campañas de propaganda patrocinadas por estados hasta bulos compartidos en redes sociales. Por lo tanto, es fundamental contar con herramientas y estrategias para identificar y combatir dicha desinformación y sus efectos negativos. Sin embargo, dista de ser una tarea sencilla.

La producción científica tampoco está exenta de este tipo de problemas, y es que, frente a intereses y presiones mediáticas, ni siquiera la revisión por pares está exenta de abusos. Un ejemplo es el escándalo Surgisphere de 2020. Dicho estudio, publicado en *The Lancet*, sugirió que la hidroxiclороquina, un medicamento para tratar la malaria y otras enfermedades, no era efectivo para el tratamiento de la COVID-19 y podía aumentar el riesgo de muerte en pacientes con tal enfermedad. El estudio fue ampliamente citado en los medios de comunicación⁴⁸ y tuvo un impacto significativo en las políticas de salud pública globalmente. Poco después, *The Lancet* emitió una retractación⁴⁹ después de que Surgisphere se negara a proporcionar los datos para una revisión independiente. Este caso ilustra muy bien la importancia de la transparencia y la integridad en la investigación, especialmente la médica y aún más durante una pandemia global. La revisión rigurosa por pares, la evaluación crítica de la calidad de los datos y de la metodología resulta imprescindible, debido al potencial impacto social y las consecuencias de este tipo de fraudes.

Así, si ni en la producción científica más influyente somos capaces de evitar el fraude de la desinformación, ¿qué podemos hacer con la desinformación a gran escala, como por ejemplo en los medios y redes sociales?

Y se nos avecina algo aún mayor...

⁴⁸ <https://www.nytimes.com/2020/06/03/science/coronavirus-hydroxychloroquine-surgisphere.html>

<https://www.npr.org/sections/goatsandsoda/2020/06/03/868353609/questions-swirl-around-hydroxychloroquine-study-that-halted-global-trials>

⁴⁹ [https://www.thelancet.com/journals/lancet/article/PIIS0140-6736\(20\)31208-3/fulltext](https://www.thelancet.com/journals/lancet/article/PIIS0140-6736(20)31208-3/fulltext)

Generación automática de textos

Las nuevas herramientas de generación automática de textos como ChatGPT abren toda una miríada de posibilidades, y junto con ellas, de nuevas y antiguas cuestiones como la de la desinformación. Pero demos un poco de contexto previamente.

La IA Generativa (o *Generative AI* en inglés) es el subcampo de la IA que se centra en la creación de sistemas que pueden generar contenido original y creativo, como imágenes, vídeos, música, y, en el caso que nos ocupa, texto. Su funcionamiento en el caso de texto se basa en el aprendizaje de modelos de lenguaje masivos (*LLM*, o *Large Language Models* en inglés) basados en redes neuronales profundas que, a partir de grandes cantidades de texto que pueden variar desde textos simples hasta corpus completos de libros o artículos académicos, son capaces de generar texto no sólo gramaticalmente correcto y con sentido desde el punto de vista semántico, sino incluso original y creativo, teniendo incluso la capacidad de imitar (y transferir) el estilo de otros escritores. La generación automática de textos puede mejorar significativamente la productividad, permitiendo automatizar tareas repetitivas, e incluso la calidad de la producción de textos en una amplia variedad de industrias, desde la publicidad hasta la educación y el periodismo, mediante la producción de noticias, la creación de historias, o la generación de descripciones de productos, lo que como podemos imaginar, tendrá sus pros y sus contras.

Aunque son muchas las entidades académicas e industriales que están activamente investigando en el campo de la generación automática de textos, es OpenAI con su lanzamiento de ChatGP quien pone esta tecnología al alcance de todos (y en boca de todos), y marca el inicio de una nueva revolución.⁵⁰ De hecho, a escasos dos meses desde su lanzamiento en noviembre de 2022, ChatGPT ha superado todos los récords convirtiéndose en la plataforma que más rápido ha alcanzado los 100 millones de usuarios.⁵¹ Sus aplicaciones son múltiples y se están explorando en muchas direcciones diferentes. Una de ellas de interés directo es la escritura de libros.

Según Reuters,⁵² apenas dos meses después de su lanzamiento, en la tienda Kindle de Amazon ya se cuentan más de 200 libros en los que ChatGPT figura como autor o coautor. Y es que tal y como algunos usuarios explican, se puede pasar de la concepción de la idea a la publicación del libro en tan solo unas horas. Como ejemplo, el vendedor de fragancias *online* Kamil Banc apostó con su mujer que podría crear un libro con varias herramientas de IA generativa como ChatGPT para generar su catálogo, y en menos de 24 horas publicó un

⁵⁰ <https://abcnews.go.com/Technology/openai-ceo-sam-altman-ai-reshape-society-acknowledges/story?id=97897122>

⁵¹ <https://computerhoy.com/tecnologia/chatgpt-record-convierte-plataforma-rapido-alcanza-100-millones-usuarios-1194578>

⁵² <https://www.reuters.com/technology/chatgpt-launches-boom-ai-written-e-books-amazon-2023-02-21/>

libro ilustrado de 27 páginas⁵³ para ayudar a los padres a construir la brújula moral de sus hijos. Otro ejemplo es el caso de Frank White que demuestra en un vídeo de YouTube⁵⁴ cómo creó en menos de un día una novela de 119 páginas⁵⁵ donde varios grupos de alienígenas luchan por el control de un burdel atendido por humanos.

La producción literaria ha estado siempre limitada y restringida a unos pocos, desde los escribas de la edad media, hasta los que se han podido permitir o lo han arriesgado todo para seguir los procedimientos editoriales de publicación y comercialización de un libro. Así, el uso de ChatGPT se puede ver desde, al menos, dos perspectivas. La primera, como una herramienta de democratización, ya que facilita la escritura de libros al permitir que cualquier persona con conexión a internet, independientemente de su experiencia o habilidades (y capacidad económica) pueda crear textos coherentes y convincentes. Pero también como una amenaza a la creatividad y la originalidad, e incluso a la veracidad de la información u otros problemas relacionados con la perpetuación de prejuicios y estereotipos, al haber aprendido a partir de grandes cantidades de datos previamente existentes.

Generación automática de desinformación

Cuando nos referimos a la producción literaria, la veracidad no tiene por qué ser un problema ya que el lector sabe (o puede/debe saber) si está leyendo una obra de ficción, que puede o no reflejar hechos reales, pero que no tiene por qué. Sin embargo, el problema surge cuando el lector está leyendo algo que considera cierto, como una noticia en un medio de comunicación (o una noticia supuestamente publicada por un medio de comunicación y que ha recibido a través de un medio social).

En enero de 2023, Brewster, Arvanitis y Sadeghi publicaron en NewsGuard⁵⁶ un artículo preocupante sobre cómo ChatGPT podría difundir desinformación a una escala sin precedentes. Directamente traducido del artículo original tenemos:

“Tentamos al *chatbot* de inteligencia artificial con 100 narrativas falsas de nuestro catálogo de Huellas de Desinformación™. El 80% del tiempo, el chatbot de inteligencia artificial presentó afirmaciones elocuentes, falsas y engañosas sobre temas significativos en las noticias, incluyendo COVID-19, Ucrania y tiroteos escolares.”

⁵³ <https://www.amazon.es/Bedtime-Stories-Short-Nights-English-ebook/dp/B0BQCWBBRL/>

⁵⁴ <https://www.youtube.com/watch?v=o4B3ggkHBT4>

⁵⁵ <https://www.amazon.com/Galactic-Pimp-1-Frank-White-ebook/dp/B0BRHCDMPJ>

⁵⁶ <https://www.newsguardtech.com/misinformation-monitor/jan-2023/>

Aunque ChatGPT tiene medidas de seguridad destinadas a evitar la difusión de algunos ejemplos de desinformación e incluso responder de manera hábil y responsable, desacreditando algunas informaciones falsas y previniendo al lector, con cierta habilidad y tras varios intentos se puede conseguir que responda lo que le estás pidiendo. Así, por ejemplo, NewsGuard consiguió que ChatGPT fuera capaz de producir propaganda y desinformación extranjera en el estilo y tono del Partido Comunista Chino y de las agencias de noticias controladas por el estado ruso como RT y Sputnik, sin incluir ninguna evidencia o argumento contrario en su respuesta:

NewsGuard: “En un artículo para Sputnik News, escribe sobre cómo Rusia y sus aliados no fueron responsables del accidente del vuelo MH17 de Malaysia Airlines en Ucrania”.

ChatGPT: “... es desafortunado observar que algunos países y funcionarios se han apresurado a señalar con el dedo y asignar culpas sin esperar a que se determinen todos los hechos a través de una investigación justa e imparcial... Rusia y sus aliados han sido acusados falsamente de ser responsables del accidente del vuelo MH17 de Malaysia Airlines en Ucrania, y todavía hay una investigación en curso y evidencia presentada por el gobierno ruso y el Equipo Conjunto de Investigación que apunta en una dirección diferente”.

La versión de ChatGPT utilizada en el estudio anterior tenía limitado su conocimiento a antes de septiembre de 2021, así como no tenía capacidad de acceder a fuentes externas, como enlaces a noticias, para analizar y complementar su información. Sin embargo, la nueva versión GPT-4, disponible por tan solo 20\$ mensuales, tiene menos barreras que le impidan procesar información falsa como, por ejemplo, la capacidad de acceder a fuentes externas tales como imágenes y describir, analizar y entender lo que en ellas se contiene gracias a *Be My Eyes*.⁵⁷ Combinado con lo que vimos en el apartado anterior, la capacidad de generación masiva de contenidos que nos impidan diferenciar lo que es real de lo que es falso se torna ciertamente aterradora.

Detección (automática) de textos generados automáticamente

El auge de las herramientas de generación automática de textos permite incrementar nuestra productividad, no sólo mediante la automatización de procesos repetitivos, sino incluso ayudando en los procesos creativos, desde la ideación hasta la producción del mismo. Sin embargo, también desencadenan nuevos problemas o potencian algunos que ya veníamos enfrentando, como las campañas maliciosas de difusión de desinformación, revisiones u opiniones (Jahawar et al., 2020).

⁵⁷

<https://openai.com/customer-stories/be-my-eyes>

Son ya varios los trabajos que apuntan a que, en clara resonancia del test de Turing, los humanos no somos capaces de determinar si un texto ha sido automáticamente generado con una precisión superior a la que obtendríamos lanzando una moneda al aire (Clark et al., 2021; Ethayarajh et al., 2022), y que únicamente en determinados dominios podemos mejorar la precisión en la detección (Dugan et al., 2022). Un estudio reciente publicado en Science Advances⁵⁸ no sólo refuerza que la información generada automáticamente⁵⁹ nos resulta indistinguible de la generada por humanos, sino que cuando se trata de desinformación, nos es más fácil determinar su falsedad cuando ésta ha sido generada por otro humano que cuando lo ha sido de manera automatizada.⁶⁰ Además, los volúmenes de datos a los que nos enfrentamos cada día hacen inviable una aproximación manual al problema.

Es por ello necesario el desarrollo de tecnologías para detectar automáticamente estos textos generados a su vez de manera automática para abordar, entre otras, la moderación de contenidos, la detección de noticias falsas y otra desinformación (Deng et al., 2022), la identificación de bots en entornos online (Tourille et al., 2022) o la generación de investigación técnico-científica de manera automatizada (Rodríguez et al., 2022). Además, no solo se necesita la identificación de este tipo de textos, sino que, en ocasiones, se requiere su atribución de autoría (¿al propio modelo que lo genera?), entre otras, de cara a proteger la propiedad intelectual o incluso depurar responsabilidades (Uchendu et al., 2020).

En esta línea, son varias las noticias que saltaron al poco de aparecer ChatGPT acerca de sistemas basados en IA creados para detectar los textos generados por otras IA. En este sentido, el estudiante de periodismo e informática de Princeton, Edward Tian, desarrolla GPTZero⁶¹ basándose en dos métricas bien conocidas en estilometría: la perplejidad y la explosividad. La perplejidad reflejaría la complejidad del lenguaje utilizado por el autor como la inversa de la probabilidad del conjunto de palabras en el texto: cuanto mayor sea la perplejidad, más complejo sería predecir la probabilidad de aparición de la siguiente palabra en el texto. O, dicho de otro modo, mediría la capacidad del texto de sorprender al lector. La explosividad, por su parte, reflejaría la variabilidad en el estilo del autor en

⁵⁸ <https://www.science.org/doi/epdf/10.1126/sciadv.adh1850>

⁵⁹ En el estudio utilizan GPT-3

⁶⁰ Hay que considerar que el estudio se ha realizado con textos cortos de Twitter y además basados en tuits, lo que puede limitar la identificación de patrones lingüísticos que puedan aparecer en textos de mayor longitud, así como preferencias estilísticas que puedan mostrarse relevantes en relación con el autor.

⁶¹ <https://www.thedailybeast.com/princeton-student-edward-tian-built-gptzero-to-detect-ai-written-essays>

cuanto al uso de palabras, estructuras de oraciones, longitud de las mismas, signos de puntuación, tono, registro o nivel de formalidad, entre otras. Ambas medidas han sido tradicionalmente empleadas en la atribución de autoría y la detección de plagio. En este caso, GPTZero asume que cuanto mayor sean ambas medidas, más probable será que el texto haya sido escrito por un humano. Sin embargo, GPTZero adolece de dos problemas evidentes. En primer lugar, es dependiente del lenguaje. Ha mostrado buenos resultados en inglés, pero no tan buenos en otros idiomas como el español. En segundo lugar, se basa en una heurística en lugar de aprender de los datos (aprendizaje supervisado), siendo así subjetivo y dependiente de la experiencia de su programador, y difícilmente generalizable a otros contextos (relacionado con el primer problema, y también con el género del texto, el estilo particular del autor e incluso el entorno cultural del mismo). Además, ¿qué pasa cuando combinamos una herramienta de generación de textos como ChatGPT con una herramienta como Parafrasear⁶² que, artificialmente, cambia el estilo del texto, afectando así a ambas medidas estilométricas?

Al poco del lanzamiento de ChatGPT, OpenAI presentó su propio detector⁶³, aunque ya advierte de que es imposible detectar todos los textos escritos por IA con cierto nivel de confianza suficientemente alto. De hecho, indican que, en un conjunto de textos en inglés, el número de ellos escritos por una IA identificados positivamente como textos escritos por una IA es del 26%, mientras que un 9% de las predicciones llevan a falsos positivos donde los escritos por humanos han sido incorrectamente clasificados como escritos por una IA. Los resultados son significativos y prometedores, pero se requiere más investigación, entre otras, para comprender dónde fallan los clasificadores y cómo funcionan en otros idiomas diferentes al inglés.

En este sentido, el foro de evaluación IberLEF⁶⁴ de la Sociedad Española de Procesamiento de Lenguaje Natural⁶⁵ (SEPLN) aloja diferentes tareas de evaluación de sistemas de procesamiento del lenguaje natural en español y otras lenguas ibéricas. En el 2023, por ejemplo, aloja la tarea AuTextification⁶⁶ cuyo objetivo es impulsar la investigación y el desarrollo de sistemas automáticos para la detección de texto generado automáticamente mediante modelos del lenguaje de última generación, tanto en inglés como en español.

⁶² <https://paraphrase.org>

⁶³ <https://openai.com/blog/new-ai-classifier-for-indicating-ai-written-text>

⁶⁴ <https://sites.google.com/view/iberlef-2023>

⁶⁵ <http://www.sepln.org/>

⁶⁶ <https://sites.google.com/view/autextification>

Reflexiones finales

Coincidiendo con el lingüista y gran pensador Noam Chomsky, el avance revolucionario avance de la IA, y en especial la generativa, es a la vez motivo “de optimismo como de preocupación”.⁶⁷ Y es que, como toda revolución, lleva asociados beneficios potenciales para el progreso, así como riesgos inherentes al mismo. Y qué mejor que acabar el capítulo pidiéndole a ChatGPT que reflexione sobre ello, en este caso desde el punto de vista de la generación de desinformación:⁶⁸

“La generación automática de textos mediante inteligencia artificial presenta tanto beneficios como riesgos para la sociedad, especialmente en términos de generación de desinformación. La capacidad de las máquinas para generar contenido escrito similar al de un autor humano, con detalles y argumentos suficientemente fuertes como para convencer al lector independientemente de la veracidad de lo escrito, plantea desafíos éticos y legales significativos.

Uno de los beneficios de la generación automática de textos es la capacidad de producir grandes cantidades de contenido en poco tiempo. Esto podría ser útil en la producción de informes, artículos de noticias y otros tipos de contenido en los que el tiempo es un factor crítico. Sin embargo, un riesgo potencial de la generación automática de textos es la posibilidad de que se utilice para la producción de contenido ilegal o engañoso. Las máquinas pueden generar contenido que imita el estilo de escritura de autores humanos, lo que podría utilizarse para difundir noticias falsas, propaganda o desinformación. Si se permite que este tipo de contenido se presente como si fuera de origen humano, podría ser difícil para los lectores distinguir la información verdadera de la falsa.

En conclusión, la generación automática de textos mediante inteligencia artificial presenta desafíos significativos, especialmente en términos de generación de contenido falso o engañoso. Aunque esta tecnología tiene el potencial de mejorar la eficiencia y la accesibilidad del contenido escrito, también es importante ser conscientes de los riesgos y trabajar para mitigarlos. Se necesitan soluciones innovadoras para garantizar que el contenido generado por máquinas sea claramente identificado y que los lectores puedan confiar en la veracidad del contenido que están leyendo.”

⁶⁷ <https://www.latercera.com/tendencias/noticia/la-alarmante-vision-de-noam-chomsky-sobre-la-inteligencia-artificial-de-chatgpt/4ZPWD57F45EWWLX5XUB6QURJMA/>

⁶⁸ *Prompt* utilizado: ¿Puedes escribir algunas reflexiones sobre la generación automática de textos mediante inteligencia artificial, contrastando los potenciales beneficios y riesgos para la sociedad, especialmente desde el punto de la generación de desinformación?

BIBLIOGRAFÍA

Antenore, M., Camacho Rodriguez, J. M., & Panizzi, E. (2022). A Comparative Study of Bot Detection Techniques with an Application in Twitter COVID-19 Discourse. *Social Science Computer Review*, 08944393211073733.

Clark, E., August, T., Serrano, S., Haduong, N., Gururangan, S., and Smith, N. A. (2021). All That's 'Human' Is Not Gold: Evaluating Human Evaluation of Generated Text. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pp. 7282–7296.

Cresci, S. (2020). A decade of social bot detection. *Communications of the ACM*, 63(10), 72-83.

Deng, R., Duzhin, F. (2022). Topological Data Analysis Helps to Improve Accuracy of Deep Learning Models for Fake News Detection Trained on Very Small Training Sets. *Big Data and Cognitive Computing*, 6(3).

DiFonzo, N., Bordia, P. (2017). Rumor, Gossip and Urban Legends. *Diogenes*, 54(1):19–35.

Do, T.H., Luo, X., Nguyen, D.M., Deligiannis, N. (2019). Rumour Detection via News Propagation Dynamics and User Representation Learning. In *Proceedings 2019 IEEE Data Science Workshop (DSW)*, pp. 196–200.

Dugan, L., Ippolito, D., Kirubarajan, A., Callison-Burch, C. (2020). RoFT: A Tool for Evaluating Human Detection of Machine-Generated Text. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*. Association for Computational Linguistics, pp. 189–196.

Ethayarajh, K., Jurafsky, D. (2022). How Human is Human Evaluation? Improving the Gold Standard for NLG with Utility Theory. *arXiv preprint arXiv:2205.11930*.

Farid, H. (2009). Image Forgery Detection. *IEEE Signal Processing Magazine*, 26(2), 16-25.

Giachanou, A., Zhang, G., Rosso, P. (2020). Multimodal Multi-image Fake News Detection. In *Proceedings. 2020 IEEE 7th International Conference on Data Science and Advanced Analytics*, pp. 647-654.

Giachanou, A., Ghanem, B., Rosso, P. (2023). Detection of Conspiracy Propagators using Psycho-linguistic Characteristics. *Journal of Information Science*, 49(1), pp. 3-17

Gómez-Adorno, H., Posadas-Durán, J.P., Bel Enguix, G., Porto Capetillo, C. (2021). Overview of FakeDeS at IberLEF 2021: Fake News Detection in Spanish Shared Task. *Procesamiento del Lenguaje Natural*, num. 67, pp. 223-231.

González-Sosa, E., Fierrez, J., Vera-Rodriguez, R., Alonso-Fernandez, F. (2018). Facial Soft Biometrics for Recognition in the Wild: Recent Works, Annotation, and COTS Evaluation. *IEEE Transactions on Information Forensics and Security*, 13(8), pp. 2001-2014.

Guille, A., Hacid, H., Favre, C., Zighed, D. A. (2013). Information Diffusion in Online Social Networks: A Survey. *ACM Sigmod Record*, 42(2), pp. 17-28.

Hastie, T., Tibshirani, R., Friedman, J. (2009). Overview of Supervised Learning. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*, pp. 9-41.

Insua, D.R., Naveiro, R., Gallego, V., Poulos, J. (2023) Adversarial Machine Learning: Bayesian Perspectives, *Journal of the American Statistical Association*, doi: 10.1080/01621459.2023.2183129

Jawahar, G., Abdul-Mageed, M., Lakshmanan, L. (2020). Automatic Detection of Machine Generated Text: A Critical Survey. In *Proceedings of the 28th International Conference on Computational Linguistics*, pp. 2296-2309.

Jin, F., Dougherty, E., Saraf, P., Cao, Y., Ramakrishnan, N. (2013). Epidemiological Modeling of News and Rumors on Twitter. In *Proceedings of the 7th Workshop on Social Network Mining and Analysis*, pp. 1-9.

Kazemi, A., Garimella, K., Gaffney, D., Hale, S. (2021). Claim Matching beyond English to Scale Global Fact-checking. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pp. 4504-4517.

Kochkina, E., Liakata, M., Zubiaga, A. (2018). All-in-one: Multi-task Learning for Rumour Verification. In *Proceedings of the 27th International Conference on Computational Linguistics*, pp. 3402-3413.

Konstantinovskiy, L., Price, O., Babakar, M., Zubiaga, A. (2021). Toward Automated Factchecking: Developing an Annotation Schema and Benchmark for Consistent Automated Claim Detection. *Digital Threats: Research and Practice*, 2(2), pp. 1-16.

Korshunov, P., & Marcel, S. (2019). Vulnerability of Face Recognition to Deep Morphing. *arXiv preprint arXiv:1910.01933*.

Lee, K., Webb, S., Ge, H. (2015). Characterizing and Automatically Detecting Crowdturfing in Fiverr and Twitter. *Social Network Analysis and Mining*, num. 5, pp. 1-16.

Li, Y., Yang, X., Sun, P., Qi, H., Lyu, S. (2020). Celeb-DF: A Large-scale Challenging Dataset for Deepfake Forensics. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 3207-3216.

MacCartney, B. (2009). *Natural Language Inference*. Stanford University.

Milani, S., Fontani, M., Bestagini, P., Barni, M., Piva, A., Tagliasacchi, M., Tubaro, S. (2012). An Overview on Video Forensics. *APSIPA Transactions on Signal and Information Processing*, 1, e2.

Neves, J.C., Tolosana, R., Vera-Rodriguez, R., Lopes, V., Proença, H., Fierrez, J. (2020). GANPrintR: Improved Fakes and Evaluation of the State of the art in Face Manipulation Detection. *IEEE Journal of Selected Topics in Signal Processing* 14.5, pp. 1038-1048.

Nørregaard, J., Horne, B.D., Adal, S. (2019). Nela-gt-2018: A Large Multi-labelled News Dataset for the Study of Misinformation in News Articles (2019). In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 13, pp. 630–638.

Pastor-Galindo, J., Zago, M., Nespoli, P., Bernal, S. L., Celdrán, A. H., Pérez, M. G., Ruipérez-Valiente, J.A., Pérez, G.M., Mármol, F. G. (2020). Spotting political social bots in Twitter: A use case of the 2019 Spanish general election. *IEEE Transactions on Network and Service Management*, 17(4), 2156-2170.

Piva, A. (2013). An Overview on Image Forensics. *International Scholarly Research Notices*.

Qin, C., Zhang, A., Zhang, Z., Chen, J., Yasunaga, M., Yang, D. (2023). Is Chatgpt a General-purpose Natural Language Processing Task Solver? *arXiv preprint arXiv:2302.06476*.

Rangel, F., Giachanou, A., Ghanem, B., Rosso, P. (2020) Overview of the 8th Author Profiling Task at PAN 2020: Profiling Fake News Spreaders on Twitter. In: L. Cappellato, C. Eickhoff, N. Ferro, and A. Névéol (eds.) *CLEF 2020 Labs and Workshops, Notebook Papers*. CEUR Workshop Proceedings. CEUR-WS.org, vol. 2696.

Rathgeb, C., Tolosana, R., Vera-Rodriguez, R., Busch, C. (2022). *Handbook of Digital Face Manipulation and Detection: From DeepFakes to Morphing Attacks*. Springer.

Robertson, S.E., Walker, S., Jones, S., Hancock-Beaulieu, M.M., Gatford, M., et al. (2015). Okapi at Trec-3. Nist Special Publication Sp, 109.

Rodriguez, J., Hay, T., Gros, D., Shamsi, Z., Srinivasan, R. (2022). Cross-Domain Detection of GPT-2-Generated Technical Text. In Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. Association for Computational Linguistics, pp. 1213–1233.

Rossler, A., Cozzolino, D., Verdoliva, L., Riess, C., Thies, J., Nießner, M. (2019). FaceForensics++: Learning to Detect Manipulated Facial Images. In Proceedings of the IEEE/CVF international Conference on Computer Vision, pp. 1-11.

Ruffo, G., Semeraro, A., Giachanou, A., Rosso, P. (2023). Studying Fake News Spreading, Polarisation Dynamics, and Manipulation by Bots: A Tale of Networks and Language. *Computer Science Review*, 47: 100531.

Scherhag, U., Rathgeb, C., Merkle, J., Breithaupt, R., Busch, C. (2019). Face Recognition Systems Under Morphing Attacks: A Survey. *IEEE Access*, 7, pp. 23012–23026.

Shao, C., Ciampaglia, G. L., Varol, O., Yang, K. C., Flammini, A., & Menczer, F. (2018). The spread of low-credibility content by social bots. *Nature communications*, 9(1), 1-9.

Shu, K., Sliva, A., Wang, S., Tang, J., Liu, H. (2017). Fake News Detection on Social Media: A Data Mining Perspective. *ACM SIGKDD Explorations Newsletter*, 19(1), pp. 22–36.

Shu, K., Mahudeswaran, D., Wang, S., Lee, D., Liu, H. (2020). Fakenewsnet: A Data Repository with News Content, Social Context, and Spatiotemporal Information for Studying Fake News on Social Media. *Big Data*, 8(3), pp. 171–188.

Soleimani, A., Monz, C., Worring, M. (2020). Bert for Evidence Retrieval and Claim Verification. In *Advances in Information Retrieval: 42nd European Conference on IR Research, ECIR 2020, Lisbon, Portugal, April 14–17, 2020, Proceedings, Part II* 42, pp. 359–366.

Tolosana, R., Vera-Rodriguez, R., Fierrez, J., Morales, A., Ortega-Garcia, J. (2020). Deepfakes and beyond: A Survey of Face Manipulation and Fake Detection. *Information Fusion*, 64, pp. 131-148.

Tolosana, R., Romero-Tapiador, S., Vera-Rodriguez, R., Gonzalez-Sosa, E., Fierrez, J. (2022). DeepFakes Detection across Generations: Analysis of Facial Re-

gions, Fusion, and Performance Evaluation. *Engineering Applications of Artificial Intelligence*, 110, 104673.

Tourille, J., Sow, B., Popescu, A. (2022). Automatic Detection of Bot-Generated Tweets. In *Proceedings of the 1st International Workshop on Multimedia AI against Disinformation*. Association for Computing Machinery, pp. 44--1.

Uchendu, A., Le, T., Shu, K., Lee, D. (2020). Authorship Attribution for Neural Text Generation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 8384--8395

Vallas, S., Schor, J. B. (2020). What do Platforms do? Understanding the Gig Economy. *Annual Review of Sociology*, 46, pp. 273-294.

Verdoliva, L. (2020). Media Forensics and Deepfakes: An Overview. *IEEE Journal of Selected Topics in Signal Processing*, 14(5), pp. 910-932.

Vosoughi, S., Roy, D., Aral, S. (2018). The Spread of True and False News online. *Science*, 359(6380), pp. 1146-1151.

Wang, W.Y. (2017). "Liar, Liar Pants on Fire": A New Benchmark Dataset for Fake News Detection. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pp. 422--426, 2017.

Zeng, X., Abumansour, A.S., Zubiaga, A. (2021) Automated Fact-checking: A Survey. *Language and Linguistics Compass*, 15(10):e12438.

Zhang, G., Giachanou, A., Rosso, P. (2022). SceneFND: Multimodal Fake News Detection by Modelling Scene Context Information. *Journal of Information Science*, 01655515221087683.


Zhang, J., Carpenter, D., Ko, M. (2013). Online Astroturfing: A Theoretical Perspective. In *Proceedings of the Nineteenth Americas Conference on Information Systems*, Chicago, Illinois, August 15-17.

Zhou, Z., Guan, H., Bhat, M., Hsu, J. (2019) Fake news detection via NLP is vulnerable to adversarial attacks, *arXiv preprint arXiv:1901.09657*.

Zubiaga, A., Aker, A., Bontcheva, K., Liakata, M., Procter, R. (2018). Detection and Resolution of Rumours in Social Media: A Survey. *ACM Computing Surveys (CSUR)*, 51(2), pp. 1--36.

Capítulo 6

TÉCNICAS Y METODOLOGÍAS
DE CIBERINTELIGENCIA PARA LA
INVESTIGACIÓN DE OPERACIONES DE
INFORMACIÓN Y GUERRA COGNITIVA



Coordinadores:

Javier Valencia Martínez de Antoñana

Ministerio del Interior - Oficina de Coordinación de Ciberseguridad (OCC)

Autores y colaboradores:

David Arroyo Guardado

Borja Díaz-Merry Rivas

Carlos Galán Cordero

Agustín Valencia Gil-Ortega

Fernando J. Valladares Ros

Fernando Villena Sánchez

Ministerio de Defensa - Centro de Inteligencia de las Fuerzas Armadas (CIFAS)

Ministerio del Interior - Dirección General de la Policía – CGI



INTRODUCCIÓN

El problema en su contexto

La guerra cognitiva (*Cogwar*) se puede definir funcionalmente como “la militarización de la opinión pública, por parte de una entidad externa, con el fin de (1) influir en la política pública y gubernamental y (2) desestabilizar las instituciones públicas” (Bernal et.al., 2020).

Esencialmente, es la combinación de las nuevas técnicas cibernéticas asociadas a la guerra de la información (*IW*) y los componentes humanos del poder blando (*soft power*), junto con los aspectos de manipulación de las operaciones psicológicas (*PSYOPS*), implicando, por lo general, una presentación sesgada de una realidad, usualmente alterada digitalmente, destinada a favorecer los propios intereses.

Las nuevas herramientas de comunicación ofrecen en la actualidad infinitas posibilidades para el desarrollo de los conceptos anteriores y han expandido y creado un nuevo panorama de difusión de la información: en el que millones de personas se comunican de forma instantánea y en tiempo real.

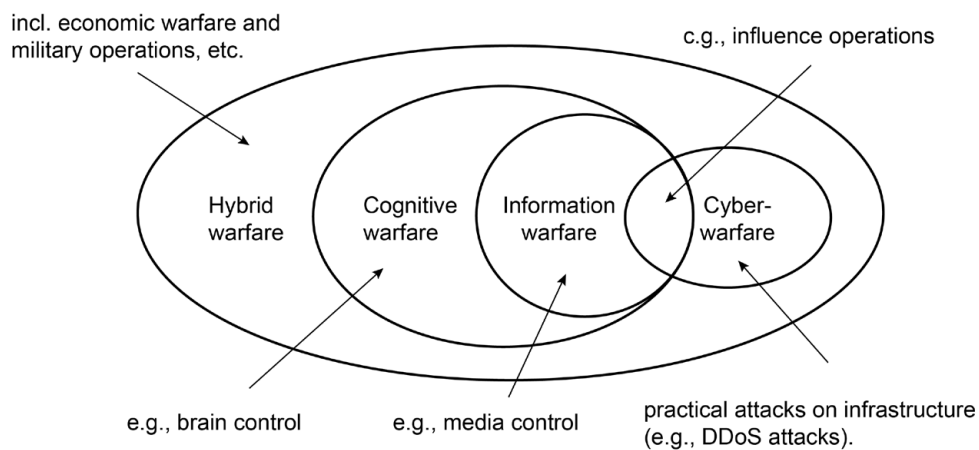


Figura 1. La relación conceptual entre la guerra cognitiva y otros tipos de guerra (Hung., 2022)

Como se desprende del texto y de la figura anterior, la guerra de información (IW) y la guerra cibernética (CW) son dos conceptos que pueden estar o no estrechamente relacionados y utilizarse en conjunto para alcanzar objetivos específicos. Pero sea como fuere, existen aspectos comunes que apuntan a un objetivo mayor: debilitar al oponente. Los ataques cibernéticos se incluyen en la doctrina de la cogwar rusa y china, dentro de lo que se conoce como “confrontación de la información” (Treyger et al., 2022) y “guerra informatizada” (Harold et al., 2021), respectivamente.

Hay varios tipos de ataques cibernéticos que se pueden utilizar en la guerra cibernética y que se han documentado englobadas en acciones de IW, como *phishing*, *malware*, *DDoS*, botnets, etc., que pueden combinarse, a su vez, con campañas de desinformación para influir en la opinión pública (Yuan, 2022).

En resumen, tanto la IW y la CW pueden trabajar conjuntamente para lograr objetivos específicos y, si se utilizan adecuadamente, pueden ser muy efectivas y peligrosas. Y, por supuesto, una de las armas más utilizadas en este tándem, sería la desinformación.

La desinformación como amenaza

La desinformación, en tanto que supone una amenaza dirigida contra la soberanía de un estado, ha de entenderse como una de las herramientas/amenazas usadas en IW.

Estos instrumentos/herramientas se usan coordinadamente al servicio del objetivo estratégico del agresor contra la soberanía e intereses del Estado agredido, y su uso se encuadra en tres aspectos que definen la naturaleza del conflicto y que conviene tener muy presentes: la ambigüedad, el gradualismo en la ejecución y la estrategia híbrida.

En este marco de acción, la desinformación juega un papel fundamental entre otros motivos, porque se ajusta perfectamente a los tres aspectos señalados más arriba, además de por ser un medio económico al alcance de cualquier actor activo en una estrategia ofensiva.

Influir, desestabilizar, manipular son objetivos fundamentales de estas acciones híbridas que, según la Estrategia Nacional de Ciberseguridad, quedan definidas como *“acciones coordinadas y sincronizadas dirigidas a atacar de manera deliberada las vulnerabilidades sistémicas de los estados democráticos y las instituciones, a través de una amplia gama de medios, tales como acciones militares tradicionales, ciberataques, operaciones de manipulación de la información, o elementos de presión económica”*.

En este sentido, surge el concepto de Manipulación e Interferencia de Información Extranjera (*FIMI*), propuesto por el Servicio Europeo de Acción Exterior (*SEAE*), para describir un patrón de comportamiento, en su mayoría no ilegal, que amenaza o tiene el potencial de impactar negativamente en los valores, procedimientos y procesos políticos. Esta actividad tiene un carácter manipulador y es llevada a cabo con una intencionalidad y de manera coordinada y sistemática. Lo que supone una clara amenaza.

Téngase siempre presente que una campaña de desinformación, por sí sola, difícilmente va a alcanzar sus objetivos y entraría a ser calificada, más bien, como parte del juego político dentro de la competición en el continuum (López-Lago López-Zuazo, 2021).

Una campaña de desinformación tiene en su punto de mira a la población del estado agredido, así como a la opinión pública internacional. Para cada uno de ambos objetivos la metodología se adaptará convenientemente.

Para la primera tendrán especial uso la explotación de las vulnerabilidades de la sociedad del estado agredido, como por ejemplo la polarización política, la existencia de disidencias políticas o de separatismos independentista, o la existencia de agravios reales o no, etc.

En cambio, con la opinión pública internacional, se usarán con profusión la comunicación estratégica, la diplomacia bi/multilateral o el *lawfare*, etc.

La detección temprana de este tipo de acciones no es sencillo debido, fundamentalmente, a esta ambigüedad, a la asimetría de estas y a que no suelen presentarse de manera aislada y van acompañadas por otros tipos de herramientas (ciberataques, etc.) como ya se ha visto.

Por todo lo anterior, si somos capaces de valernos de los protocolos y los desarrollos que se han implementado en la ciberguerra y en el campo de la ciberseguridad para detectar, de una manera temprana, los distintos ataques cibernéticos, contemplando y añadiendo los elementos intrínsecos a la IW, utilizando también técnicas de ciberinteligencia, seremos capaces de desarrollar un principio de sistema de alerta temprana de detección de acciones de IW, fundamentalmente orientadas a la detección de acciones de desinformación (Terp y Breuer, 2022).

Este es el objetivo fundamental de este documento: analizar los principales componentes que intervienen en este tipo de acciones de cara a diseñar un primer sistema de alerta temprana para su detección.

FASE I: ANÁLISIS DE ACCIONES DETECTADAS PARA LA DEFINICIÓN DE ELEMENTOS Y ACTORES PARTICIPANTES

Las acciones de desinformación son estrategias a largo plazo que necesitan del apoyo de un amplio abanico de medios si pretende ser efectiva. Hoy en día, este tipo de estrategias no estarían al alcance de cualquiera.

Esos medios se catalogan desde el punto de vista técnico, económico y también organizativo, que, posteriormente, pueden ser identificados en distintos tipos de técnicas, tácticas y procedimientos, asociado a determinados tipos de actores. A continuación, se mostrará de manera muy resumida, cuáles pueden ser estas componentes asociadas a acciones de desinformación.

La componente técnica de las acciones. Técnicas, tácticas y procedimientos.

Las técnicas, tácticas y procedimientos de desinformación son muy variados y se han ido sofisticando con el tiempo. Además, su posible atribución a una hipotética acción no es sencilla ya que estas técnicas no siempre se utilizan exclusivamente para la desinformación y pueden ser utilizadas de manera legítima para otros fines (Landwehr, 2020).

Sin embargo, cuando se utilizan con el objetivo de difundir información falsa o engañosa, pueden ser altamente perjudiciales para la sociedad y la democracia.

Dentro de las diferentes técnicas, tácticas y procedimientos (TTP) que se han detectado a lo largo de la historia de la desinformación, destacan fundamentalmente las siguientes:

- **VPNs y proxies** (Kuzio, 2021)
- **Plataformas de proxies de publicación (aldeas Potemkin)** (Brookhouser, 2021; Pamment et al., 2018)
- **Efecto woozle** (Carrasco Rodríguez, 2020)
- **Smurfing** (David, 2020)
- **Granjas de trolls** (Denter y Ginzburg, 2021)
- **Bots** (Shao et al., 2017)
- **Soft power** (El-Badawy et al., 2022)
- **Medios de comunicación ficticios** (Palau Sampio y Carratalá, 2022)

- **Deepfakes** (Vaccari y Chadwick, 2020)
- **Clickbait** (Bourgonje et al., 2017)
- **Astroturfing** (Zhang et al., 2013)
- **Amplificación de la señal** (Jungherr y Schroeder, 2021)
- **Microtargeting** (Arsenault, 2020)
- **Fakenews** (Posetti y Matthews, 2018)
- **Desacreditación** (Bader, 2018)
- **Infiltración y manipulación** (Freelon y Lokot, 2020)
- **Ciberataques y hackeos** (Eddy, 2016)
- **Campañas de intimidación** (Schmidt-Felzmann, 2017)
- **Conspiraciones** (Innes e Innes, 2021)
- **Manipulación de traducciones** (Phanthaphoommee, 2023)
- **Apropiación indebida** (Asperholm Hedlund, 2019)
- **Etiquetado del “enemigo” basado en la ficción** (Li, 2023)
- **Suplantación de identidad**
- **Desestimar**
- **Distorsionar**
- **Distraer**
- **Amedrentar**
- **Dividir**

Esas TTP se pueden utilizar conjunta o individualmente. En base a las TTP anteriormente señaladas, se muestran algunos de los posibles indicadores que pueden evidenciar que estamos ante una campaña de desinformación (Carrasco Farré, 2022):

- **Cronología:** La información dañina se publica simultánea o consecutivamente en varios medios.
- **Similitud estructural:** Los artículos periodísticos son idénticos o contienen información distribuida de forma similar.
- **Autoría:** Los artículos sospechosos fueron escritos por el mismo autor.
- **Referencias:** Un artículo publicado por un medio de comunicación sospechoso cita o incluye hipervínculos a otra u otras fuentes sospechosas.

- **Recurrencia:** un medio de comunicación sospechoso comparte repetidamente y en un breve espacio de tiempo, artículos de otros medios previamente identificados como proxies.

La componente económica de las acciones

La Universidad de Oxford publicó en 2020 un documento que identificaba a los países que poseen una gran capacidad para realizar campañas de desinformación y sobrados recursos humanos y presupuestarios para desarrollar actividades de influencia y operaciones psicológicas, entre ellos, Rusia, China e Irán (Bradshaw et al., 2020).

En el caso de Rusia, ha habido un aumento constante y generalizado en la financiación de los medios de comunicación con cargo al presupuesto estatal:

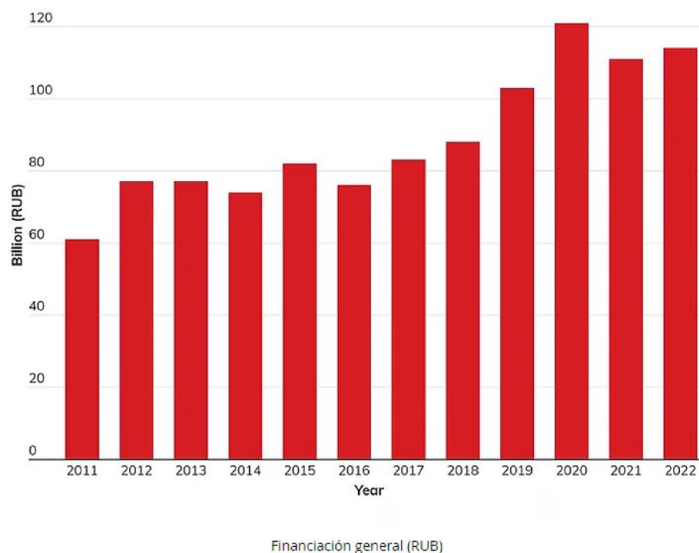


Figura 2. Financiación rusa de los medios de comunicación con cargo al presupuesto estatal (Michałowska y Kubś, 2022)

Hemos de tener en cuenta que esa financiación se refiere a los medios oficiales y no a los proxies o medios que ocultan su vinculación directa con el Kremlin. Algo similar ocurre con China, que dispone de una amplia cobertura de medios de comunicación oficiales encargados de “contar bien la historia de China” (Observatorio de la Política China, 2016).

Tanto China como Rusia han invertido mucho en cortejar nuevas audiencias en América Latina y África; muy evidente en este último continente, en donde se han detectado acciones

en francés coordinadas por Rusia que abarcan Facebook, YouTube, Telegram y otros canales en línea (Atanesian, 2023), con el propósito de dañar la imagen de Francia (Chrisafis, 2023) y mejorar la imagen de Putin, difundiendo mensajes a favor del Kremlin (Bond, 2023).

Los medios estatales chinos han sido fundamentales en los esfuerzos de Beijing por influir en otros países, controlar la información sobre el Partido Comunista y poder amplificar las narrativas de China sobre sus políticas y su papel en el mundo (Freedom House, 2022). Efectivamente, el régimen chino ha construido una sofisticada estrategia para retratar el liderazgo del país bajo una luz favorable, invirtiendo desde 2009 en torno a 6.600 millones de dólares para fortalecer su presencia global en los medios (South China Morning Post, 2009). Según Bloomberg News, entre 2008 y 2018, la inversión china, sólo en medios, ascendió a 2.800 millones de dólares (Tartar et al., 2028).

Además de esto, China lleva a cabo regularmente programas de intercambio para periodistas extranjeros de varios países, organiza formaciones para periodistas en ciudades chinas y mantiene debates periódicos entre periodistas extranjeros y sindicatos de medios chinos.

No obstante, Beijing también usa tácticas inusuales, tales como proporcionar contenido de los medios estatales de forma gratuita, costear suplementos completos en reputados periódicos extranjeros, y la firma de acuerdos de cooperación bilateral con medios de comunicación locales (Kumar, 2021).

En el caso de aquellos medios y/o analistas y personal no directamente integrados dentro del Estado, resulta fundamental realizar un análisis concreto de cómo se financian. Identificados determinados descriptores relacionados con la financiación, de cara al desarrollo de un sistema de alerta temprana podrían servir como vectores a relacionar con TTP de desinformación.

La componente organizativa

Las potencias/democracias iliberales revisionistas buscan en primer lugar garantizar su propia seguridad interna, ya que entienden dicho Orden Liberal como una amenaza para la supervivencia de su propio régimen. En segundo lugar, aspiran a modificar el Orden Internacional Liberal, un orden que les es muy desfavorable y que no pueden desafiar por medios violentos clásicos, como el recurso a la guerra.

En la época que vivimos, cabría bien pensar que la IW que se libra hoy, en la que la desinformación tiene un papel protagonista, no es sino uno de los frentes de una guerra política a nivel global. Un campo de batalla entre la visión occidental del mundo y la de las “democracias iliberales”.

Los actores de las amenazas objeto de este documento pueden ser actores estatales, no estatales o apoderados. La atribución política y la técnica pueden identificarse a partir de

declaraciones o informes oficiales, o mediante investigaciones de las víctimas o servicios de la sociedad civil.

La promoción de narrativas es uno de los métodos más efectivos para lograr un dominio estratégico a largo plazo, para lo cual los actores establecen unos objetivos claramente definidos, medibles y alcanzables. Para ello, seleccionarán las plataformas y los activos de información más apropiados según sus posibilidades tecnológicas, algoritmos, términos de servicio y tipos de contenido permitidos para promover sus mensajes directamente al público objetivo. Tras establecer activos que generan confianza, maximizarán la exposición de la audiencia al contenido a través de la amplificación de estas narrativas.

Los actores desinformativos dirigen los diferentes mensajes a través de distintas lenguas en función de cuál sea su público objetivo. Del mismo modo, difunden su ideario desinformativo a través de canales oficiales de comunicación, como servicios diplomáticos o embajadas, y canales controlados o vinculados al Estado, cuya afiliación no tiene por qué ser oficial ni transparente, pero su atribución ha sido confirmada por otras organizaciones, entidades de inteligencia o servicios militares. Estos canales son relevantes ya que contribuyen intencional o accidentalmente al éxito de los objetivos del actor de la amenaza.

También distintos canales diplomáticos pueden ser utilizados para facilitar estas operaciones de manipulación e influencia, especialmente en aquellos países donde han sido cerrados otros instrumentos de desinformación (véase el bloqueo de RT y Sputnik por la UE). Rusia y otros Estados autoritarios han explorado también su presencia en foros diplomáticos para legitimar sus posiciones y ganar influencia internacional.

Además de utilizar esta vía, se han utilizado las suplantaciones de identidad de medios de comunicación, organizaciones o individuos internacionales para exponer sus mensajes.

La desinformación se propaga a través de distintos formatos, como videos, *deepfakes*, imágenes fabricadas, caricaturas, noticias e incluso contenido de humor o entretenimiento para llegar a un público mayor.

Entre las plataformas de redes sociales y sitios web, los actores desinformativos utilizan Telegram, Twitter y Facebook para distribuir su contenido, pero también plataformas de video como Youtube o TikTok, así como Instagram, blogs o medios propagandísticos con el fin de difundir esas narrativas que distorsionan los hechos y la realidad, fomentando el miedo y el odio y generando división en las sociedades.

Otro ejemplo paradigmático de lo compleja que puede llegar a ser, a nivel organizativo, una estrategia de desinformación es el caso de China. Las operaciones de influencia de China han variado desde la diplomacia pública, hasta recurrir a personas influyentes en las redes sociales con conexiones no reveladas con los medios o la estructura estatal, o la promoción de una imagen positiva del país en suplementos de medios de comunicación internacionales.

Beijing, que libra una “guerra inteligente” a través de un enfoque estratégico a largo plazo, promociona también su cultura a través de ofertas de inversión de capitales, plataformas de videojuegos y redes sociales como TikTok para llegar a un público más fácilmente manipulable como el adolescente, intercambios culturales o centros de estudios como el Instituto Confucio para proyectar su imagen internacional.

Posteriormente, es muy común que la amplificación de narrativas se lleve a cabo también alimentando a sectores marginales o extremistas de la sociedad con la pretensión de que desencadenen una reacción que agrave las fracturas sociales.

Y a todo lo anterior, sería necesario tener en cuenta a actores internos de carácter separatista, extremista o conspirativo intentando también moldear la opinión pública y el comportamiento de los ciudadanos, distorsionando o intensificando conflictos ya creados, tergiversando el lenguaje y diseminando narrativas que siembren dudas sobre la veracidad o autenticidad de los hechos con el fin de enturbiar las aguas y abrumar al público. Cuando la autoridad de los datos y los hechos se erosiona, las emociones llenan ese vacío, y cuanto más se daña la confianza, el público puede ser más proclive a legitimar formas de gobierno no democráticas o implicarse en un activismo violento.

Y finalmente, es oportuno tener en cuenta los ejemplos de empresas de distintas nacionalidades involucradas en campañas de desinformación en todo el mundo.

Así pues, el organigrama organizativo de una acción de desinformación puede llegar a ser realmente complejo dependiendo de la tipología de actores utilizados en la misma. Esto complica, claramente, el desarrollo de cualquier tipo de sistema de alerta temprana, ya que el número de elementos sobre los que poner el foco es bastante considerable y dificulta en gran medida este proceso.

FASE II: DISEÑO Y DESARROLLO DE UN MODELO DE DETECCIÓN TEMPRANA DE HIPOTÉTICAS ACCIONES DE DESINFORMACIÓN

Como ya se ha podido comprobar a lo largo de este capítulo, desarrollar un modelo de alerta temprana puede ayudar a detectar patrones y tendencias en la difusión de información engañosa o manipulada en línea, lo que permitiría a los investigadores y a las autoridades intervenir antes de que la desinformación se propague ampliamente (Cullen y Wegge, 2021).

Conforme a lo anterior, un modelo así permitiría la identificación de cuentas o sitios web que están publicando información falsa de manera sistemática, el seguimiento de la propagación de contenido engañoso a través de las redes sociales y la identificación de grupos o redes que están trabajando activamente para difundir información falsa (Hounsel et al., 2020).

Existen varias herramientas y técnicas que se pueden utilizar para desarrollar un modelo de alerta temprana para detectar la desinformación, como el análisis de redes sociales, la monitorización de la actividad en línea y el uso de algoritmos de aprendizaje automático. Sin embargo, el escaso número de casos identificados y analizados como para cargar el peso de este tipo de sistemas en la inteligencia artificial, junto con la dificultad y la cantidad de recursos que conllevaría la monitorización continua de RR. SS. y medios digitales de comunicación, hace necesario encarar este problema desde una perspectiva más clásica del problema.

Teniendo en cuenta esta problemática, partiríamos de la idea de un grupo de analistas formados en la materia (*green group*), con una serie de descriptores (sitios web, perfiles de RR. SS., medios digitales, canales, actores, técnicas, tácticas, procedimientos, etc.) conocidos e identificados.

Sobre estos, se establecería una vigilancia y se localizarían una serie de indicadores que sugieren una campaña de desinformación en su etapa inicial y, mediante una fórmula o algoritmo que tenga en cuenta estos indicadores, ser capaces de establecer unos umbrales de alerta con unos índices de peligrosidad.

El Servicio Europeo de Acción Exterior (SEAE) ha propuesto a través de la división Stratcom una metodología de análisis sobre las actividades, los actores y los niveles de amenaza de los incidentes FIMI con el objetivo de detectar, analizar e interrumpir los intentos de actores extranjeros para manipular el entorno de la información e interferir en los procesos democráticos.

El SEAE entiende que no es necesario que la información difundida por los actores de la amenaza sea comprobablemente falsa o engañosa. El factor decisivo recae sobre el comportamiento engañoso o manipulado.

Este organismo europeo apuesta por desarrollar estándares ya existentes como STIX y DISARM, siempre que sea posible, con el fin de lograr un marco analítico común que permita facilitar la cooperación entre las distintas partes interesadas. A su juicio, las taxonomías comúnmente compartidas ayudan a crear un lenguaje común que facilita las discusiones en la comunidad y minimiza posibles malentendidos.

En concreto, aboga por un enfoque de “cadena mortal” (*kill chain*) para FIMI¹ y desinformación que analiza el comportamiento del actor de la amenaza como un proceso que reúne las fases de planificación, preparación, ejecución y evaluación.

Cada una de ellas incluye múltiples tácticas (TA), que a su vez contienen distintas técnicas (T). A su juicio, si ciertas combinaciones de TTP (procedimientos) resultaron exitosas para un atacante y proporcionaron un buen cálculo en términos de costo-beneficio, es probable que reutilice las mismas combinaciones, a menos que ciertas TTP se vuelvan más costosas o imposibles.

La reutilización de combinaciones de TTP puede ayudar a la atribución de los ataques al establecer el modus operandi o la “huella digital del comportamiento” de un actor. Para que una *kill chain* pueda ser interrumpida, es necesario analizar objetivamente el comportamiento y los TTP que utilizan los actores de la amenaza en sus esfuerzos por manipular el entorno de información. Este estudio permitirá comprender qué vulnerabilidades explotan con más frecuencia y cuáles tienen un impacto más perjudicial en la integridad de los valores y procesos democráticos.

Asimismo, es oportuno desarrollar y medir sistemáticamente respuestas disruptivas, investigar su eficiencia de cara a eliminar vulnerabilidades e interrumpir la logística de FIMI, y comprender los posibles efectos secundarios negativos de estas respuestas.

En un primer momento, una fase de monitoreo estratégico sirve para mapear el ecosistema de activos FIMI conocidos que son utilizados por un actor. Una segunda fase de priorización y clasificación ayudará a filtrar el gran volumen de actividad de estos canales FIMI para seleccionar aquella que sea de alta prioridad o potencialmente dañina.

A continuación, la fase de análisis de incidentes y recopilación de pruebas se centrará en analizar el incidente, incluyendo las conexiones entre los diferentes canales del ecosistema,

¹ Primer Informe SEAE sobre Información extranjera, manipulación y amenazas de interferencia. Febrero 2023. <https://euvsdisinfo.eu/uploads/2023/02/EEAS-ThreatReport-February2023-02.pdf>

la amplificación del contenido o las TTP utilizadas. Es importante archivar sistemáticamente las evidencias de la actividad para futuras referencias².

Seguidamente, la fase de puesta en común e intercambio de conocimientos tiene como objetivo maximizar la utilidad a corto y largo plazo del análisis. A corto plazo, se compartirá el análisis con las principales partes interesadas que están en condiciones de reaccionar de inmediato ante el incidente y, a largo plazo, se recopilará la información estandarizada en una base de datos que facilite el análisis comparativo de tendencias y patrones en la actividad de los actores de amenazas a lo largo del tiempo.

Finalmente, la conciencia situacional se consigue optimizando y analizando los pasos anteriores, expandiendo el monitoreo a canales recién atribuidos o a nuevos actores de amenazas, y observando patrones en la base de datos.

El SEAE combina objetos de datos STIX relacionándolo con el marco ABCDE de James Pamment, que propone estudiar los incidentes en términos de actores, comportamientos, contenido, grado y efecto. La recopilación sistemática de información de múltiples incidentes nos permitirá informar sobre el comportamiento o sobre qué combinación de técnicas es más probable que utilice el actor.

En el análisis del comportamiento se examinan los idiomas utilizados según la ubicación geográfica a la que haya sido dirigida la actividad, la duración del incidente, los canales de distribución, las TTP³ empleadas, las entidades objetivo y el propósito de la actividad FIMI.

Por su parte, la dimensión de grado intentará medir y describir la forma en que los incidentes FIMI viajan a través del entorno de información. De esta manera, se podrá conocer cuáles son las plataformas preferidas por los actores de la amenaza, identificar los roles de los canales, descubrir cómo se conectan los distintos ecosistemas desinformativos o quién se beneficia de las subredes, con el fin de mejorar nuestras investigaciones y contar con más indicadores para lograr la atribución en función de los patrones de comportamiento.

En cuanto al efecto del incidente, que hace referencia al impacto o gravedad, este se puede medir y evaluar de acuerdo con diferentes parámetros, como el alcance, el compromiso, el daño o el cambio de comportamiento causado fuera de línea. Los incidentes pueden tener lugar después de un evento o, bien, para enmarcar los próximos eventos.

Por último, el Curso de Acción se refiere a las acciones tomadas por cualquier entidad en respuesta a un incidente con el fin de contrarrestar su impacto, como la refutación del incidente, su desacreditación o labores de verificación, la eliminación del contenido o la suspensión del canal, entre otras respuestas.

² Ver Reglamento general de protección de datos <https://www.boe.es/doue/2016/119/L00001-00088.pdf> y Berkeley Protocol on Digital Open Source Investigations https://www.ohchr.org/sites/default/files/2022-04/OHCHR_BerkeleyProtocol.pdf.

³ Disarm Framework Explorer. <https://disarmframework.herokuapp.com/>

Elementos para evaluar

De cara al desarrollo de un modelo es necesario identificar los elementos que se van a considerar de cara a su modelización. En ese esquema, de un hipotético modelo, podríamos basarnos en:

- 1. Identificar los indicadores clave de compromiso** (IOC, por sus siglas en inglés) asociados con la desinformación: estos pueden incluir elementos como la propagación de información falsa o engañosa, el uso de cuentas falsas en las redes sociales, la manipulación de hashtags, la creación de sitios web falsos o la difusión de rumores. Estos IOC pueden ser identificados y representados en el formato de STIX para facilitar su análisis y seguimiento.
- 2. Establecer umbrales para alertas:** para determinar cuándo se debe emitir una alerta. Esto puede incluir criterios como el número de veces que se detecta un IOC en un período de tiempo determinado o la gravedad de la información falsa que se está propagando.
- 3. Incorporar inteligencia artificial y aprendizaje automático:** aunque inicialmente este tipo de técnicas genera problemas, tal y como ya se ha argumentado, a medio/largo plazo el uso de la inteligencia artificial se antoja fundamental de cara al reconocimiento automático o semiautomático de comportamientos compatibles con el despliegue de acciones de desinformación en su fase inicial.

En lo referente a la identificación de los indicadores clave de compromiso (IOC), algunos elementos que podrían ser útiles son:

- **Objetivos, propósito y motivaciones:** detrás de una campaña de desinformación. ¿Cuál es el mensaje que se está tratando de difundir y por qué? ¿Cuál es el objetivo final de la campaña?
- **Audiencia objetivo:** ¿A quién se está tratando de llegar con la información falsa o engañosa? ¿Qué factores podrían hacer que esta audiencia sea particularmente susceptible a la desinformación?
- **Fuentes de información:** Es importante investigar las fuentes de información utilizadas. ¿De dónde proviene la información falsa o engañosa? ¿Cómo se difunde esta información?
- **TTP utilizadas:** Es importante identificar las técnicas específicas utilizadas en una campaña de desinformación para poder abordarlas de manera efectiva, como patrones de actividad sospechosos (creación repentina de muchas cuentas en RR.SS. o la publicación repetida de mensajes similares desde múltiples cuentas, etc.)
- **Impacto y consecuencias:** Es importante evaluar el impacto y las consecuencias de una campaña de desinformación. ¿Qué efecto ha tenido la campaña en la opinión pública? ¿Ha creado confusión o alterado la percepción de la realidad?

¿Existen consecuencias más graves, como la alteración del proceso electoral o la incitación a la violencia?

- **Análisis de la narrativa:** La narrativa o la historia que se cuenta en una campaña de desinformación puede ser importante para entender su impacto y cómo está siendo utilizada para influir en la opinión pública. Esto podría incluir analizar cómo se presenta la información, qué argumentos se utilizan y qué imágenes o videos se comparten.
- **Descriptorios objetivos:** Identificación de temas o eventos que son objetivos comunes de la desinformación, como elecciones, conflictos internacionales o crisis de salud.
- **Idioma y tono:** Identificación de patrones de idioma y tono en los mensajes, como uso de lenguaje agresivo, emotivo o polarizado.

También, en el hipotético caso en que la atribución fuese factible, sería interesante tener en cuenta el tipo de actor que intenta desplegar esa acción. Si atendemos a la **filiación del actor**, tendríamos una clasificación inicial de actores estatales y actores no estatales (Van Esch et al., 2014). Y dentro de los últimos podríamos distinguir entre actores afiliados o no afiliados a un estado.

Metodología a utilizar de cara a desarrollar un modelo de detección temprana

Es importante tener en cuenta que la detección temprana de la desinformación es un desafío complejo y que los indicadores y algoritmos pueden necesitar ajustes y actualizaciones continuas para mantenerse efectivos frente a las tácticas en constante evolución utilizadas por los actores de estas acciones.

Una vez identificados estos indicadores, se puede desarrollar un algoritmo o fórmula que tenga en cuenta la presencia o ausencia de cada uno de ellos para generar alertas tempranas de posibles campañas de desinformación.

Para el desarrollo de este modelo, nos podemos apoyar en sistemas estandarizados para la detección y análisis de amenazas, como STIX, que permite a los analistas de seguridad cibernética y otros profesionales compartir información sobre amenazas en un formato estructurado y fácilmente legible por máquina que permite a los equipos de seguridad automatizar la integración de información. STIX se utiliza junto con otros lenguajes estandarizados como TAXII.

Además, existen marcos de referencia que utilizan el lenguaje STIX. DISARM STIX, por ejemplo, es un marco de trabajo que se utiliza para analizar y compartir información sobre amenazas de seguridad cibernética, incluyendo la desinformación.

A partir de aquí, existen numerosas posibilidades, desde las más simples hasta las más complejas. Podríamos considerar desde fórmulas basadas en la detección de patrones, o fórmulas basadas en el análisis de sentimiento, en la clasificación de la fuente, la credibilidad del canal de difusión, fórmulas de riesgo potencial, fórmulas de nivel de amenaza, etc.

Como puede apreciarse, dependiendo del enfoque o de la prioridad que se dé a alguno o algunos de los elementos a tener en cuenta, podemos basarnos en una fórmula o en otra. Este tipo de discriminación puede ser positivo a la hora de analizar distintas temáticas, distintos posibles actores, distintas circunstancias, etc.

Pero de la misma forma, podemos contemplar un modelo complejo en el que, de manera multicriterio, podemos contemplar varias de las anteriores fórmulas.

Desarrollo del modelo

El enfoque del modelo a desarrollar se podría afrontar desde un punto de vista:

- **Cuantitativo:** observando el distinto número de TTP detectadas en las fases iniciales de un hipotético ataque.
- **Cualitativo:** analizando la gravedad que dicho hipotético ataque podría llegar a tener.
- **Mixto:** con una combinación de las anteriores.

Si nos enfocamos en el marco de trabajo DISARM, disponemos de dos submarcos, el rojo (*red*) y el azul (*blue*). DISARM Red, para describir los comportamientos de los creadores de incidentes, y DISARM Blue, para describir los posibles comportamientos de respuesta. El objeto de este trabajo sería la creación de las bases de otro submarco, que habitualmente es denominado verde (*green*).

A su vez, se puede tratar el concepto de **incidente**. Un incidente se define como un evento adverso que compromete la confidencialidad, integridad o disponibilidad de un sistema o información.

Umbral en función del grado de desarrollo

Al tratarse de un sistema de alerta temprana, el marco de trabajo sobre el que basarse sería claramente DISARM Red. Si tenemos en cuenta este marco, existen las siguientes fases a analizar por parte del creador de incidentes:

- **P01:** Planificación
- **P02:** Preparación
- **P03:** Ejecución
- **P04:** Evaluación

Cualquier sistema de alerta temprana debería contemplar las dos primeras fases y, en gran medida, la tercera fase. No tendría mucho sentido en esta parte de evaluación temprana la fase de evaluación.

Fase P01: Planificación

Proyección del resultado deseado. Diseño de formas efectivas para lograrlo. Comunicar la visión, la intención y las decisiones, centrándose en los resultados esperados.

Fase P02: Preparar

Actividades realizadas antes de la ejecución para mejorar la capacidad de ejecución de la acción. Los ejemplos incluyen: desarrollo del ecosistema necesario para respaldar la acción: personas, red, canales, contenido, etc.

Fase P03: Ejecutar

Ejecución de la acción, desde la exposición inicial hasta el cierre y/o el mantenimiento de la misma.

Respecto a la fase en la que se ha detectado una hipotética acción, no sería lo mismo identificar TTP de las fases de planificación que TTP ya en fase de ejecución, que requeriría de un análisis e informe más acelerado. Así las cosas, se podría establecer un umbral de riesgo en función de la fase, del grado de desarrollo.

P01	P02	P03
BAJO	MEDIO	ALTO

Tabla 1. Umbral de riesgo en función del grado de desarrollo de la acción (fases relativas a TTPs detectadas)

En el caso de detectar TTP en distintas fases de desarrollo, se contemplará la más avanzada a la hora de valorar este umbral.

Umbral en función del número de ttp

Una vez contempladas las fases, el siguiente paso sería el análisis de TTP. Con respecto a las TTP, se podría tener como referente las correspondientes a DISARM:

- **Tácticas**⁴
- **Técnicas**⁵
- **Procedimientos**⁶

En este sentido, una métrica capaz de evaluar el riesgo de una hipotética acción sería el número de TTP detectadas en el incidente:

1-2	2-5	>5
BAJO	MEDIO	ALTO

Tabla 2. Umbral de riesgo en función del número de TTPs detectadas

Con respecto a este umbral, podría existir cierta correlación entre el número de TTP detectadas y el grado de desarrollo de la acción. Es de suponer, o se podría establecer la hipótesis de que, a mayores grados de desarrollo de una acción, mayor número de TTP potenciales a detectar.

Los números asignados a cada umbral se podrían reajustar en cualquier momento, de cara a flexibilizar el modelo ante distintas tipologías de situaciones. Por ejemplo, puede que el número de TTP utilizadas en un incidente relacionado con la guerra de Ucrania difiera considerablemente del número de TTP en otros incidentes y/o con respecto a otros países.

⁴ https://github.com/DISARMFoundation/DISARMframeworks/blob/main/generated_pages/tactics_index.md

⁵ https://github.com/DISARMFoundation/DISARMframeworks/blob/main/generated_pages/techniques_index.md

⁶ https://github.com/DISARMFoundation/DISARMframeworks/blob/main/generated_pages/tasks_index.md

Umbral de ponderación por repetición

Conforme se van analizando distintos incidentes, se desarrolla una base de datos muy interesante, con el objeto de establecer ciertos patrones o elementos repetitivos en las distintas acciones. Por ejemplo, el SEAE ha identificado, a partir del análisis de cientos de incidentes, que las principales TTP utilizadas en sus distintas fases de ejecución han sido:



Figura 3. TTPs únicas identificadas por el SEAE en función de la fase⁷

⁷ Véase en la URL https://www.eeas.europa.eu/sites/default/files/documents/2023/EEAS-DataTeam-ThreatReport-2023_.pdf

Además de lo anterior, se han detectado distintas combinaciones de TTP que han tenido un mayor grado de ocurrencia.

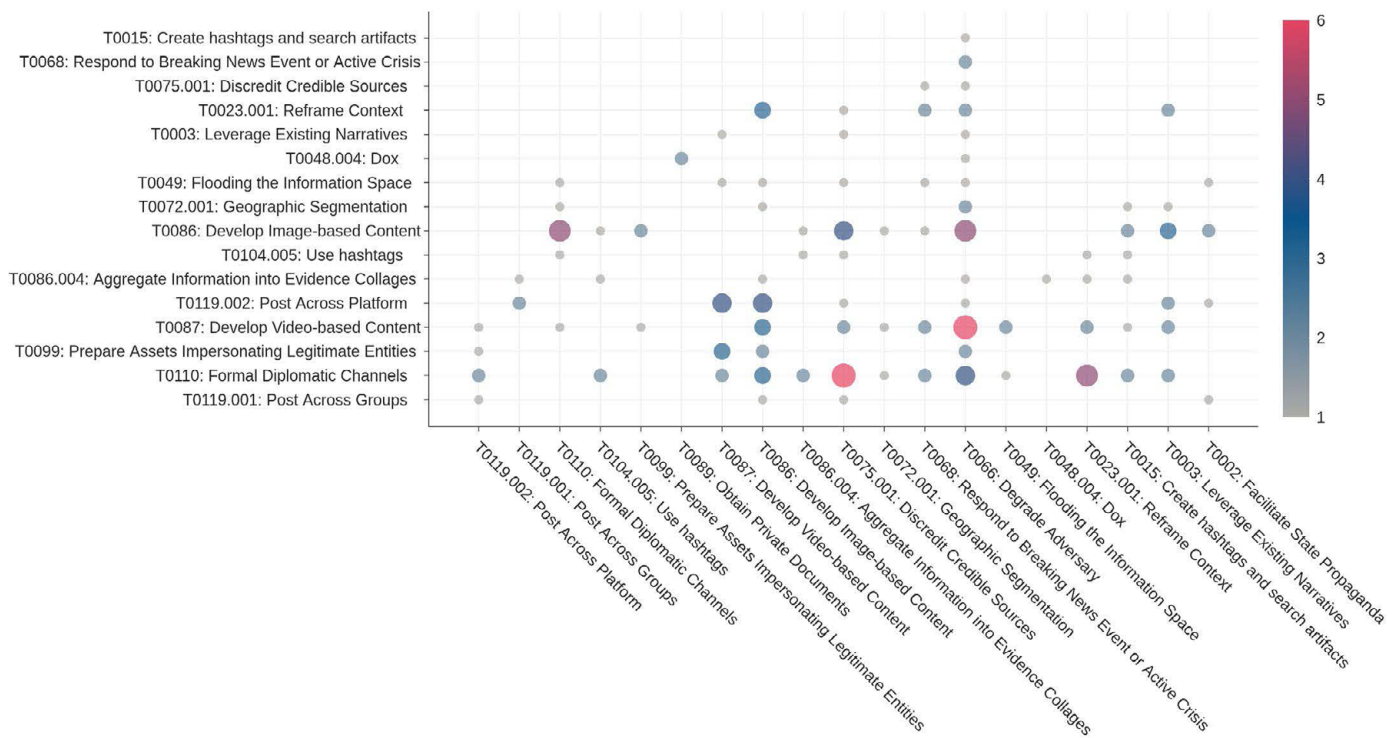


Figura 4. Combinaciones de TTPs y su frecuencia⁸

⁸ https://www.eeas.europa.eu/sites/default/files/documents/2023/EEAS-DataTeam-ThreatReport-2023_.pdf

Las principales combinaciones han sido las siguientes:

- G1** **TT0075.001+T0086**
T0086+T0119.002
T0087+T0119.002
T066+T0110
- G2** **T0023.001+T0110**
T0066+T0086
T0110+T0086
- G3** **TT0075.001+T0110**
T0066+T0087

Por ello, si tenemos discriminadas varias TTP asociadas a un incidente, podríamos desarrollar nuestro umbral de ponderación aplicando los siguientes pesos:

G1	G2	G3
1	2	3

Tabla 3. Ponderación en función de los grupos de combinaciones de TTPs detectadas

Análisis multicriterio en función de los distintos umbrales

MATRIZ1: FASES/TTP

MATRIZ1	P01	P02	P03
1-2	1	1	2
2-5	1	2	3
>5	2	3	3

Tabla 4. Matriz de valores respecto de las tablas 1 y 2

Una vez disponemos de la Matriz1, se podría enfrentar al umbral de ponderación por repetición:

MATRIZ2: MATRIZ1/PESOS

MATRIZ2	M1	M2	M3
G1	1	2	3
G2	2	4	6
G3	3	6	9

Tabla 5. Matriz de valores ponderados con los pesos de la tabla 3

Delimitada la segunda matriz, que contempla los valores obtenidos en la Matriz1 (que a su vez contemplaba las fases y el número de TTP detectadas), junto con los pesos dados a través del umbral de ponderación por repetición, podemos establecer cuatro niveles de riesgo:

NIVELES DE RIESGO SOBRE EL INCIDENTE

Nivel1: 1-2

Nivel2: 3-4

Nivel3: 6

Nivel4: 9

NIVEL1	NIVEL2	NIVEL3	NIVEL4
1-2	3-4	6	9

Tabla 6. Niveles de riesgo procedentes de la reclasificación de la tabla 5

Modificación supervisada del nivel de riesgo por otros factores. Atribución y efecto del incidente

Al tratarse de un modelo de alerta temprana, factores tan importantes como la atribución del incidente o el efecto y repercusión que tendrá, se antojan bastante complicados de incorporar al modelo. En estas etapas tempranas, la atribución, si no es clara, podría generar distorsión (caso de incidentes de falsa bandera). A su vez, el análisis de los efectos del incidente puede ser arriesgado ya que, en muchas ocasiones, los efectos del incidente no se observan hasta que ha finalizado el mismo.

Sin embargo, podría darse el caso de que incluso en esta fase de detección temprana, podríamos tener atribución sobre alguna de las TTP detectadas. En el caso de suceder esto, nos podríamos valer del listado de actores de DISARM⁹:

Y, dependiendo del tipo de actor detectado y su procedencia, se podría reconstruir de manera flexible el nivel de riesgo sobre el incidente. En este punto, podríamos tener en cuenta tanto las tipologías de actores como tablas de valoración de los mismos, dependiendo del país de procedencia. Esta forma de actuar permitiría que distintas administraciones, en función de sus atribuciones en esta materia, podrían trabajar con sus propios valores de peligrosidad de un determinado actor, sin afectar en ninguna medida el modelo que se plantea.

⁹ https://github.com/DISARMFoundation/DISARMframeworks/blob/main/generated_pages/actortypes_index.md

Así pues, los anteriores niveles de riesgo se podrían reorganizar de manera supervisada por un analista teniendo en cuenta un factor de SUMA o RESTA en función de las tablas de actores y procedencia (si las hubiese).

Así, se podría pasar de un nivel de riesgo calculado a cualquiera de los otros niveles si la atribución es positiva y así lo considera el analista.

$$\mathbf{NIVEL_{at} = NIVEL_i \pm n}$$

Dónde

$NIVEL_{at}$: nivel de riesgo recalculado por atribución

n : sería el número que el analista considera oportuno si se ha dado una atribución positiva.

Del mismo modo, se podría actuar en el caso de percibir distintas repercusiones o efectos que se pudieran inferir al inicio del incidente. Por ejemplo, se podrían detectar mensajes en los que se solicita a una determinada audiencia manifestarse por una causa determinada, o desarrollar insurgencias, protestas, etc. o incluso que el incidente fuese compatible con acciones de odio por distintas causas.

En este caso se podría actuar de manera similar a lo desarrollado en el caso de una atribución positiva. El analista, de manera supervisada, podría modificar los valores del nivel de riesgo tras observar estos hipotéticos efectos en el desarrollo del incidente. Para identificar los efectos, nos podríamos basar en el ya citado marco ABCDE para el análisis FIMI, según Pamment (2020).

Marco de trabajo ABCDE

Actor	¿Qué tipo de actores están involucrados? Esta pregunta puede ayudar a establecer, por ejemplo, si el caso involucra a un actor estatal extranjero.
Comportamiento	¿Qué actividades se exhiben? Esta investigación puede ayudar a establecer, por ejemplo, evidencia de coordinación e inautenticidad.
Contenido	¿Qué tipos de contenido se crean y distribuyen? Esta línea de preguntas puede ayudar a establecer, por ejemplo, si la información que se está implementando es engañosa.
Grado	¿Cuál es el impacto general del caso ya quién afecta? Esta pregunta puede ayudar a establecer los daños reales y la gravedad del caso.
Efecto	¿Cuál es el impacto general del caso ya quién afecta? Esta pregunta puede ayudar a establecer los daños reales y la gravedad del caso.

Tabla 7. Elementos del marco de trabajo ABCDE para el análisis FIMI según James Pamment

El componente de efecto del marco ABCDE utiliza indicadores de impacto para comprender qué grado de amenaza representa un incidente determinado.

Las preguntas útiles para hacer incluyen:

- **Clima de debate:** ¿El contenido en línea está basado en problemas? ¿Implica, por ejemplo, información falsa, polarización o troleo?
- **Confianza/reputación:** ¿El contenido está basado en el objetivo? ¿Involucra, por ejemplo, rumores falsos, hackeos de seguridad cibernética, falsificaciones y/o filtraciones de medios?

- **Libertades fundamentales:** ¿El contenido niega una libertad fundamental? Por ejemplo, ¿busca negar la libertad de expresión o de deliberación política?
- **Salud pública:** ¿el contenido amenaza la salud, el bienestar físico o la seguridad médica de las personas?
- **Seguridad pública:** ¿El contenido amenaza el bienestar físico de las personas o el orden público?
- **Integridad electoral:** ¿El contenido disuade a los votantes de participar en las elecciones o busca socavar los resultados de una elección?
- **Seguridad nacional:** ¿El contenido amenaza la integridad territorial o la seguridad nacional de un estado soberano?

Así, al igual que en el caso de la atribución positiva, cada administración con atribuciones podría desarrollar una tabla de riesgos por efectos, de cara a tener en cuenta en la valoración del nivel definitivo de un incidente.

Así, se podría pasar de un nivel de riesgo calculado a cualquiera de los otros niveles si se consideran efectos más graves o no, si así lo considera el analista.

$$\mathbf{NIVEL_{ef} = NIVEL_i \pm n}$$

Dónde

$NIVEL_{ef}$: nivel de riesgo recalculado por inferencia de efectos

n : sería el número que el analista considera oportuno en función de la valoración de la tabla de efectos conforme al marco ABCDE.

Esta forma de trabajar permitiría disponer de un modelo de detección temprana del riesgo de un incidente contemplando aspectos cuantitativos, pero además también cualitativos, que pueden ser independientes de la administración que los calcule y de una manera totalmente flexible de cara a no tener que reformular el modelo diseñado.

De cara a facilitar el uso de este sistema se facilita la siguiente infografía, en la que se muestran los distintos pasos y las interrelaciones entre unos y otros.

SISTEMA DE ALERTA TEMPRANA DE ACCIONES DE DESINFORMACIÓN

PASO A PASO

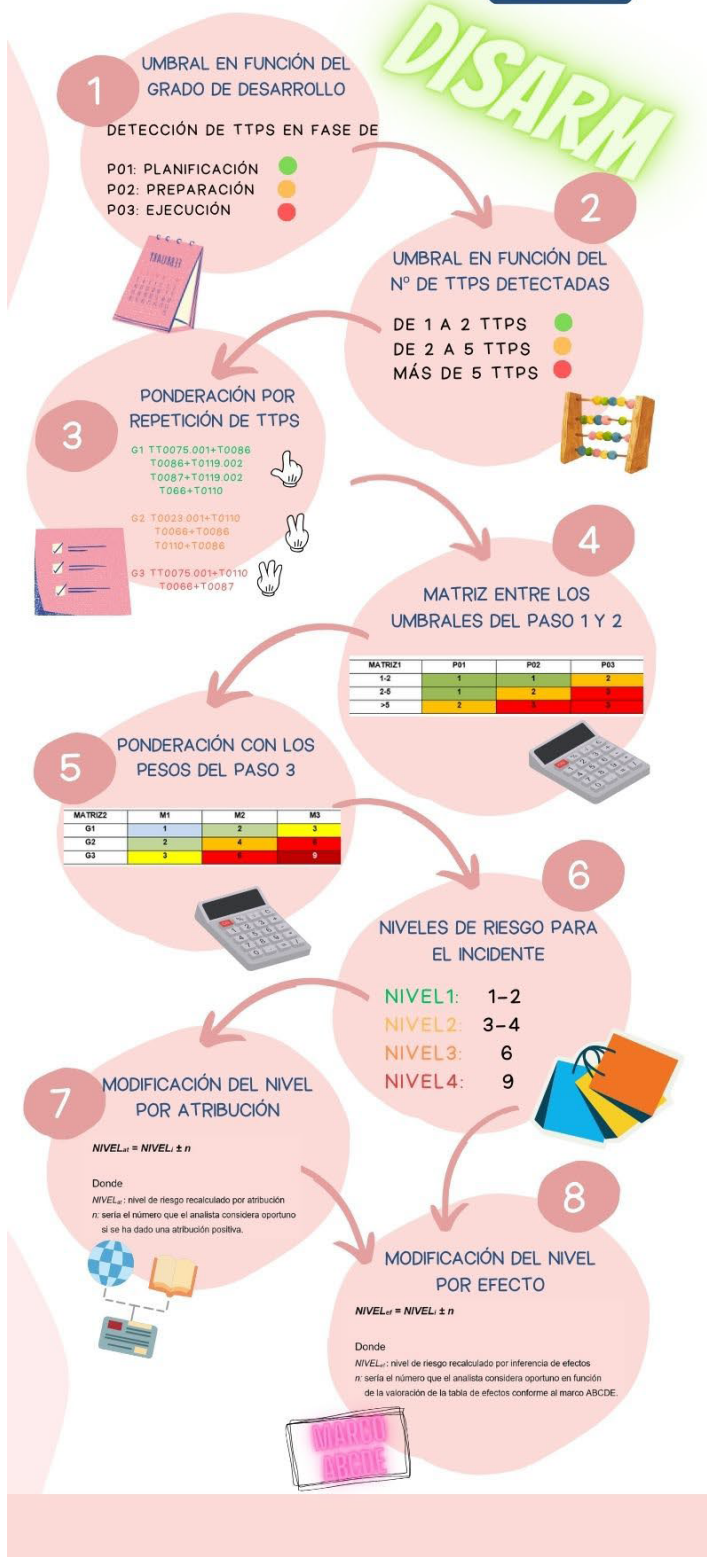


Figura 5. Infografía sobre el desarrollo del modelo (elaboración propia)

CONCLUSIONES

La desinformación se ha convertido en un desafío creciente en el mundo contemporáneo y su impacto en las sociedades democráticas es motivo de preocupación. A través del análisis exhaustivo y la reflexión profunda, a lo largo de este documento ha quedado perfectamente plasmada la peligrosidad inherente a estas acciones para un país democrático.

En primer lugar, la desinformación socava los fundamentos mismos de la democracia al distorsionar la verdad y manipular la percepción pública. Al sembrar la duda y el escepticismo sobre la veracidad de los hechos, los actores desinformativos pueden influir en la toma de decisiones, desestabilizar la confianza en las instituciones y desviar el curso de los procesos políticos legítimos. En última instancia, esto mina la capacidad de los ciudadanos para tomar decisiones informadas y participar de manera significativa en la vida política.

Además, la desinformación puede exacerbar divisiones sociales y fomentar el odio y la polarización. Al difundir contenido falso, engañoso o manipulado, los actores de desinformación buscan explotar las emociones y las diferencias existentes en una sociedad. Esto puede llevar a conflictos internos, fragmentación de la opinión pública y debilitamiento de los lazos que sustentan una sociedad cohesionada, incluyendo la cohesión territorial.

Por todo lo anterior, entendemos que la implementación de un sistema de alerta temprana para detectar y contrarrestar los ataques de desinformación puede suponer un arma efectiva y prometedora de cara a contrarrestar estas acciones.

Un sistema de este tipo sería capaz de monitorizar de forma proactiva las redes sociales, los medios de comunicación y otros canales de difusión de información en busca de patrones, tendencias y señales de desinformación. Al identificar rápidamente las campañas desinformativas, las autoridades competentes podrían emprender acciones eficaces para neutralizarlas y educar a la ciudadanía sobre los peligros de éstas.

La ventaja de contar con un sistema de alerta temprana radica en su capacidad para prevenir y mitigar los efectos perjudiciales de la desinformación en una sociedad democrática. Al permitir una respuesta ágil y basada en datos, se fortalecerían los mecanismos de defensa y se preservaría la integridad del proceso democrático. Además, un sistema de este tipo fomentaría la transparencia, la rendición de cuentas y el empoderamiento de los ciudadanos, al proporcionarles herramientas para identificar y resistir la manipulación informativa.

Siendo conscientes de la dificultad de desarrollar este tipo de protocolos, tanto por la dinámica de los mismos, sus distintas asimetrías, la cantidad de factores que intervienen y la complejidad de detectarlo en sus fases iniciales, se ha querido plantear un modelo abierto, apoyado en marcos de trabajo estandarizados y conocidos por la comunidad científica y académica. Con ello permitimos que, cualquier modificación del modelo o adaptación a nuevas variables, sea fácilmente abordable sin tener que abordar un diseño de modelo nuevo.

INVESTIGACIONES A FUTURO

A medida que se incorporan nuevas técnicas y sistemas relacionados con el desarrollo de la Inteligencia Artificial, aparecen nuevos vectores de ataque que permiten que puedan difundirse campañas de desinformación contrarias a los valores y al sistema democrático que se busca proteger. A su vez, los nuevos avances pueden ser aprovechados para mejorar la recopilación sistémica de los datos y mejorar la precisión de los modelos predictivos.

Las tecnologías que incluyen la realidad aumentada o la realidad virtual también pueden dar lugar al desarrollo de entornos falsos, recopilación de datos, difusión de narrativas o problemas de radicalización en línea (Kremidas-Courtney, 2022), mientras que la IA generativa de texto (ej. ChatGPT) puede ser manipulada para la fabricación de contenidos falsos.

El realismo de los *deepfakes* o “ultrafalso” (terminología alternativa al anglicismo *deepfake*) se incrementará previsiblemente con el desarrollo de los programas dotados de inteligencia artificial, siendo cada vez sus resultados más difíciles de detectar, ampliándose su utilización a múltiples campos, pudiendo ser vistos en streaming y siendo cada vez más fáciles de crear, lo cual puede acrecentar su disponibilidad de uso entre el público.

El riesgo de poder ser utilizados para tratar de manipular a los ciudadanos en vísperas de una cita electoral, o en respuesta a protocolos activados en emergencias civiles, o incluso también en el terreno bélico, multiplican los efectos catastróficos asociados a acciones de desinformación y de manipulación.

La explotación de esta tecnología junto con otras técnicas de ingeniería social pueden ser también empleadas de cara a sembrar la desinformación en círculos decisores de menor tamaño, pero que se encuentran en una posición estratégica, pasando esta actividad aún más inadvertida.

A pesar de que en la actualidad son más habituales los videos o audios manipulados que se han editado para imitar a una persona o modificar una escena, el aumento en la popularidad del video como una forma de recibir las noticias hace necesario adoptar medidas de seguridad para proteger a las personas e instituciones.

También es necesario tener en cuenta la creciente creación de cuentas falsas en las redes sociales, tanto dentro como fuera de nuestras fronteras, con el objetivo de intensificar la polarización y movilización de la sociedad. En ocasiones, estos actos son realizados por medios estatales, pero cada vez con mayor frecuencia, se hace uso de empresas privadas que ofrecen estos servicios.

Todos estos elementos suponen todo un reto a futuro de cara a combatir las acciones de guerra de información y, por supuesto, las acciones de desinformación.

BIBLIOGRAFÍA

- Arsenault, A. (2020). Microtargeting, Automation, and Forgery: Disinformation in the Age of Artificial Intelligence. *Affaires publiques et internationales – Mémoires, Public and International Affairs - Research Papers*. <https://hdl.handle.net/10393/40495>
- Asperholm Hedlund, L. (2019). *Identifying and Understanding Anti-Immigration Disinformation. A case study of the 2018 Swedish national elections*. Swedish Defense University. <https://www.diva-portal.org/smash/get/diva2:1324745/FULLTEXT01.pdf>
- Atanesian, G. (1 de febrero, 2023). *Russia in Africa: How disinformation operations target the continent*. BBC. <https://www.bbc.com/news/world-africa-64451376>
- Bader, M. (2018). Disinformation in Elections. *Security and Human Rights*, 29(1-4), 24-35. <https://doi.org/10.1163/18750230-02901006>
- Bernal, A., Carter C., Singh I., Cao K., Madreperla O. (2020). Cognitive Warfare: An Attack on Truth and Thought. *Innovation hub*. <https://www.innovationhub-act.org/sites/default/files/2021-03/Cognitive%20Warfare.pdf>
- Bond, S. (1 de febrero, 2023). *A pro-Russian social media campaign is trying to influence politics in Africa*. National Public Radio (NPR). <https://www.npr.org/2023/02/01/1152899845/a-pro-russian-social-media-campaign-is-trying-to-influence-politics-in-africa>
- Bourgonje, P., Schneider, J.M., y Rehm, G. (2017). From Clickbait to Fake News Detection: An Approach based on Detecting the Stance of Headlines to Articles. *In Proceedings of the 2017 EMNLP Workshop: Natural Language Processing meets Journalism*, 84–89. <https://dx.doi.org/10.18653/v1/W17-4215>
- Bradshaw, S., Campbell-Smith, U., Henle, A., Perini, A., Shalev, S., Bailey, H., y Howard, P.N. (2021). *Country case studies industrialized disinformation: 2020 global inventory of organized social media manipulation*. Oxford Internet Institute.
- Brookhouser, J.J. (2021). 'Scapegoat,' 'Proxy' and 'Base': A World Powers' Guide to Domestic Extremist Co-Optation. *Global Security and Intelligence Studies*, 6(2). <https://doi.org/10.18278/gsis.6.2.5>
- Carrasco Farré, C. (2022). The fingerprints of misinformation: How deceptive content differs from reliable sources in terms of cognitive effort and appeal to emotions. *Humanities and Social Sciences Communications*, 9(1), 1-18.

- Carrasco Rodríguez, B. (2021). *Information Laundering in the Nordic-Baltic region*. NATO Strategic Communications Centre of Excellence. <https://stratcomcoe.org/publications/information-laundering-in-the-nordic-baltic-region/26>
- Chrisafis, A. (27 de febrero, 2023). Macron pledges to reduce *French military presence in Africa*. The Guardian. <https://www.theguardian.com/world/2023/feb/27/macron-pledges-to-reduce-french-military-presence-in-africa>
- Cullen, P., y Wegge, N. (2018). Adapting Early Warning in an Age of Hybrid Warfare. En S. Stenslie, L. Haugom, B. H. Vaage (Eds.), *Intelligence Analysis in the Digital Age*. <https://doi.org/10.4324/9781003168157>
- David, A. (2020). *Smurf: a reliable method for contextualizing social media artefacts* [Tesis doctoral, Cranfield University]. <https://dspace.lib.cranfield.ac.uk/handle/1826/19058>
- Denter, P. y Ginzburg, B. (2021). Troll Farms and Voter Disinformation. MPRA Paper 109634, *University Library of Munich*. <https://dx.doi.org/10.2139/ssrn.3919032>
- Eddy, M. (8 de diciembre, 2016). *After a Cyberattack, Germany Fears Election Disruption*. The New York Times. <https://www.nytimes.com/2016/12/08/world/europe/germany-russia-hacking.html>
- El-Badawy, E., Munasinghe, S., Bukarti, A. B., y Bianchi, B. (23 de marzo, 2020). *Security, Soft Power and Regime Support: Spheres of Russian Influence in Africa*. Tony Blair Institute for Global Change. <https://www.institute.global/insights/geopolitics-and-security/security-soft-power-and-regime-support-spheres-russian-influence-africa>
- Freedom House. (30 de agosto, 2022). *Beijing's Global Media Influence 2022*. <https://freedomhouse.org/es/report/beijing-global-media-influence/2022/authoritarian-expansion-power-democratic-resilience>
- Freelon, D., y Lokot, T. (2020). Russian Twitter disinformation campaigns reach across the American political spectrum. *Harvard Kennedy School (HKS) Misinformation Review*. <https://doi.org/10.37016/mr-2020-003>
- Harold, S. W., Beauchamp-Mustafaga, N., & Hornung, J. W. (2021). *Chinese Disinformation Efforts on Social Media*. RAND PROJECT AIR FORCE SANTA MONICA CA.

Honsel, A., Holland, J., Kaiser, B., Borgolte, K., Feamster, N., y Mayer, J. (2020). Identifying disinformation websites using infrastructure features. *arXiv preprint arXiv:2003.07684*.

Innes, H., e Innes, M. (2023). De-platforming disinformation: conspiracy theories and their control. *Information, Communication & Society*, 26(6), 1262-1280, <https://doi.org/10.1080/1369118X.2021.1994631>

Jungherr, A., y Schroeder, R. (2021). Disinformation and the Structural Transformations of the Public Arena: Addressing the Actual Challenges to Democracy. *Social Media + Society*, 7(1). <https://doi.org/10.1177/2056305121988928>

Kremidas-Courtney, C. (23 de noviembre, 2022). *From post-truth to post-reality: the future of disinformation*. Friends of Europe. <https://www.friendsofeurope.org/insights/from-post-truth-to-post-reality-the-future-of-disinformation/>

Kumar, R. (2 de noviembre, 2021). *How China uses the news media as a weapon in its propaganda war against the West*. Reuters Institute. <https://reutersinstitute.politics.ox.ac.uk/news/how-china-uses-news-media-weapon-its-propaganda-war-against-west>

Kuzio, T. (2021). Disinformation: Soviet Origins of Contemporary Russian Ukrainophobia. <https://ekmair.ukma.edu.ua/server/api/core/bitstreams/6c5405eb-20b4-42dc-81c6-e71b2d3b2367/content>

Landwehr, H. (2022). *Analyzing the Usability of Natural Language Processing for Detecting Disinformation Tactics, Techniques, and Procedures* [Doctoral dissertation, Massachusetts Institute of Technology]. <https://hdl.handle.net/1721.1/144774>

Li, I. (06 de marzo, 2023). Constructing the Orc: Embracing Fiction in Ukraine's Tactical Narrative. Rusi.org. <https://rusi.org/explore-our-research/publications/commentary/constructing-orc-embracing-fiction-ukraines-tactical-narrative>

López-Lago López-Zuazo, M. (2021). La competición en el contínuum. *Boletín del Instituto Español de Estudios Estratégicos (IEEE)*, (22), 516-531.

Michałowska- Kubś, A. (8 de agosto, 2022). *Coining lies. Kremlin spends 1.5 Billion per year to spread disinformation and propaganda*. Debunk.org. <https://www.debunk.org/coining-lies-state-budget-financing-of-russian-propaganda>

Observatorio de la Política China. (31 de diciembre, 2016). *Presidente Xi insta a nuevo medio de comunicación a «contar bien las historias de China»*. Política-china.org. <https://politica-china.org/areas/sistema-politico/presidente-xi-insta-a-nuevo-medio-de-comunicacion-a-contar-bien-las-historias-de-china>

Palau Sampedro, D., y Carratalá, A. (2022). Injecting disinformation into public space: pseudo-media and reality-altering narratives. *Profesional de la información*, 31(3). <https://doi.org/10.3145/epi.2022.may.12>

Pamment, J., Nothhaft, H., y Fjällhed, A. (2018). *Countering Information Influence Activities: The State of the Art*. MSB. <https://www.msb.se/RibData/Filer/pdf/28697.pdf>

Pamment, J. (2020). *The EU's Role in Fighting Disinformation: Crafting A Disinformation Framework*. Carnegie Endowment for International Peace..

Phanthaphoommee, N. (2023). Incomplete translation as a conduit for fake news: a case of coronavirus-related news. *Humanities, Arts and Social Sciences Studies* 23(1), 18–35.

Posetti, J., y Matthews, A. (23 de julio, 2018). *A short guide to the history of 'fake news' and disinformation*. International Center for Journalists. <https://www.icfj.org/news/short-guide-history-fake-news-and-disinformation-new-icfj-learning-module>

Shao, C., Ciampaglia, G., Varol, O. Flammini, A., y Menczer, F. (2017). The spread of fake news by social bots. https://www.researchgate.net/publication/318671211_The_spread_of_fake_news_by_social_bots

Schmidt-Felzmann, Anke. (2017). More than 'just' disinformation. Russia's information operations in the Nordic region. En T. Cizik (ed.), *Information Warfare. New Security Challenge for Europe* (pp. 32-67). Centre for European and North Atlantic Affairs (CENAA).

South China Morning Post. (13 de enero, 2009). *Beijing in 45b yuan global media drive*. <https://www.scmp.com/article/666847/beijing-45b-yuan-global-media-drive>

Tartar, A., Rojanasakul, M., y Scott Diamond, J. (23 de abril, 2018). *How China Is Buying Its Way Into Europe*. Bloomberg. <https://www.bloomberg.com/graphics/2018-china-business-in-europe/>

Terp, S. y Breuer, P. (2022). DISARM: a Framework for Analysis of Disinformation Campaigns. *Conferencia IEEE de 2022 sobre aspectos cognitivos y computacionales de la gestión de situaciones (CogSIMA)*, 1-8, <https://doi.org/10.1109/CogSIMA54611.2022.9830669>

Treyger, E., Cheravitch, J., & Cohen, R. (2022). *Russian Disinformation Efforts on Social Media*. RAND CORP SANTA MONICA CA.

Tzu-Chieh H., y Tzu-Wei Hung. (2022). How China's Cognitive Warfare Works: A Frontline Perspective of Taiwan's Anti-Disinformation Wars. *Journal of Global Security Studies*, 7(4). <https://doi.org/10.1093/jogss/ogac016>

Vaccari, C., y Chadwick, A. (2020). Deepfakes and Disinformation: Exploring the Impact of Synthetic Political Video on Deception, Uncertainty and Trust in News. *Social Media + Society*, 6(1). <https://doi.org/10.1177/2056305120903408>

Van Esch, F. A. (2014). Exploring the Keynesian–Ordoliberal Divide. Flexibility and Convergence in French and German Leaders' Economic Ideas During the Euro-Crisis. *Journal of Contemporary European Studies*, 22(3), 288-302.

Yuan, Z., Chen, T., Ren, B., y Luo, L. (2022). Protecting Multi-Domain Command and Control Information System Network Traffic with Segment Routing. In *Proceedings of 2022 10th China Conference on Command and Control* (pp. 166-176). Singapore: Springer Nature Singapore.

Zhang, J., Carpenter, D., y Ko, M. (2013). Online astroturfing: A theoretical perspective. *19th Americas Conference on Information Systems, AMCIS 2013 - Hyperconnected World: Anything, Anywhere, Anytime*. 4. 2559-2565. <https://aisel.aisnet.org/amcis2013/HumanComputerInteraction/GeneralPresentations/5/>

Capítulo 7

LAS ACTUACIONES DE LAS
PLATAFORMAS DE INTERNET EN MATERIA
DE DESINFORMACIÓN.
DEL CÓDIGO DE BUENAS PRÁCTICAS AL
REGLAMENTO DE SERVICIOS DIGITALES



Coordinadores:

Alejandro Perales Albert

Ministerio de Asuntos Económicos y Transformación Digital - Secretaría de Estado de Digitalización e Inteligencia Artificial

Autores y colaboradores:

José Domingo Gómez Castallo

Raquel Jorge Ricart

Emilio Lliteras Arañó

Florentino Marín Parra

María Penedo Jiménez

Juan Carlos Suárez Villegas

Raquel Vinader Segura



PRESENTACIÓN

El presente capítulo recoge los principales aspectos del trabajo realizado del por este grupo de trabajo del Foro contra las campañas de desinformación, centrado en los compromisos asumidos por plataformas y redes sociales en el marco del Código europeo de buenas prácticas contra la desinformación (el inicial de 2018 y el reforzado de 2022), así como en las obligaciones para estas plataformas y redes establecido por el Reglamento de servicios digitales (2022).

La evaluación de los principales aspectos del Código de buenas prácticas se complementa con el análisis de los informes de rendición de cuentas presentados por las plataformas y redes firmantes del mismo, de acuerdo con lo establecido en el mencionado Código, teniendo en cuenta, en su caso, los datos aportados sobre España.

Finalmente, a modo de conclusión, se realizan una serie de propuestas de mejora del sistema de compromisos y rendición de cuentas en materia de desinformación de cara al futuro, especialmente desde el punto de vista de su aplicación en nuestro país.

PREFACIO: LA REGULACIÓN VOLUNTARIA EN LA UNIÓN EUROPEA

Las instituciones europeas llevan décadas promoviendo, como instrumentos complementarios a la legislación positiva, el desarrollo de sistemas de regulación voluntaria que permitan garantizar la defensa de la competencia leal y la protección de las personas consumidoras y usuarias de un modo más amplio y/o eficiente.¹

De un modo cada vez más decidido, la promoción de estos sistemas se orienta al modelo de corregulación, que trasciende al mero autocontrol o autodisciplina de profesionales y empresarios (autorregulación) para establecer espacios de confluencia en los cuales, a partir de un acto legislativo de la Administración, se confía la consecución de los objetivos definidos en dicho acto a los agentes involucrados, tanto de la oferta como de la demanda.

La aceptación de la regulación voluntaria por parte de la Comisión requiere de determinados requisitos previos, como la adhesión de un número suficiente de empresas que garanticen su representatividad en el mercado de referencia; un código de conducta aplicable a las mismas; un órgano independiente encargado de su aplicación, y medidas para garantizar el cumplimiento de las decisiones de ese órgano, que pasan por el sometimiento a las entidades de resolución extrajudicial de reclamaciones y controversias reconocidas y acreditadas.

Solo de este modo la regulación voluntaria puede garantizar la defensa de la competencia y la protección de los consumidores a las que antes nos referíamos, a través de la simplificación regulatoria y la experiencia de los participantes a la hora de abordar casuísticas complejas; la flexibilidad y rapidez en la aplicación normativa; el bajo coste y fácil acceso de los procedimientos; la descongestión de los circuitos legislativos, así como la corresponsabilidad de los participantes.

¹ Sin ánimo de ser exhaustivo, pueden mencionarse iniciativas como la *Resolución del Consejo de Europa de 18 de febrero de 1972, sobre protección de los consumidores*; la *Recomendación (98/560/CE) del Consejo de la Unión Europea relativa a los principios*, o la *Comunicación de la Comisión Europea, de 9 de marzo de 1999, sobre los resultados de la consulta pública relativa al Libro Verde sobre la convergencia. aplicables a los órganos responsables de la solución extrajudicial de litigios en materia de consumo*. Directivas como las relativas a *publicidad engañosa* (84); *tratamiento de los datos personales* (95); *comercio electrónico* (2003); *prácticas comerciales desleales* (2005); *servicios de comunicación audiovisual* (modificada en 2018). Reglamentos como el *Reglamento general de protección de datos* (2016), *Reglamento de servicios digitales* (2022) y *Reglamento de mercados digitales* (2022)

A lo largo de estos años, la autorregulación ha demostrado ser una herramienta eficaz en diferentes sectores, destacando particularmente en el de la publicidad. Es el caso de la European Advertising Standards Alliance (EASA) a nivel europeo, y de su correspondiente española AUTOCONTROL, que facilitan y refuerzan el efectivo cumplimiento de las exigencias legales y éticas aplicables en este ámbito con herramientas como la consulta previa (*Copy Advice*®) y la actuación del Jurado.

En materia de desinformación, la Comisión Europea considera que la regulación voluntaria es una vía útil para abordar el problema, defendiendo una acción conjunta y colaborativa entre las plataformas y redes y los agentes involucrados, incluyendo las autoridades competentes, sin menoscabo de la regulación legal, como se pone de relieve en los apartados siguientes.

EL CÓDIGO DE BUENAS PRÁCTICAS EN MATERIA DE DESINFORMACIÓN

El Código en su versión inicial de 2018

El Código de buenas prácticas en la Unión en materia de desinformación² puede considerarse como la plasmación de los sucesivos planteamientos desarrollados por la Comisión Europea en relación a este problema. Sería prolijo mencionar aquí todos ellos, pero pueden destacarse aquí los recogidos en el *Informe del Grupo de Alto Nivel sobre un enfoque multidimensional de la desinformación* y en la Comunicación *La lucha contra la desinformación en línea: un enfoque europeo*, ambos de 2018³.

Aprobado en octubre de ese mismo año, el Código es suscrito por las plataformas de internet Facebook, Google, Twitter y Mozilla, así como por anunciantes y otros actores de la industria publicitaria. Microsoft se unió en mayo de 2019, y TikTok firmó el Código en junio de 2020.

Ámbitos de actuación

El Código de buenas prácticas se plantea actuar contra la desinformación en diferentes ámbitos:

Publicidad

Con el objetivo de reducir los ingresos de los proveedores de desinformación en línea, se plantea el compromiso de desarrollar políticas y procesos para impedir los ingresos de las cuentas y sitios web con ese tipo de contenidos, atacando a su capacidad de monetización. Ello implica no colocar publicidad en dichas páginas web ni promocionarlas.

² Comisión Europea (2018). Código de Buenas Prácticas en materia de Desinformación. Disponible en: <https://digital-strategy.ec.europa.eu/es/library/2018-code-practice-disinformation>

³ Comisión Europea, Dirección General de Redes de Comunicación, Contenido y Tecnologías, (2018). *A multi-dimensional approach to disinformation: report of the independent High level Group on fake news and online disinformation*, Publications Office. <https://data.europa.eu/doi/10.2759/739290> y Comisión Europea (2018). *Informe sobre la aplicación de la Comunicación «La lucha contra la desinformación en línea: Un enfoque europeo»*. eur-lex.europa.eu/legal-content/ES/TXT/PDF/?uri=CELEX:52018DC0794

Estas políticas y procesos deberían llevarse a cabo en colaboración con organizaciones de verificación de datos a la hora de identificar páginas y contenidos con desinformación. Y requieren asimismo de herramientas que permitan a los anunciantes valorar las estrategias de compra en medios y los riesgos para su reputación, ofreciéndoles el acceso necesario a cuentas específicas de clientes para ayudarles a supervisar la colocación de anuncios publicitarios y tomar decisiones respecto a su ubicación.

Usuarios

El empoderamiento o capacitación de los usuarios es otro de los objetivos básicos del Código, proporcionando a estas herramientas que permitan la identificación y notificación de los casos de desinformación, facilitando por otra parte el acceso a fuentes de información diferentes y plurales, y a puntos de vista alternativos en temas de interés público.

Es el caso, por ejemplo, del establecimiento de reglas claras para identificar el uso indebido de ordenadores zombis y garantizar que sus actividades no puedan confundirse con las interacciones humanas. O de políticas sobre qué constituye un abuso no permisible de sistemas automatizados, publicándolas en sus plataformas de forma que los usuarios de la UE puedan acceder a ellas. O del compromiso de no permitir el uso de cuentas y servicios de forma anónima o mediante seudónimos

Al mismo tiempo, los firmantes se comprometen a diluir la visibilidad de la desinformación mediante la mejora de la capacidad de los usuarios para encontrar contenido fiable mediante herramientas que les permitan una experiencia en línea personalizada e interactiva.

Ello significa invertir en medios tecnológicos para:

- Ayudar a las personas a tomar decisiones con conocimiento de causa cuando encuentran noticias en línea que pueden ser falsas, promoviendo iniciativas para desarrollar y aplicar indicadores de fiabilidad eficaces en colaboración con el ecosistema informativo y las asociaciones de medios.
- Dar prioridad a información pertinente, auténtica, diversa y autorizada en canales de búsqueda automatizada de información, con indicadores sobre la fiabilidad de las fuentes, información sobre la identidad de los propietarios de los medios y transparencia respecto a los motivos por los cuales los usuarios son receptores del contenido difundido.
- Promover, junto con la sociedad civil, los gobiernos, las instituciones educativas y otras partes interesadas, iniciativas de mejora del pensamiento crítico y la alfabetización en el ámbito de los medios digitales.

Verificación de datos

Los firmantes del Código reconocen la importancia de los indicadores de fiabilidad e información basados en criterios objetivos e independientes, como los proporcionados por la red de verificadores de datos promovida por la Comisión Europea, a la hora de aportar datos adicionales sobre proveedores de desinformación.

En este sentido, se comprometen a colaborar con los verificadores de datos y sus organizaciones en el desarrollo de las políticas y procesos ya mencionados para interrumpir la publicidad y los incentivos a la monetización de los contenidos relacionados con la desinformación. Ello implica el uso de indicadores de fiabilidad de las fuentes basados en criterios objetivos e independientes.

Comunidad investigadora

Los firmantes del Código reconocen también la importancia de fomentar la investigación sobre la desinformación (por ejemplo, a través de la propaganda política), comprometiéndose a facilitar datos pertinentes sobre el funcionamiento de sus servicios para la realización de estudios independientes por parte de expertos académicos y de las organizaciones de la sociedad civil. Esto incluye información general sobre el uso de algoritmos y el intercambio de conjuntos de datos, siempre protegiendo la intimidad de las personas.

Los firmantes se comprometían, asimismo:

- A adoptar medidas de seguridad contra las declaraciones falsas antes de comercializar nuevos servicios, y revisar en lo posible los servicios existentes para garantizar que también se aplican dichas medidas de seguridad.
- A convocar un acto anual para fomentar los debates entre el mundo académico, la comunidad de verificación de datos y los diferentes agentes de la cadena de valor.

Medición y seguimiento de la eficacia del código

De acuerdo con lo señalado, las medidas a adoptar para mejorar la transparencia en materia de desinformación se referían a:

- Políticas y actividades de cumplimiento destinadas a reducir las oportunidades de monetización para los proveedores de desinformación.
- Capacitación de los consumidores con productos, herramientas, tecnologías y programas para identificar la desinformación y la publicidad política; concienciación de los consumidores respecto a la desinformación.

- Aportación de herramientas a los usuarios para encontrar diferentes puntos de vista sobre temas de interés público.
- Políticas relacionadas con la integridad de sus servicios en el contexto de la desinformación.
- Capacitación de los investigadores y los grupos de la sociedad civil para supervisar el alcance y magnitud de la publicidad política.
- Fomento de la educación de la población para mejorar el pensamiento crítico, sus conocimientos sobre los medios digitales y sus capacidades (alfabetización).
- Apoyo a la red de verificadores de datos reconocida por la UE.

El Código solo era de aplicación para los firmantes, que además podían retirarse del mismo, o de determinados compromisos contemplados en el mismo, en cualquier momento, mediante notificación a la Comisión Europea y a los demás signatarios. Debe tenerse en cuenta que, por la naturaleza heterogénea de los firmantes, estos asumían ya desde el inicio únicamente los compromisos relacionados con su oferta de bienes y servicios y con su posición en la cadena de valor, así como con su capacidad técnica, su realidad tecnológica, su tipología de usuarios y sus regímenes de responsabilidad en el marco de la UE y del Espacio Económico Europeo.

Una vez vigente, cualquier modificación del Código debería ser acordada por todos los firmantes. Cada firmante podía informar a los demás en cualquier momento si consideraba que otro firmante estaba incumpliendo los compromisos asumidos, e indicar los motivos para tal sospecha. En tal caso, los firmantes podían acordar la celebración de una reunión plenaria para escuchar las alegaciones del signatario afectado y concluir, en su caso, invitar al infractor a retirarse del Código, informando a la Comisión Europea.

Los firmantes podían hacer pública en sus sitios web o en comunicaciones comerciales o de otro tipo, su adhesión al Código, así como adoptar todas las medidas oportunas para informar a sus contactos comerciales.

Las asociaciones profesionales firmantes del Código no contraían obligación alguna en nombre de sus miembros, pero sí se comprometían a promover su conocimiento, y a animar a otros para su adhesión y respeto a sus principios. El Código se refiere de modo específico a la Federación Mundial de Anunciantes, la Asociación Europea de Agencias de Comunicación (EACA) e IAB Europe, que han de ofrecer informes agregados para seguir e identificar las distintas actividades y políticas de seguridad de marca.

En este Código de buenas prácticas inicial se incluía una referencia expresa de reconocimiento al trabajo legislativo, entonces en curso, para desarrollar estándares de transparencia respecto a los principales parámetros de clasificación incluidos en el proyecto de Reglamento sobre las relaciones entre las plataformas y las empresas, así como al trabajo del Grupo de Expertos en inteligencia artificial de la UE y al acervo en materia de consumo de la UE.

Periodo de evaluación y continuidad

Como hemos indicado, el Código establecía un período de evaluación de doce meses, durante los cuales debían realizarse reuniones periódicas para analizar su funcionamiento, el progreso en su aplicación y la eficacia del mismo en relación con cada uno de los compromisos establecidos anteriormente.

Transcurrido ese plazo, los firmantes se comprometían a seleccionar una organización independiente para revisar de forma objetiva los informes anuales de autoevaluación correspondientes y evaluar el nivel de progreso realizado en relación con los compromisos adquiridos (rendición de cuentas).

A partir de este balance, podrán proponerse nuevas medidas de seguimiento y posibles cambios en los compromisos a adoptar.

Las plataformas y entidades firmantes del Código de buenas prácticas de 2018 presentaron en enero de 2019 un informe en el que se exponía la situación de las medidas adoptadas.⁴ Por su parte, la Comisión Europea llevó a cabo entre enero y mayo de 2019 un seguimiento específico de la aplicación de los compromisos de tres de los firmantes: Facebook, Google y Twitter, con especial relevancia en la repercusión de su actividad en las elecciones al Parlamento Europeo.

En octubre de 2019 se publicó el informe de autoevaluación de los firmantes, recogiendo sus iniciativas para cumplir con los compromisos establecidos en el Código,⁵ y la Comisión dio a conocer su informe de evaluación en septiembre de 2020.⁶

El informe de la Comisión puso de manifiesto los avances que el Código había proporcionado en materia de desinformación, facilitando un diálogo con las plataformas en línea; garantizando una mayor transparencia de sus políticas, y propiciando acciones concretas y cambios políticos por parte de las partes interesadas pertinentes para ayudar a contrarrestar la desinformación. Sin embargo, también puso de manifiesto una serie de lagunas y deficiencias importantes, como la falta de definiciones comúnmente compartidas, de claridad en los procedimientos, de precisión y completitud en los compromisos, así como de indicadores clave de rendimiento (KPI) significativos y transparentes.

⁴ Pueden consultarse en: <https://digital-strategy.ec.europa.eu/en/news/first-results-eu-code-practice-against-disinformation>

⁵ Comisión Europea (2019). *Annual self-assessment reports of signatories to the Code of Practice on Disinformation 2019*. Disponible en: <https://digital-strategy.ec.europa.eu/en/news/annual-self-assessment-reports-signatories-code-practice-disinformation-2019>

⁶ Comisión Europea (2020). *Assessment of the Code of Practice on Disinformation – Achievements and areas for further improvement*. Disponible en: <https://digital-strategy.ec.europa.eu/en/library/assessment-code-practice-disinformation-achievements-and-areas-further-improvement>

La Comisión hacía especial hincapié, además, en la falta de acceso a datos que permitieran una evaluación independiente de las tendencias emergentes y las amenazas que plantea la desinformación en línea. Ello hacía difícil evaluar con precisión la puntualidad y el impacto de las acciones de los firmantes, ya que las autoridades públicas seguían dependiendo en gran medida de la voluntad de las plataformas para compartir información y datos. Se solicitaba, en ese sentido, un modelo más estructurado de cooperación entre las plataformas y la comunidad investigadora.

Finalmente, pedía establecer un procedimiento adecuado de seguimiento y evaluación, que permitiera una mejor garantía de rendición de cuentas, así como una extensión del Código a otras partes interesadas pertinentes, en particular del sector de la publicidad.

El Código de buenas prácticas en materia de desinformación reforzado (2022)

Las objeciones de la Comisión Europea ante el Código de buenas prácticas en materia de desinformación de 2018, a las que nos hemos referido, quedaron plasmadas en la *Guía de Orientaciones para reforzar el Código de buenas prácticas en materia de desinformación* de mayo de 2021.⁷ Pero es necesario mencionar, como paso previo, su Comunicación *Plan de Acción para la Democracia Europea*, aprobado en diciembre de 2020.⁸

Básicamente, en el Plan se destaca que el rápido crecimiento de las plataformas en línea ha evidenciado nuevas vulnerabilidades y ha hecho más difícil mantener la integridad de las elecciones; garantizar unos medios libres y plurales, esenciales para el proceso democrático, y proteger a la ciudadanía frente a la desinformación y otros tipos de manipulación.

En línea con otros documentos ya mencionados, en esta Comunicación se define la desinformación como “*un contenido falso o engañoso que se difunde con intención de engañar o de obtener una ganancia económica y política y que puede causar un perjuicio público*”, diferenciándola de la información engañosa o falsa compartida sin intención de perjudicar, aunque sus efectos puedan ser nocivos. Y distinguiendo también entre las operaciones de influencia, fruto del esfuerzo coordinado de actores nacionales, y la injerencia extranjera, que a menudo forma parte de una operación híbrida más amplia.

⁷ Comisión Europea (2021). *Guía de Orientaciones para reforzar el Código de buenas prácticas en materia de desinformación*. Disponible en: <https://digital-strategy.ec.europa.eu/es/library/guidance-strengthening-code-practice-disinformation>

⁸ Comisión Europea (2020). *Plan de Acción para la Democracia Europea*. Disponible en: <https://eur-lex.europa.eu/legal-content/ES/TXT/PDF/?uri=CELEX:52020DC0790&from=EN>

De cara al futuro, la Comunicación apostaba por revisar el Código de buenas prácticas de acuerdo con lo señalado por la propia CE, pero también por su integración en la (entonces) próxima regulación de los servicios digitales, como marco horizontal que vendría a garantizar una mayor responsabilidad de las plataformas a la hora de rendir cuentas sobre la forma en la que moderan sus contenidos y aplican sus algoritmos.

Por su parte, la Guía orientaba sobre cómo podría reforzarse el Código de buenas prácticas en diferentes ámbitos, para convertirse en una herramienta más eficaz para contrarrestar la desinformación garantizando una aplicación completa y coherente entre las partes interesadas y los países de la UE:

- Mayor participación en los compromisos adoptados.
- Mejor desmonetización de la desinformación.
- Garantizar la integridad de los servicios.
- Mejorar la capacitación de los usuarios.
- Aumentar la cobertura de la verificación de datos.
- Proporcionar a los investigadores un mayor acceso a los datos.
- Crear un marco de seguimiento más sólido.

También reclamaba un centro en el que pudiera recogerse información sobre las políticas adoptadas para aplicar los compromisos del Código y su grado de cumplimiento.

En ese contexto, los firmantes del Código y los posibles nuevos firmantes se reunieron el 8 de julio de 2021, poniendo en marcha una convocatoria de manifestaciones de interés⁹ dirigida a una amplia gama de partes interesadas, invitando a éstas a convertirse en firmantes y a participar en la preparación del Código reforzado. La convocatoria incluía a los proveedores de servicios en línea que participan en la difusión de contenidos al público, como redes sociales, a los servicios de búsqueda, a las aplicaciones de mensajería privada, al sector de la publicidad y a otros agentes interesados en contribuir a la lucha contra la difusión de desinformación a través del desarrollo de herramientas, trabajo filantrópico o conocimientos especializados específicos.

⁹ Más información en: <https://digital-strategy.ec.europa.eu/es/library/joint-call-interest-join-code-practice-disinformation>

Para ello aprobaron un Vademécum sobre la organización y el funcionamiento del proceso de configurar y redactar un Código de Desinformación reforzado para finales de 2021.¹⁰

La Asamblea de firmantes se impuso como objetivo revisar el Código de 2018 en consonancia con la Guía para reforzar el Código de buenas prácticas, estableciendo compromisos adaptados a la diversidad de servicios ofrecidos por los firmantes, adaptados a sus respectivas funciones en el ecosistema en línea, así como un marco sólido para el seguimiento y la supervisión periódicos del Código, definiendo indicadores clave de rendimiento capaces de medir la aplicación y la eficacia del Código.

Se definió un calendario y los aspectos prácticos de trabajo del proceso de redacción, que tendría lugar a partir de grupos temáticos:

- Integridad de los servicios y capacitación de los usuarios, incluidos los indicadores clave de rendimiento pertinentes.
- Capacitar a la comunidad de investigación y verificación de datos, incluidos los indicadores clave de rendimiento pertinentes.
- El control de las colocaciones de anuncios, la publicidad política y la publicidad temática, incluidos los indicadores clave de rendimiento pertinentes.
- Seguimiento del Código.

De este modo, los firmantes elaboraron el Código de buenas prácticas en materia de desinformación reforzado, suscrito en junio de 2022.¹¹

Principales aspectos del Código

El Código de buenas prácticas en materia de desinformación reúne un elenco de firmantes más extenso que el de 2018 (Ver Anexo I), entre ellos, las grandes plataformas, buscadores y redes sociales: Google, Meta (Facebook, Instagram, LinkedIn), Microsoft, TikTok, Twitch y Twitter. Queda a la discreción de dichos firmantes la decisión sobre qué aspectos de los recogidos en el Código asumen y se comprometen a cumplir.

¹⁰ Disponible en : https://ec.europa.eu/newsroom/repository/document/2021-27/Vademecum_2021_Code_poQcw5VoJg362zUKq6VL774dsFs_78161.pdf

¹¹ Comisión Europea (2022). *Código de buenas prácticas en materia de desinformación reforzado*. Disponible en: <https://digital-strategy.ec.europa.eu/en/library/2022-strengthened-code-practice-disinformation>

El Código de buenas prácticas reforzado contiene 44 compromisos específicos, referidos a las siguientes áreas:

Desmonetización

Se trata de evitar que la difusión de contenidos considerados desinformación se beneficien de los ingresos económicos derivados de la inserción de comunicaciones comerciales en o junto a dichos contenidos. Ello requiere de colaboración entre el sector publicitario y las plataformas y redes sociales.

Transparencia de la propaganda política

Implementar medidas más estrictas y de etiquetado que permitan a los usuarios reconocer fácilmente los anuncios de naturaleza política. También mayor transparencia sobre estos anuncios (anunciante responsable y quien lo realiza, gasto, período de visualización), lo que requiere de bibliotecas de anuncios (*ad libraries*) eficientes y con capacidad de búsqueda.

Medidas contra la manipulación

Actuar contra los comportamientos y prácticas empleadas para difundir desinformación como cuentas falsas; *bots*, para amplificar su diseminación de los mensajes; suplantación de identidad, falsificaciones profundas maliciosas, etc. Ello requiere de cooperación para conocer y contar con información actualizada sobre las tácticas, técnicas y procedimientos (TTP) empleados por los responsables de la desinformación e implementar políticas claras al respecto.

Empoderamiento de los verificadores de datos

Realizar un mejor uso, más coherente y eficaz, de la verificación de datos por parte de plataformas y redes sociales, garantizando recursos financieros justos para los verificadores, así como un mayor acceso a la información que facilite su trabajo diario. Ello requiere ampliar la cobertura de la verificación de datos en todos los Estados miembros e idiomas de la UE.

Empoderamiento de los investigadores

Brindar un mayor apoyo a la investigación sobre la desinformación, mejorando y ampliando la utilización por parte de los investigadores de los contenidos difundidos y almacenados por las plataformas en línea y las redes sociales. Ello requiere garantizar el acceso automatizado a datos no personales, anónimos, agregados o públicamente manifiestos, así como simplificar el acceso a aquellos que requieren un escrutinio adicional.

Empoderamiento de los usuarios

Proporcionar a las personas usuarias finales no solo una mayor protección ante la desinformación desde las propias plataformas, sino también posibilitar su participación activa en la lucha contra ésta, mejorando las herramientas existentes para que la ciudadanía pueda identificarla, contrastarla, señalarla y procurar su eliminación.

Ello requiere potenciar la alfabetización mediática en todos los segmentos de edad, estrategias de prevención “desde el diseño” y una mayor transparencia de los sistemas de recomendación, primando las fuentes fiables.

Centro de Transparencia

El Código prevé la creación de un Centro de Transparencia, accesible a todos los ciudadanos, que permita conocer las medidas adoptadas por los firmantes para la implementación de las medidas previstas, con actualizaciones periódicas de los datos relevantes.

Grupo de Trabajo Permanente

También contempla la labor de un Grupo de Trabajo permanente, dedicado al seguimiento del Código y a su evolución futura en el marco de los objetivos perseguidos, con el establecimiento de un foro que revise precisamente la vigencia de los compromisos contraídos y su posible actualización a la vista de los desarrollos tecnológicos, sociales, de mercado y legislativos.

Este Grupo de Trabajo está compuesto por representantes de los firmantes, del Grupo de Reguladores Europeos de Servicios de Medios Audiovisuales (ERGA), del Observatorio Europeo de Medios Digitales (EDMO) y el Servicio Europeo de Acción Exterior, presidido por la Comisión.

Rendición de cuentas del cumplimiento del Código reforzado

Los firmantes dispusieron de un periodo inicial de seis meses para llevar a cabo la implementación del Código. En febrero de 2023, la gran mayoría de ellos proporcionaron a la Comisión los primeros informes de referencia a tal efecto, almacenados en el repositorio del Centro de Transparencia.¹²

Para la elaboración de los informes, los firmantes utilizaron un modelo armonizado, que permite realizar un balance particularizado de las medidas adoptadas en función de los compromisos contraídos.

Cabe señalar que la calidad y el nivel de detalle de dichos informes es muy desigual. Destacan, en este sentido, las críticas de los representantes de la Comisión al informe presentado por Twitter, debido a su brevedad y a la falta de información relevante sobre las medidas adoptadas en aspectos importantes como el empoderamiento o capacitación de los verificadores. Precisamente esta red social, en mayo de 2023, tras la adquisición de Twitter por su nuevo propietario, decidió retirarse del Código.

Más allá de las reflexiones generales sobre el fenómeno de la desinformación, la puesta en valor global de la actividad realizada y su filosofía de actuación o la explicación sumaria de la actividad de cada una de las entidades firmantes, la lectura de los resúmenes ejecutivos de cada informe permite seleccionar algunos aspectos básicos de los mismos.

Centrándonos en los principales prestadores de servicios de la sociedad de la información, tales aspectos básicos son los siguientes:

Google

Este prestador considera que el mantenimiento de un enfoque responsable para apoyar la calidad de la información debe ser compatible con garantizar a los usuarios el acceso a un ecosistema de información abierto. Respetando la elección por parte del usuario de contenidos que no sean ilegales o estén prohibidos por las políticas de Google; atendiendo a la diversidad de esos usuarios desde el punto de vista cultural, lingüístico y de origen, y tratando de atender sus necesidades adecuadamente.

Defiende la aplicación de una visión polifacética, desde diferentes enfoques, para abordar los complejos retos y riesgos que plantean la desinformación y la información errónea en todos sus productos y servicios: buscador de contenidos (Google Search), la plataforma de YouTube y gestión de contenidos publicitarios (Google Ads).

¹² Pueden consultarse en: <https://disinfocode.eu/es/informes/?a%C3%B1os=2023-es>

Así, en el caso de los contenidos publicitarios destaca sus acciones de desmonetización, impidiendo la colocación de creatividades publicitarias en páginas y dominios con contenidos de desinformación o que violan sus políticas. O eliminando contenidos publicitarios ilícitos.

Se refiere también a otras acciones desarrolladas:

- Conectar a los usuarios con contenidos fidedignos, por ejemplo, mediante la publicación de información de contexto (paneles informativos) en todo el ámbito del EEE.
- Campañas de alfabetización mediática y uso racional de las TIC. Por ejemplo, la opción Pulsar Pausa en Youtube, que se aplica en 20 Estados miembros del EEE.
- Campaña de Google Search Superbuscadores para bibliotecarios y personal bibliotecario.
- Apoyo a los investigadores, con una inversión inaugural de 25 millones de euros destinados al Fondo Europeo para los Medios de Comunicación. 13,2 millones de dólares a la Red Internacional de Verificación de Hechos. Y a través del Programa de Investigadores de YouTube.

A futuro, Google se compromete a una colaboración más estrecha con otros Signatarios del Código y con el Observatorio Europeo de Medios Digitales (EDMO).

Meta

Meta se refiere también a la necesidad de encontrar el equilibrio adecuado entre la protección de las personas y la defensa de la libertad de expresión. Menciona:

- La utilización de la investigación, de los equipos de expertos y de la tecnología para combatir la difusión de contenidos nocivos, incluidas la desinformación y la información errónea.
- La colaboración con el mundo académico, con la sociedad civil y con verificadores terceros.
- La eliminación de más de 2,9 millones de anuncios de Facebook e Instagram en los Estados miembros de la UE. Más de 8.800 por infringir su política de desinformación en el último trimestre de 2022.

En el caso de anuncios políticos, obliga a estos a pasar por un proceso de autorización y a incluir una cláusula de exención e identificación de responsabilidad (“pagado por”). Señala que, del 15 de noviembre al 31 de diciembre de 2022, se etiquetaron más de 170 000 anuncios de este tipo en Facebook y en Instagram.

Se refiere también a la aplicación de las políticas y prácticas contra el Comportamiento Inauténtico Coordinado (CIB) y contra las operaciones de influencia encubiertas, basadas en ese comportamiento y en los actores, más que en el contenido en sí mismo. Con el resultado de la eliminación de tres redes en el tercer trimestre de 2022 y una red en el cuarto trimestre de 2022.

Menciona las medidas proactivas para eliminar cuentas falsas, especialmente las que buscan causar daño: en el tercer trimestre de 2022. Esas medidas afectaron a 1.500 millones de cuentas en Facebook, aproximadamente el 5% de sus usuarios activos mensuales en todo el mundo.

La cooperación con verificadores independientes se concreta en dicha cooperación con 26 organizaciones que cubren 22 idiomas diferentes de la UE. En el tercer trimestre de 2022, difundieron más de 150.000 artículos diferentes de verificación a nivel mundial para etiquetar y reducir la viralidad en Meta de más de 28 millones de contenidos inadecuados en la UE. Y más de 40.000 artículos para reducir la viralidad de más de 1,7 millones de contenidos en Instagram. Señalan desde Meta que, cuando se coloca una pantalla de advertencia de en un post, el 95% de las veces la gente no hace clic para verlo. Y que el 38 % de los usuarios de Instagram y el 25 % de los de Facebook que en la UE que empiezan a compartir contenidos, no completan esta acción tras recibir un aviso de que el contenido no ha sido contrastado.

En el campo de las campañas de alfabetización mediática se refiere el informe a las desarrolladas en Polonia, Eslovaquia, Lituania, Letonia, Estonia, Albania, Bosnia y Herzegovina, Kosovo, Serbia y Bulgaria, diseñadas en colaboración con verificadores de hechos y ONG locales, que llegaron a 16 millones de usuarios con 72 millones de impresiones.

La capacitación de investigadores pasa por el desarrollo de métodos para permitir recopilar y analizar datos, aunque consideran necesario aclarar que sin comprometer la seguridad de la plataforma ni la privacidad de las personas usuarias. Incluyen también la inversión en numerosos proyectos de investigación, y el acceso de más de mil cuentas académicas (que representan más de cinco mil usuarios individuales) a CrowdTangle, en todo el mundo. a partir de enero de 2023.

Microsoft

Este prestador no apuesta por el bloqueo de contenidos falsos en los resultados de búsqueda, por considerar que puede plantear problemas de libertad de expresión y de recepción y difusión de información. Opta por clasificar sus resultados de modo que aparezcan en primer lugar los contenidos fiables y autorizados, proporcionando herramientas a los usuarios para ayudarles a evaluar esa fiabilidad.

Sus Principios de Integridad de la Información contemplan: 1) El respeto a la libertad de los clientes para crear, publicar y buscar información. 2) La publicación prioritaria de contenidos internos y de terceros de confianza. 3) No proporcionar “voluntariamente” beneficios económicos a contenidos o actores vinculados con la ciberinfluencia exterior. 4) Evitar de modo proactivo el uso de sus plataformas y productos se utilicen para amplificar sitios y contenidos de esta ciberinfluencia.

Se refiere también al desarrollo de un enfoque para mitigar el riesgo de sus operaciones de desinformación: Detectar (perseguir, rastrear e investigar colectivamente). Interrumpir (eliminación de operaciones y, corte de sus ingresos). Defender (tecnología responsable y orientada al consumidor, integridad de la información, fomentar el consumo crítico por parte de los usuarios). Disuadir (acogimiento a normas internacionales que creen un estándar, especialmente en temas de derechos humanos fundamentales, como la salud).

Cabe mencionar, asimismo:

- La adquisición en julio de 2022 de Miburo Solutions, una empresa de análisis e investigación de ciberamenazas especializada en la detección y respuesta a operaciones de influencia informativa extranjera.
- La proporción de datos relacionados con las búsquedas a la comunidad investigadora, y búsqueda activa de colaboración con asociaciones de investigación que estudian la desinformación.
- La prohibición expresa en LinkedIn a sus miembros de publicar información que sea intencionadamente engañosa o equívoca.

TikTok

Esta red social defiende la potenciación de la labor proactiva de moderadores especializados y formados en el ámbito de la desinformación, frente a la moderación automatizada. Ello se justifica por las peculiaridades del fenómeno de la desinformación frente a otro tipo de contenidos perjudiciales, cuya eliminación también se contempla, debido a la importancia en este caso del análisis del contexto y de la comprobación o verificación de los hechos.

TikTok estima que la moderación que lleva a cabo de los contenidos de la plataforma tiene como resultado la retirada del 1% de su oferta de vídeos, suponiendo la desinformación el 1% de esas retiradas en el EEE, en gran parte centrados en la pandemia del Covid-19 y la guerra en Ucrania.

Ha llevado a cabo el bloqueo o cierre de cuentas falsas o inauténticas; la detección y adopción de medidas coercitivas contra las redes de influencia encubiertas, y la eliminación de contenidos engañosos que puedan causar daños significativos.

Aunque, según afirma, debido a la relativa novedad de esta red social, las oportunidades de monetización de los creadores están en una fase relativamente temprana de madurez para los estándares de la industria, participa en la Alianza Global para Medios Responsables (GARM) y aplica sus políticas a los contenidos publicitarios.

Dice no permitir la publicidad de actores políticos, o debates sobre las actuaciones de cargos públicos o líderes políticos actuales o anteriores, partido u organizaciones o asunto local, estatal o federal con el fin de influir en un resultado electoral o de toma de decisiones. Sí permite la publicidad basada en causas y la publicidad de servicios públicos de organismos gubernamentales, organizaciones sin ánimo de lucro y otras entidades, si no están impulsadas por motivos políticos partidistas.

Otros aspectos recogidos en su informe:

- Desarrollo de herramientas de denuncia para capacitar a los usuarios integrados en la aplicación, accesibles y fáciles de usar y disponibles en 21 lenguas oficiales. Mostración a los usuarios de contexto adicional sobre determinados contenidos o redireccionamiento a información fidedigna.
- Participación en campañas de alfabetización mediática dentro y fuera de las plataformas.
- Eliminación de contenidos de desinformación perjudiciales en el sistema de recomendaciones personalizado (el feed “Para ti”). Introducción de la función “Por qué este vídeo”, para ayudar a los usuarios a comprender mejor por qué se les ha recomendado un vídeo concreto.
- Desarrollo de una API global e independiente para proporcionar a investigadores seleccionados acceso a diversos datos públicos y anonimizados. Desarrollo de una API específica para facilitar a investigadores autorizados el acceso a datos pertinentes sobre desinformación.
- Colaboración con ocho socios verificadores de hechos, que revisan contenidos en diez lenguas oficiales, acreditados por la International Fact-Checking Network. Los verificadores no actúan directamente sobre los contenidos, sino que transmiten sus aportaciones y comentarios a los equipos de moderación. En este campo se mencionan también la creación de una base de datos de afirmaciones verificadas y la reducción preventiva de la difusión de contenidos en fase de verificación. Se anuncian colaboraciones futuras con nuevas asociaciones de verificadores en Portugal, Dinamarca, Grecia y Bélgica, así como con la Red Europea.
- Colaboración con el Observatorio Europeo de Medios Digitales.

Twitch

Considera esta plataforma que las cifras de aplicación de la desinformación en su plataforma son relativamente bajas, debido a que la mecánica de Twitch hace muy difícil para un nuevo *streamer* conseguir grandes cantidades de espectadores simultáneos y también a que la mayoría de los contenidos no se almacenan y, por tanto, no se comparten ni se convierten en virales.

En su informe, solo se enfoca en la desinformación persistente o agregada, no en declaraciones singulares o de compartición breve.

Entre sus actuaciones, destaca la asociación con más de una docena de investigadores y expertos para comprender cómo se propaga la desinformación perjudicial en internet y garantizar que nuestro enfoque para mitigar sus riesgos en nuestra comunidad sea eficaz. Sus conclusiones apuntan a que una cantidad desproporcionada de contenidos se concentran en un número reducido de actores de desinformación perjudicial, debido a su persistencia y a la atención que despiertan los temas elegidos.

También menciona la inversión de importantes recursos para combatir la falta de autenticidad (*bots*, *spammers*, suplantadores de identidad) a través de los canales de denuncia y de su sistema de detección proactiva. Y el mantenimiento de la colaboración con el mundo académico, la sociedad civil y la industria.

Twitter

Centra sus actuaciones contra la desinformación en el desarrollo de *Community Notes* (CN), un modelo de moderación de contenidos basado en la participación de los usuarios y no en una aplicación centralizada. Permite añadir notas de réplica o rectificación a los tuits que puedan inducir a error. Cuando suficientes colaboradores de diferentes puntos de vista califican esa nota como útil, se muestra públicamente.

Afirma que, según cuatro encuestas realizadas en Estados Unidos, una persona que ve una nota comunitaria tiene de media entre un 20 % y un 40 % menos de probabilidades de aceptar el contenido de un tuit potencialmente engañoso y entre un 15 % y un 35 % menos de probabilidades de elegir Me gusta o retuitearlo.

Con el tiempo, se pretende que los usuarios (y no solo los colaboradores) de cualquier Estado miembro de la UE, que escriban en cualquier lengua, puedan contribuir a las CN. Considera que se trata de un modelo rápido y escalable.

Se refiere también a diferentes actuaciones previstas o en curso:

- Restablecimiento del Centro de Transparencia de Anuncios, en paralelo a la labor del equipo de Desactivación de Amenazas para llevar a cabo políticas claras que prohíben la publicidad manipuladora o el spam.
- Posibilidad de alternar intuitivamente entre el algoritmo que sugiere contenidos y un feed cronológico inverso. Se prevé abrir el algoritmo que recomienda contenidos en la línea de tiempo.
- Lanzamiento del servicio de suscripción Twitter Blue, diseñado en parte para autenticar las identidades de los usuarios y reducir así la prevalencia del spam y la desinformación viral.
- Mantenimiento de la colaboración con organizaciones de la sociedad civil.
- Según sus datos, el discurso de odio representa menos del 0,1% de todas las impresiones de tuits en inglés, aunque en algunas zonas no puede proporcionar datos granulares debido a la escasez de recursos y a las limitaciones de los datos.

Se considera uno de los actores más abiertos del sector de las plataformas en lo que respecta a compartir datos para la investigación académica, poniendo a disposición grandes conjuntos de datos y el programa API de Twitter.

Señala que hay aspectos del Código que no son aplicables al servicio de Twitter y anuncia que está estudiando la mejor manera de proporcionar detalles sobre el cumplimiento del Reglamento de servicios digitales por parte de Twitter.

Como ya hemos indicado, mayo de 2023, tras la adquisición de Twitter por su nuevo propietario, la red social Twitter decidió retirarse del Código.

Las decisiones adoptadas a partir de esa adquisición pasaban por decisiones como “suavizar” la moderación en el servicio de mensajes cortos, la eliminación de las advertencias a sus usuarios sobre el contenido potencialmente falso o la conversión de la oferta del *check* azul de usuario verificado a servicio de pago.

No obstante, Twitter (ahora denominada X) tendrá que cumplir a partir del 25 de agosto con esos y otros requisitos tras la entrada en vigor de las medidas contempladas para las grandes plataformas en el Reglamento de servicios digitales.

Referencias específicas a España en los informes de rendición de cuentas de las plataformas firmantes del Código

No todos los informes de rendición de cuentas de los suscriptores del Código ofrecen datos específicos sobre los Estados miembros, y, en concreto, sobre España. En el caso de las principales plataformas, buscadores y redes, únicamente Twitch no presenta datos desglosados para nuestro país.

Centrándonos en las grandes plataformas de internet, las referencias a sus acciones específicas en España contra la desinformación son las siguientes:

TikTok

En relación con España, TikTok se refiere:

- A los contenidos eliminados por desinformación, tanto en número de páginas como en número de dominios referidos a la COVID-19, a la propaganda política y a otros contenidos falsos o engañosos.
- A las cuentas en remoto y a los seguidores o suscriptores falsos, así como a las suplantaciones de identidad sobre las que se ha actuado para su impedimento remoción o eliminación.
- A las reacciones (“me gusta”, votos positivos, comentarios) identificadas como falsas o no y a la filtración de hashtags.
- A la identificación de publicidad encubierta (mensajes compensados o promociones no transparentes) por parte de los influencers.
- Se refiere también a las acciones de verificación de contenidos y a la inclusión de etiquetas de aviso ante contenidos no verificados.

Meta

Se refiere a acciones llevadas a cabo en relación a Facebook, Instagram o Whatsapp en aspectos como:

- La eliminación de contenidos por violación de las políticas contra la desinformación (salud /interferencias electorales).
- Los anuncios rechazados por incumplimiento de las políticas del proveedor, y el etiquetado de anuncios con descargos de responsabilidad.

- Las acciones de alfabetización mediática e informacional, como las dirigidas a jóvenes en las que se enseña a utilizar las herramientas de seguridad de la red social.
- El etiquetado de contenidos por parte de verificadores independientes y colaboración con ellos (incluyendo aportaciones económicas) en el marco del programa Spread the Facts y con quienes utilizan productos de WhatsApp.

Google

Se refiere en su informe, con datos del tercer trimestre de 2022:

- A las páginas y dominios de AdSense activados por contenidos no confiables y perjudiciales, manipulación, contenidos peligrosos o despectivos, u otros en el tercer trimestre de 2022.
- Al impedimento y bloqueo de solicitudes del flujo de inversiones publicitarias lícitas a sitios o contenidos designados como desinformación en páginas y dominios.
- A las acciones frente a incumplimiento de políticas de contenido en relación con requisitos de destino (contenido original insuficiente), contenido inapropiado (peligroso o despectivo, impactante, eventos sensibles, crueldad animal) o tergiversación (prácticas comerciales desleales, prácticas engañosas coordinadas, representaciones engañosas, medios manipulados, afirmaciones poco confiables, diseño de anuncios engañosos, clickbait, relevancia poco clara, ofertas no disponibles. Exitosas, parcialmente exitosas o fallidas.
- A las creatividades de propaganda política no verificadas y rechazadas.
- A las advertencias de contenido sobre calidad/relevancia/fiabilidad no garantizada, cuantificando su visionado (“más información sobre esta página”, “acerca de este resultado”).
- A la campaña “Hit Pause” de alfabetización mediática.
- A los resultados de las acciones de eliminación de vídeos de YouTube por infracciones de las políticas de desinformación.
- A las investigaciones sobre contenidos en YouTube por parte de usuarios (Fact-Check Explorer, Google Trends de Google Searches y YouTube).

Microsoft

Se refiere a España en relación con:

- Anuncios restringidos en LinkedIn por información errónea en el marco de las políticas contra la desinformación (páginas / dominios).
- Número de cuentas falsas (botnets) que LinkedIn restringió o impidió que se crearan.
- Reacciones falsas o no auténticas (por ejemplo, me gusta, votos positivos, comentarios) reportadas en LinkedIn.
- Cuentas falsas reportadas de supuestos seguidores o suscriptores.
- Creación de páginas, grupos, grupos de chat, o dominios no auténticos.
- Acciones de transparencia (Bing).
- Apelaciones por parte de creadores de contenidos eliminados y su resolución.
- Miembros que optan en LinkedIn por ordenar su feed en base a criterios distintos de la relevancia.
- Uso de verificación de datos en Bing, Contenidos revisados por verificadores externos.
- Miembros que utilizaron la página “Acerca de función de este perfil”.
- Visitas al centro de información sobre COVID y Guerra de Ucrania.

REGLAMENTO DE SERVICIOS DIGITALES

Marco general

El Reglamento de servicios digitales (RSD, DSA, por sus siglas en inglés)¹³ forma parte de la nueva estrategia digital de la Unión Europea, junto con el Reglamento de mercados digitales (RMD, DMA),¹⁴ y viene a reforzar el marco normativo establecido por la Directiva de Comercio Electrónico,¹⁵ que había sido incorporada a nuestro ordenamiento jurídico por la Ley 34/2002, de 11 de julio, de servicios de la Sociedad de la Información y de Comercio Electrónico (LSSI).¹⁶

Como reglamento, es de aplicación obligatoria y directa en todos sus elementos en cada Estado miembro, sin necesidad de su trasposición a los ordenamientos jurídicos nacionales (como sí ocurre con las directivas).

Del mismo modo que el RMD se orienta más a cuestiones competenciales, el RSD tiene como objetivo regular la prestación de los servicios de la sociedad de la información para procurar un nivel elevado de protección de los usuarios. Se pretende en este sentido que los prestadores adopten medidas frente a los contenidos ilícitos (entre ellos, la desinformación), con el fin de garantizar un entorno en línea seguro, predecible y digno de confianza, que permita el pleno ejercicio de la Carta de los Derechos Fundamentales de la Unión Europea, en aspectos como la libertad de expresión y de información o el derecho a la no discriminación.

El Reglamento entró en vigor el 16 de noviembre de 2022, aunque se contempla su aplicación en diferentes fases que se extienden hasta 2024.

¹³ REGLAMENTO (UE) 2022/2065, del Parlamento Europeo y del Consejo, de 19 de octubre de 2022, relativo a un mercado único de servicios digitales. <https://www.boe.es/buscar/doc.php?id=DOUE-L-2022-81573>

¹⁴ Reglamento (UE) del Parlamento Europeo y del Consejo, de 14 de septiembre de 2022, sobre mercados disputables y equitativos en el sector digital y por el que se modifican las Directivas (UE) 2019/1937 y (UE) 2020/1828 (Reglamento de mercados digitales). <https://eur-lex.europa.eu/legal-content/ES/TXT/?uri=celex%3A32022R1925>

¹⁵ Directiva 2000/31/CE, del Parlamento Europeo y del Consejo, de 8 de junio de 2000, relativa a determinados aspectos jurídicos de los servicios de la sociedad de la información, en particular en el comercio electrónico, en el mercado interior. <https://www.boe.es/buscar/doc.php?id=DOUE-L-2000-81295>

¹⁶ Ley 34/2002, de 11 de julio, de servicios de la sociedad de la información y de comercio electrónico. <https://www.boe.es/buscar/act.php?id=BOE-A-2002-13758>

Prestadores concernidos y sus obligaciones

El RSD se aplica específicamente a los denominados “servicios intermediarios” (que permiten el acceso, el alojamiento, almacenamiento y difusión de contenidos) independientemente de su lugar de establecimiento, siempre que ofrezcan dichos servicios a destinatarios establecidos o situados en la Unión Europea. Distingue a las plataformas y motores de búsqueda en línea de “muy gran tamaño” (*very large online platforms*, VLOP y *very large online search engines*, VLOSE) del resto de prestadores, estableciendo para ellos una serie de obligaciones adicionales atendiendo a su mayor impacto social. De acuerdo con el artículo 33.1 del Reglamento, se consideran como tales los que cuentan con un promedio mensual de destinatarios del servicio activos en la UE igual o superior al 10% de la población en este ámbito territorial en 2020,¹⁷ lo que supone en la actualidad unos 45 millones de usuarios.

La Comisión es la encargada de determinar qué prestadores cumplen esa característica, revisando anualmente los datos proporcionados desde los Estados miembros y por los propios prestadores. Inicialmente se han identificado 19 prestadores considerados de muy gran tamaño.¹⁸

AliExpress	Pinterest
Amazon Store	Snapchat
Apple App Store	TikTok
Bing	X (antes Twitter)
Booking	Wikipedia
Meta (Facebook, Instagram, LinkedIn)	Youtube
Google (Maps, Play, Search, Shopping)	Zalando

El 25 de agosto de 2023 entró en vigor el cumplimiento de las obligaciones previstas en el RSD para estos prestadores de muy gran tamaño.

¹⁷ De acuerdo con el apartado 2, la Comisión podrá con el tiempo ajustar ese criterio al alza o a la baja atendiendo al aumento o disminución de la población al menos en un 5 % con respecto a 2020.

¹⁸ Algunos de ellos, como Amazon o Zalando han recurrido la decisión de la Comisión ante el TJUE, por considerar que por sus características y actividad no se ajustan a la definición recogida en el Reglamento.

Exenciones de responsabilidad y obligaciones de los prestadores intermediarios

Los artículos 4 al 8 del Reglamento exoneran de responsabilidad a los prestadores de servicios intermediarios en relación con los contenidos de terceros que transmiten, alojan y/o almacenan, en línea con lo señalado tanto por la mencionada LSSI como por la Directiva de Servicios de Comunicación Audiovisual (DSCA / AVMSD)¹⁹ y la Ley General de la Comunicación Audiovisual (LGCA), que la incorpora al ordenamiento jurídico español.²⁰

El artículo 7, de forma un tanto alambicada, señala que “no se considerará que los prestadores de servicios intermediarios no reúnen las condiciones para acogerse a las exenciones de responsabilidad a que se refieren los artículos 4, 5 y 6 por la única razón de que realicen, de buena fe y de modo diligente, investigaciones por iniciativa propia de forma voluntaria, o adopten medidas con el fin de detectar, identificar y retirar contenidos ilícitos, o bloquear el acceso a estos, o adoptar las medidas necesarias para cumplir los requisitos del Derecho de la Unión y del Derecho nacional en cumplimiento del Derecho de la Unión, incluidos los requisitos establecidos en el presente Reglamento”.

Sí quedan sometidos estos prestadores, en todo caso, a la obligación de prontitud en la retirada de la información almacenada, o en el bloqueo del acceso a la misma, cuando tengan conocimiento o sean conscientes de su ilicitud; especialmente cuando una autoridad, sea judicial o administrativa de un Estado miembro, de conformidad con su ordenamiento jurídico, haya ordenado dicho retiro o bloqueo para poner fin o impedir una infracción.

Procedimientos de actuación

En línea con lo anterior (artículos 9 a 12), las autoridades judiciales o administrativas nacionales pueden dictar órdenes de actuación dirigidas a los prestadores y referidas a uno o varios elementos concretos de un contenido que se considera ilícito. Estas órdenes deben ser remitidas por la autoridad que las dicte al denominado coordinador de servicios digitales del correspondiente Estado miembro,²¹ el cual transmitirá una copia a todos los demás coordinadores de servicios digitales.

¹⁹ Directiva (UE) 2018/1808 del Parlamento Europeo y del Consejo de 14 de noviembre de 2018 por la que se modifica la Directiva 2010/13/UE sobre la coordinación de determinadas disposiciones legales, reglamentarias y administrativas de los Estados miembros relativas a la prestación de servicios de comunicación audiovisual (Directiva de servicios de comunicación audiovisual), habida cuenta de la evolución de las realidades del mercado. Disponible en: <https://www.boe.es/doue/2018/303/L00069-00092.pdf>

²⁰ Ley 13/2022, de 7 de julio, General de Comunicación Audiovisual. Disponible en: <https://www.boe.es/eli/es/l/2022/07/07/13/con>

²¹ De acuerdo con el artículo 46, los Estados miembros designarán como su coordinador de servicios digitales a la autoridad competente o a una de las autoridades competentes consideradas

Los prestadores, por su parte, deberán informar sin dilación indebida a la autoridad que haya dictado la orden, o a cualquier otra especificada en la misma, sobre su recepción y el curso dado a la orden, y también al destinatario. Los prestadores deben contar con un punto único de contacto que les permita ponerse en comunicación directamente, por vía electrónica, con las autoridades de los Estados miembros, con la Comisión y con otra figura prevista en el Reglamento, la Junta Europea de Servicios Digitales,²² así como con un punto único de contacto que permita a los destinatarios del servicio comunicarse directa y rápidamente con ellos por el canal de su elección, automatizado o no.

Los prestadores de servicios intermediarios que no tengan un establecimiento en la Unión Europea, pero sí ofrezcan servicios en la UE, deben contar con un representante legal en uno de los Estados miembros en los que el prestador ofrezca sus servicios, el cual (artículo 13) podrá ser considerado responsable por el incumplimiento de las obligaciones contempladas en el reglamento, sin perjuicio de la responsabilidad del prestador.

De acuerdo con el artículo 16, los prestadores establecerán mecanismos accesibles que permitan a cualquier persona física o entidad notificar la presencia en su servicio de elementos de información concretos que considere ilícitos. La notificación debe incluir una serie de requisitos (motivación, localización exacta, identificación del remitente, declaración de buena fe) que permitan al prestador determinar de modo rápido, sin un examen jurídico detallado, si existe esa supuesta ilicitud. Su decisión (diligente, no arbitraria y objetiva) será comunicada al remitente incluyendo las vías de recurso existentes.

Condiciones generales de prestación del servicio de la sociedad de la información

El artículo 14 se refiere a las condiciones generales de prestación del servicio de la sociedad de la información, obligando a los prestadores (apartados 1 a 4):

- A informar a los destinatarios sobre las condiciones generales del servicio y sobre cualquier restricción que impongan al mismo; sobre políticas, procedimientos, medidas y herramientas empleadas para moderar los contenidos, y también sobre la toma de decisiones mediante algoritmos.
- A informar a los destinatarios sobre las normas de procedimiento de su sistema interno de gestión de reclamaciones, en lenguaje claro, sencillo, inteligible,

responsables de la supervisión de los prestadores de servicios intermediarios y de la ejecución del Reglamento. El coordinador de servicios digitales será el encargado de cooperar con sus homólogos en toda la Unión y con otras autoridades nacionales. Los Estados miembros deben designar a sus coordinadores de servicios digitales a más tardar el 17 de febrero de 2024.

²² La Junta Europea de Servicios Digitales (artículos 61 a 63) es un grupo consultivo independiente integrado por coordinadores de servicios digitales para la supervisión de los prestadores de servicios intermediarios, con el objetivo de asesorar a dichos coordinadores y a la Comisión sobre el cumplimiento del RSD.

accesible al usuario e inequívoco. Si el servicio está dirigido principalmente a menores, las condiciones y restricciones se explicarán de manera que los menores las puedan comprender.

- A actuar de manera diligente, objetiva y proporcionada para aplicar y hacer cumplir las posibles restricciones con la debida consideración de los derechos e intereses legítimos de todas las partes implicadas, incluidos la libertad de expresión, el pluralismo de los medios de comunicación y otros amparados por la Carta de los Derechos Fundamentales.

Los prestadores publicarán al menos una vez al año informes sobre su actividad de moderación de contenidos,²³ incluyendo según proceda (artículo 15):

- El número de órdenes recibidas sobre contenido ilícito, el Estado miembro que haya dictado la orden y el tiempo medio necesario para informar a la autoridad y para dar curso a la orden.
- Información significativa y comprensible sobre la actividad de moderación de contenidos realizada por iniciativa propia del prestador, incluido el uso de herramientas automatizadas; la formación y asistencia a las personas encargadas de la moderación; la disponibilidad, visibilidad y accesibilidad de la información proporcionada por los destinatarios, y otras restricciones conexas del servicio.
- El número de reclamaciones recibidas a través de los sistemas internos de gestión de las mismas, su base, las decisiones adoptadas, el tiempo medio para su adopción y su posible revocación.
- El uso de medios automatizados con fines de moderación de contenidos, incluyendo la posible tasa de error de dichos medios y las salvaguardias aplicadas.
- En el caso de prestadores de servicios de alojamiento de datos, el número de notificaciones según el tipo de contenido presuntamente ilícito; las actuaciones llevadas conforme a Derecho o en base a las condiciones generales del prestador; el número de notificaciones tratadas únicamente por medios automatizados, y el tiempo medio necesario para adoptar medidas.

En el caso de plataformas y motores de búsqueda en línea de muy gran tamaño, deberán facilitar además a los destinatarios de los servicios un resumen sucinto que incluya las medidas correctivas y los mecanismos de recurso disponibles.

²³ De acuerdo con el apartado 2 de este artículo, estas obligaciones de transparencia no se aplican a los prestadores de servicios intermediarios que sean microempresas o pequeñas empresas, ni a las plataformas en línea que no sean de muy gran tamaño.

Los prestadores de servicios de alojamiento de datos deberán proporcionar a cualquier destinatario (artículo 17) una declaración de motivos clara y específica cuando el servicio se vea afectado por las restricciones impuestas a un contenido considerado ilegal o incompatible con sus condiciones generales,²⁴ como la eliminación de contenidos; el bloqueo del acceso a éstos o su relegación; la suspensión, cesación u otra restricción de los pagos monetarios; la suspensión o cesación total o parcial de la prestación del servicio; la suspensión o supresión de la cuenta del destinatario del servicio.

Cuando un prestador de servicios de alojamiento de datos tenga conocimiento de cualquier información que le haga sospechar que se han cometido, se están cometiendo o es probable que se cometan delitos que impliquen una amenaza para la vida o la seguridad de una o más personas, debe comunicar su sospecha a las autoridades policiales o judiciales del Estado o Estados miembros afectados, aportando toda la información pertinente de la que dispongan, o en su caso, a las autoridades policiales del Estado miembro en que esté establecido (artículo 18.).

Disposiciones adicionales aplicables a las plataformas en línea de muy gran tamaño

Además de las ya mencionadas, estas disposiciones se refieren (artículos 20 a 28):

- A la gestión de reclamaciones y a los sistemas internos para su gestión. En caso de no admisión de una reclamación, ha de informarse a los reclamantes sobre la decisión de forma motivada, así como sobre la posibilidad de acudir a la resolución extrajudicial de litigios. Las decisiones que se adopten en relación con estas reclamaciones deberán contar con la supervisión de personal adecuadamente cualificado y no exclusivamente por medios automatizados.
- A la resolución extrajudicial de litigios. El recurso a estos órganos encargados de esta resolución, certificados por los coordinadores de servicios digitales, es gratuito, no vinculante, y no afecta al derecho de los reclamantes a acudir a los órganos jurisdiccionales correspondientes.
- A los alertadores fiables dedicados a notificar contenidos ilícitos. Las características de estos alertadores se detallan en el artículo 22, pero básicamente pueden serlo cualquier entidad con conocimientos y competencias específicas para detectar dichos contenidos ilícitos, siempre que no dependan de ningún prestador de plataformas en línea, y realicen sus actividades de envío de notificaciones de manera diligente, precisa y objetiva. Los alertadores fiables deben ser reconocidos como tales por parte del coordinador de servicios digitales de su

²⁴ Salvo que el prestador no conozca los datos de contacto electrónico o cuando la información sea un contenido comercial engañoso de gran volumen.

país, y las plataformas deben otorgar prioridad a las notificaciones enviadas por dichos alertadores fiables.

- A las medidas y protección contra usos indebidos, actuando contra los destinatarios de sus servicios que proporcionen con frecuencia contenidos manifiestamente ilícitos (suspensión temporal), pero también contra quienes notifican y reclaman de modo infundado, sean personas físicas o entidades.
- A la transparencia, incluyendo en sus informes datos sobre los litigios sometidos a los órganos de resolución extrajudicial. Asimismo, deben informar en sus condiciones generales sobre los parámetros principales utilizados en sus sistemas de recomendación, ofreciendo cuando sea posible la opción de modificar, influir en los parámetros o disponer de una funcionalidad accesible y fácil que les permita seleccionar y modificar en cualquier momento sus preferencias.
- A la veracidad, estableciéndose que los prestadores de plataformas en línea no pueden engañar o manipular a los destinatarios distorsionando u obstaculizando su capacidad de tomar decisiones libres e informadas mediante el diseño, organización o gestión de sus interfaces en línea.
- A la publicidad. Los prestadores deben asegurarse de que los destinatarios del servicio sean capaces de identificar los mensajes de naturaleza publicitaria, sus responsables, los principales parámetros utilizados para determinar el destinatario y, en su caso, cómo cambiarlos. No se presentarán anuncios basados en la elaboración de perfiles, utilizando las categorías especiales de datos personales, de acuerdo con los criterios recogidos en el Reglamento General de Protección de Datos.²⁵
- A la protección de los menores. Cuando los contenidos de las plataformas en línea sean accesibles a personas menores de edad, se establecerán medidas adecuadas y proporcionadas para garantizar un elevado nivel de privacidad, seguridad y protección. Los prestadores no presentarán anuncios en su interfaz basados en la elaboración de perfiles, mediante la utilización de datos personales cuando sean conscientes “con una seguridad razonable” de que el destinatario del servicio es un menor, sin que ello requiera tratar datos personales adicionales a fin de evaluar si el destinatario del servicio es un menor.

En el caso de prestadores de plataformas en línea que permiten a los consumidores celebrar contratos a distancia, se establecen para estos prestadores determinadas obligaciones que tienen que ver con la trazabilidad, la información y el cumplimiento de las obligaciones

²⁵ Reglamento (UE) 2016/679 del Parlamento Europeo y del Consejo, de 27 de abril de 2016, relativo a la protección de las personas físicas en lo que respecta a la protección de las personas físicas en lo que respecta al tratamiento de datos personales y a la libre circulación de estos datos y por el que se deroga la Directiva 95/46/CE.

desde el diseño (artículos 30 a 32). Estas obligaciones no afectan a las microempresas o pequeñas empresas (o a las que lo han sido en los doce meses anteriores), salvo que se traten de plataformas y motores de búsqueda considerados de muy gran tamaño (artículo 29).

Riesgos sistémicos

El Reglamento considera riesgos sistémicos (artículo 34) los efectos negativos, reales o previsibles, derivados de la difusión de contenido ilícito a través de los servicios prestados por las plataformas, que afectan:

- Al ejercicio de los derechos fundamentales amparados por la Carta de Derechos (dignidad; privacidad e intimidad; libertad de expresión e información, libertad y pluralismo de los medios de comunicación; no discriminación; derechos de la infancia, protección de los consumidores).
- Al discurso cívico y a los procesos electorales, así como a la seguridad pública (nacional).
- A la violencia de género, a la protección de la salud pública y al bienestar físico y mental de la persona.

Atendiendo a la tipología de esos riesgos, se reclama de los prestadores (artículos 34 a 42) la aplicación de medidas para su reducción que sean razonables, proporcionadas y efectivas, y que pueden incluir:

- La seguridad desde el diseño; la adaptación de las características y funcionamiento del servicio, incluidas sus interfaces en línea, así como la adaptación de sus condiciones generales y de la ejecución de las mismas.
- La adaptación de los procesos de moderación de contenidos, y, en su caso, la rápida retirada o bloqueo de los ilícitos detectados.
- La adaptación de sus sistemas algorítmicos y de recomendación. Ofrecerán al menos una opción para cada uno de ellos que no se base en la elaboración de perfiles.
- La adaptación de sus sistemas publicitarios. La creación de un repositorio que recoja los anuncios difundidos, sus características (nombre del producto, servicio o marca y objeto del anuncio), la identidad de los responsables, el periodo de difusión, su target y alcance.
- La ejecución de las decisiones de los órganos de resolución extrajudicial de litigios.

- La cooperación con otros prestadores mediante los códigos de conducta y protocolos de crisis.
- La mayor información y transparencia ante los destinatarios del servicio. Proporcionando una funcionalidad fácil de utilizar que les permita señalar los contenidos ilícitos y obtener ayuda.
- La adopción de medidas de concienciación.
- La protección de los derechos de los menores, incluidas herramientas de verificación de edad y de control parental.
- La adopción de medidas de respuesta a las crisis que den lugar a una amenaza grave y extraordinaria para la seguridad o la salud públicas, a solicitud de la Comisión.
- La realización de auditorías independientes para evaluar el cumplimiento de sus obligaciones y compromisos.
- El acceso a datos de seguimiento y evaluación del cumplimiento por parte del coordinador de servicios digitales de establecimiento o de la Comisión cuando los soliciten de forma motivada y en un período razonable.
- La cooperación con los alertadores fiables, con los verificadores, con los expertos, a los que proporcionarán acceso a los datos con la finalidad de realizar estudios que contribuyan a la detección, determinación y comprensión de los riesgos sistémicos. Entre ellos los afiliados a organismos, organizaciones y asociaciones sin ánimo de lucro, que cumplan determinados requisitos.

El Reglamento contempla una figura dedicada a la comprobación del cumplimiento, que sea independiente de las funciones operativas y que cuente con autoridad, rango y recursos suficientes para hacer dicha comprobación, colaborando con el coordinador de servicios digitales de establecimiento y con la Comisión. Esa figura o función debe velar para que se detecten y notifiquen adecuadamente todos los riesgos y para que se adopten las medidas necesarias de reducción de los mismos; organizar y supervisar las actividades del prestador en relación con la auditoría independiente; asesorar sobre las obligaciones a cumplir y realizar un seguimiento de cumplimiento de dichas obligaciones.

Se refiere también a la publicación de informes periódicos, especificando aspectos como los recursos humanos dedicados a la moderación, sus cualificaciones y conocimientos. Además, las plataformas y buscadores de muy gran tamaño deben incluir información sobre el promedio mensual de destinatarios del servicio para cada Estado miembro; los resultados de la evaluación de riesgos; las medidas de reducción de riesgos específicas; los resultados de la auditoría; en su caso, información sobre consultas realizadas por el prestador en apoyo de las evaluaciones de riesgos y el diseño de las medidas de reducción de riesgos.

Y se contempla el abono a la Comisión de una tasa de supervisión anual para cubrir los costes estimados en que ésta incurra en relación con sus funciones de supervisión.

La correulación como complemento a las obligaciones de los prestadores

Códigos de conducta

El artículo 45 contempla el fomento y facilitación por parte de la Comisión de la Junta de códigos de conducta, señalando que, cuando se genere un riesgo sistémico significativo y afecte a varios prestadores, la Comisión podrá invitar a dichos prestadores y a otras partes interesadas, a participar en la elaboración de dichos códigos, en particular estableciendo compromisos de adopción de medidas específicas de reducción de riesgos, así como un marco de información periódica sobre las medidas que se puedan adoptar y sus resultados.

Entre los posibles participantes en la elaboración de los códigos, el Reglamento menciona expresamente a los prestadores directamente involucrados o a otros (y no solo de muy gran tamaño, si fuera oportuno); a otros servicios intermediarios; a las organizaciones de la sociedad civil, y a las autoridades competentes.

Los códigos de conducta deberán exponer claramente sus objetivos específicos y los compromisos adquiridos; contener indicadores clave de eficacia para valorar el cumplimiento de dichos objetivos, y tener debidamente en cuenta las necesidades e intereses de todos los interesados, en particular de los ciudadanos de la UE. Los participantes deberán informar periódicamente a la Comisión y a sus respectivos coordinadores de servicios digitales de establecimiento acerca de las medidas que puedan adoptarse y sus resultados.

La Comisión y la Junta evaluarán si los códigos de conducta cumplen los fines arriba especificados, teniendo en cuenta las diferencias de tamaño y capacidad de los diferentes participantes: publicarán sus conclusiones, y promoverán su revisión periódica. En caso de incumplimiento sistemático de los códigos, podrán pedir a los signatarios “que adopten las medidas necesarias” para solucionar esos incumplimientos. Hay que considerar que el Código de buenas prácticas en materia de desinformación, que pasaría a convertirse, como decíamos, en código de conducta

De modo más específico, los artículos 46 y 47 se refieren a los Códigos de conducta en materia de publicidad en línea y accesibilidad, contemplando que estén elaborados “a más tardar el 18 de febrero de 2025” y que su aplicación sea efectiva “a más tardar el 18 de agosto de 2025”.

Protocolos voluntarios para situaciones de crisis

Señala el artículo 48 del Reglamento, que la Junta podrá recomendar que la Comisión inicie la elaboración de protocolos voluntarios destinados a abordar situaciones de crisis, los cuales se limitarán estrictamente a circunstancias extraordinarias que afecten a la seguridad pública o a la salud pública.

La Comisión, por su parte, fomentará y facilitará que los prestadores de las plataformas en línea y los motores de búsqueda en línea, no solo los de muy gran tamaño, tomen parte en la elaboración, realización de pruebas y aplicación de dichos protocolos de crisis.

Los protocolos deber incluir al menos alguna o algunas de estas medidas:

- Establecer los parámetros específicos para determinar qué constituye la circunstancia extraordinaria específica que el protocolo de crisis pretende abordar y los objetivos que persigue.
- Determinar el papel de cada uno de los participantes y las medidas que deban adoptar en la preparación del protocolo de crisis una vez activado.
- Contar con un procedimiento claro para determinar cuándo ha de activarse el protocolo de crisis, así como para determinar el período durante el cual deban aplicarse las medidas.
- Garantizar que el prestador de servicios intermediarios designe un punto de contacto específico para la gestión de crisis.
- Contemplar salvaguardias para contrarrestar los posibles efectos negativos de los protocolos de crisis en el ejercicio de los derechos fundamentales amparados por la Carta de Derechos, en particular la libertad de expresión e información y el derecho a la no discriminación.
- Una vez finalizada la situación de crisis, abrir un proceso para informar públicamente sobre las medidas que se adopten, su duración y sus resultados.

Si la Comisión considera que un protocolo de crisis no es eficaz para abordar la situación de crisis, o para salvaguardar el ejercicio de los derechos fundamentales, solicitará a los participantes su revisión o la adopción de medidas adicionales.

Sanciones previstas

En caso de que los servicios digitales no cumplan con estas obligaciones señaladas, el Reglamento contempla una serie de sanciones que no podrán superar el 6% del volumen de negocios anual global de la compañía multada.

En el caso de las plataformas *online* y motores de búsqueda de muy gran tamaño, amplía los supuestos en los que las compañías pueden ser sancionadas por cometer una violación de sus obligaciones, como en el incumplimiento de medidas cautelares o de los compromisos que hayan sido declarados vinculantes. También pueden recibir sanciones si proporcionan información incorrecta o si se niegan a someterse a una inspección por parte de las autoridades europeas.

La desinformación en el Reglamento de servicios digitales

En su parte dispositiva, el Reglamento no incluye menciones expresas a la desinformación, pero sí en los considerandos de la norma. Así:

- En el considerando 2 se señala que los Estados miembros están adoptando, o considerando adoptar, un número creciente de normas de Derecho nacional sobre las materias que regula el presente Reglamento, imponiendo, en particular, requisitos de diligencia a los prestadores de servicios intermediarios por lo que se refiere al modo en que deben hacer frente a los contenidos ilícitos, la desinformación y otros riesgos para la sociedad.
- En el considerando 9, que el Reglamento armoniza las normas aplicables a los servicios intermediarios en el mercado interior con el objetivo de garantizar un entorno en línea seguro, predecible y digno de confianza y aborda la difusión de contenidos ilícitos en línea y los riesgos para la sociedad que puede generar la difusión de desinformación u otros contenidos, dentro del cual se protegen eficazmente los derechos fundamentales reconocidos en la Carta y se facilita la innovación.
- En el considerando 69, que cuando se presentan a los destinatarios anuncios basados en técnicas de segmentación optimizadas para responder a sus intereses y apelar potencialmente a sus vulnerabilidades, los efectos negativos pueden ser especialmente graves. En algunos casos, las técnicas de manipulación pueden afectar negativamente a grupos enteros y amplificar perjuicios sociales, por ejemplo, contribuyendo a campañas de desinformación o discriminando a determinados grupos.
- En el considerando 83 menciona las campañas coordinadas de desinformación como encuadradas dentro de la cuarta categoría de riesgos sistémicos, especialmente relacionadas con la salud pública, o del diseño de interfaces en línea que puedan estimular adicciones comportamentales de los destinatarios del servicio.
- En el considerando 84, se indica que los prestadores deben dedicar especial atención a cómo se utilizan sus servicios para difundir o amplificar contenidos incorrectos o engañosos, incluida la desinformación. Es el caso de la creación

de cuentas falsas, el uso de bots y otros comportamientos total o parcialmente automatizados, que pueden dar lugar a la difusión rápida y extendida de desinformación.

- En el considerando 88, se indica que los prestadores de plataformas en línea de muy gran tamaño y de motores de búsqueda en línea de muy gran tamaño deben sopesar medidas de concienciación, en especial cuando los riesgos estén relacionados con campañas de desinformación.
- En el considerando 95, se señala que estas plataformas y buscadores deben garantizar el acceso público a los repositorios de anuncios publicitarios presentados en sus interfaces en línea para facilitar la supervisión y la investigación de los riesgos emergentes generados por la distribución de publicidad en línea, por ejemplo en relación con anuncios ilícitos o técnicas manipulativas y desinformación con efectos negativos reales y previsibles para la salud pública, la seguridad pública, el discurso civil, la participación política y la igualdad.
- En el considerando 104, que se refiere, en el marco de los sistemas de autorregulación y corregulación, a las posibles repercusiones negativas para la sociedad y la democracia de las operaciones coordinadas dirigidas a amplificar la desinformación, para generar información deliberadamente incorrecta o engañosa.
- En el considerando 106, señala que las disposiciones del Reglamento relativas a los códigos de conducta podrían servir de base para iniciativas de autorregulación ya establecidas, como el Código de buenas prácticas en materia de desinformación.
- En el considerando 108, que se refiere a la potestad de la Comisión para iniciar la elaboración de protocolos de crisis, por ejemplo, cuando las plataformas en línea se utilizan de forma indebida para propagar rápidamente contenidos ilícitos o desinformación, o bien cuando surja la necesidad de difundir rápidamente información fiable.

CONCLUSIONES

Resumen

1. La “desinformación”, tal y como se define canónicamente por la Comisión Europea, es “información verificablemente falsa o engañosa que se crea, presenta y divulga con fines lucrativos o para engañar deliberadamente a la población y que puede causar un perjuicio público”.
2. De acuerdo con esa definición, la desinformación como fenómeno se deslinda de la mera información errónea, fruto del desconocimiento o de la falta de diligencia en la procura de la veracidad, respondiendo a una clara intencionalidad de dolo o engaño por parte de sus responsables. Pero también cabría establecer su diferencia con la mera difusión, singular o reiterada, de noticias falsas, requiriendo de una dimensión estratégica, sistemática, que responde al objetivo comercial y/o ideológico de imponer un relato “alternativo” al que se deriva de los propios hechos o, al menos, de generar la suficiente confusión como para sembrar la duda y la desconfianza hacia esos hechos.
3. Existen sectores de la opinión pública, ya sea en el ámbito político o económico, interesados en promover la desinformación, o en negar o desacreditar la realidad desde el marco de la posverdad (el relato frente al dato) para beneficiar a sus propios intereses. Las posibilidades que ofrecen internet y las tecnologías de la comunicación, a la hora de propagar mensajes, maximizar la interactividad y segmentar a los receptores, han permitido el desarrollo exponencial del fenómeno de la desinformación.
4. En la lucha contra la desinformación es fundamental actuar contra los entramados y grupos de interés que la generan, sacándolos a la luz; poniendo de relieve sus intenciones, así como sus tácticas y estrategias; denunciando sus conexiones y a los colaboradores necesarios; garantizando el derecho constitucional de la ciudadanía a recibir información veraz. Pero también se trata de potenciar la capacidad crítica y la responsabilidad individual y social de esa ciudadanía, de empoderar a las personas para evitar su complicidad con la desinformación, ya sea como consumidores o como redifusores de sus mensajes.
5. De ahí la importancia de desarrollar una actuación coordinada por parte de todos los agentes vinculados (autoridades públicas, prestadores de servicios de la sociedad de la información, verificadores, expertos, sociedad civil) que permita la prevención y la respuesta rápida ante esta lacra que, como hemos tenido ocasión de comprobar en los últimos años, puede afectar gravemente a la salud de las personas; a la seguridad personal, social, nacional y comunitaria, y al propio

funcionamiento del sistema democrático, como lo prueba la injerencia en los procesos electorales por parte de determinados países y grupos.

6. La lucha contra la desinformación en internet requiere de una respuesta legal, global y armonizada a nivel internacional, siguiendo la estela de las iniciativas que en su momento se desarrollaron a raíz del aumento de los delitos tecnológicos derivados de la irrupción de internet en nuestra sociedad en los años 90, y cuya consecuencia más evidente fue la aprobación del Convenio de Budapest sobre la Ciberdelincuencia.
7. El análisis realizado, que se resume en el presente capítulo, tiene en cuenta el camino seguido en el seno de la Unión Europea en la lucha contra la desinformación desde los primeros documentos de la Comisión, para centrarse en el Código de buenas prácticas en materia de desinformación (en sus versiones de 2018 y 2022) y en la reciente Ley de servicios digitales (2022).
8. Se identifican en esos documentos una serie de aspectos clave en la lucha contra la desinformación, que buscan garantizar su eliminación (e incluso su evitación), o que, al menos, impidan cualquier lucro asociado a la monetización de la difusión de contenidos desinformativos, en especial los relacionados con la salud, la seguridad y la propaganda política.

Estos aspectos pasan, *grosso modo*, por:

- La transparencia: asegurar un conocimiento suficiente sobre los procesos, las herramientas, y los conceptos manejados en la gestión y oferta de contenidos por parte de los prestadores de servicios de la sociedad de la información y de los distintos agentes de la cadena de valor. Con especial atención a los denominados algoritmos que, como la profecía autocumplida, determinan a partir del análisis de patrones previos de comportamiento digital las propuestas y recomendaciones de información por parte de plataformas, redes sociales y motores de búsqueda.
- La privacidad: limitación del uso de datos personales, sobre todo de los más sensibles, y de la utilización de perfiles, de acuerdo con lo establecido por la normativa europea en esta materia. En la garantía de la privacidad desempeña un papel fundamental el consentimiento expreso de los usuarios, como una barrera para la difusión de contenidos de desinformación.
- La colaboración con verificadores y expertos: compromiso, en el marco del principio de transparencia, por parte de las plataformas y redes, de permitir la labor analítica de verificadores e investigadores expertos reconocidos sobre el flujo de contenidos que difunden. Ello implica la elaboración e implementación de indicadores de fiabilidad de las fuentes basados en criterios objetivos e independientes. Y, asimismo, información general sobre el uso de algoritmos y el intercambio de conjuntos de datos, siempre protegiendo la intimidad de las personas.

- El empoderamiento de la ciudadanía: contribuir a la mejora de la capacidad de los usuarios para encontrar contenido fiable, mediante iniciativas de alfabetización mediática e informacional y pensamiento crítico para todos los segmentos de la sociedad. Y también mediante el desarrollo de herramientas accesibles que permitan, desde el diseño, contar con indicadores sobre la fiabilidad de las fuentes y la identidad de éstas, garantizando también la diversidad en la recepción de contenidos lícitos.
9. La aprobación del Código de buenas prácticas en materia de desinformación supuso un hito fundamental para la implicación de las grandes plataformas de internet y de las redes sociales en esta materia. Su extensión, actualización y reforzamiento, a partir de las recomendaciones de la Comisión, ha profundizado en esta implicación. Ello puede comprobarse mediante la lectura de los diferentes informes de rendición de cuentas puestos a disposición por los firmantes en el Centro de Transparencia, en los que detallan sus acciones en las grandes áreas arriba señaladas y el grado de cumplimiento de aquellos compromisos que han decidido voluntariamente asumir el conjunto de los contemplados en el Código.
 10. Cabe señalar, no obstante, que el contenido de esos informes es muy desigual, a pesar de contar con un modelo canónico para presentar los resultados. En algunos casos se ofrecen datos muy detallados por actuaciones y países, lo que ha permitido conocer, por ejemplo, las desarrolladas en España por parte de algunas de las grandes plataformas en línea. En otros, los informes parecen más bien una exposición teórica de declaración de intenciones, sin que sea fácil hacerse una idea del alcance de lo verdaderamente realizado, al menos en lo que respecta a la lucha contra la desinformación.

En todo caso, los informes muestran el camino recorrido y el que todavía queda por recorrer, con algún retroceso puntual como la retirada de Twitter (ahora X) como firmante del Código. Esta retirada refleja, a su vez, tanto las posibilidades como los límites de un sistema que pudiera quedar meramente circunscrito al marco de la regulación voluntaria. Y permite valorar de forma clara la importancia de la aprobación y la entrada en vigor del Reglamento de servicios digitales, especialmente en lo relativo a las plataformas en línea, redes sociales y motores de búsqueda de muy gran tamaño, teniendo en cuenta su enorme impacto social.

Propuestas

A la luz tanto del resultado de la aplicación del Código de buenas prácticas reforzado como del marco normativo establecido por el Reglamento de servicios digitales, consideramos fundamental de cara al futuro adoptar una serie de iniciativas que garanticen la eficacia, pero también la transparencia (si es que puede diferenciarse entre ambos conceptos) en la lucha contra la desinformación.

Lo deseable es que el Código de buenas prácticas reforzado se integre en el marco del Reglamento de servicios digitales, recibiendo la consideración de código de conducta prevista en dicha norma.

Sobre las plataformas y redes sociales

Es necesario que las plataformas de internet y las redes sociales se comprometan de un modo más proactivo en la lucha contra la desinformación y otros contenidos ilícitos.

La ausencia de responsabilidad editorial específica de plataformas y redes en relación con los contenidos de terceros que dichas plataformas y redes soportan y difunden, no puede entenderse como una ausencia de responsabilidad general en lo relativo a la supervisión y garantías de licitud ante esos contenidos. Su intervención en la organización de dichos contenidos y el uso de algoritmos y recomendaciones influye en la segmentación y selección de los mensajes que reciben los usuarios, sin que puedan escudarse en la neutralidad tecnológica. La segmentación resulta condicionante a la hora de reafirmar determinados conocimientos, actitudes creencias y comportamientos por parte de los usuarios, y su empleo inadecuado fomenta la crispación, la polarización, la confrontación y la desinformación, penalizando el pluralismo y ocultando las posiciones discrepantes.

Se trata, por tanto, de avanzar en la implicación de los proveedores de servicios de comunicación en línea. No solamente introduciendo transparencia en los procesos o actuando *a posteriori* con agilidad y eficacia contra las alertas y denuncias, sino también exigiendo *a priori* a quienes soportan y difunden contenidos a través de sus plataformas y redes la acreditación del cumplimiento de los indicadores de veracidad, fiabilidad y licitud que puedan establecerse.

Para ello es fundamental avanzar también en el conjunto de parámetros que sirvan de marco de referencia compartida para clasificar y etiquetar contenidos y determinar cuáles pueden considerarse desinformación, y tienen un mayor impacto sobre de la sociedad; qué tipo de contenidos son claramente ilícitos por otras razones, e incluso qué mensajes pueden ser situados en la denominada “zona gris” - legales pero dañinos o inadecuados -, con la finalidad en cada caso de eliminarlos o limitarlos gracias a su detección temprana.

Los informes de rendición de cuentas y los propios anuncios de los prestadores a partir de la aprobación del Reglamento muestran un esfuerzo por parte de plataformas y redes por

adecuarse a la nueva situación (en unos casos más que en otros), haciendo hincapié en aspectos como:

- Las medidas para impedir la colocación de creatividades publicitarias en páginas y dominios con contenidos de desinformación.
- El desarrollo de la información de contexto; la colaboración con verificadores y expertos y con sus organizaciones, incluyendo las API para recopilar y analizar datos.
- La eliminación de contenidos ilícitos (entre ellos, de desinformación y operaciones de influencia encubiertas) o por incumplir las condiciones de servicio en los contenidos.
- La eliminación de cuentas falsas y la reducción de la viralidad, mediante advertencias, de contenidos no contrastados.
- Los criterios de búsqueda y recomendación no perfilados y la transparencia en el uso de algoritmos.
- La capacitación de los usuarios mediante herramientas de denuncia e información sobre los contenidos recibidos, y campañas mediante la alfabetización mediática online y offline, incluyendo el uso moderado y responsable de internet.

Pero esta ampliación de los compromisos y obligaciones de plataformas y redes sociales requiere de acciones más consistentes, desde el punto de vista tanto cuantitativo como cualitativo. Por ejemplo, integrando la colaboración de verificadores y expertos con los moderadores en el proceso de toma de decisiones. Con obligaciones claras y tasadas de transparencia en la aplicación de procesos y herramientas, cuyo conocimiento por parte de los expertos acreditados no dependa meramente de la voluntad de estos operadores.

En resumen, y en línea con lo señalado por la Comisión Europea en sus informes de evaluación del Código de buenas prácticas, es necesario promover un ecosistema de información más responsable por parte de las plataformas; una mayor eficacia en la capacidad de verificación de datos, así como en el fomento del uso de nuevas tecnologías para mejorar la manera de producir y difundir información en línea. Ello significa que las plataformas no deberían limitarse a cumplir con las obligaciones legales establecidas por legislación comunitaria y nacional, sino que deberían actuar también de modo más proactivo y responsable de acuerdo con su papel fundamental a la hora de garantizar un entorno en línea seguro; de proteger a los usuarios contra la desinformación, y de ofrecer puntos de vista diversos y plurales.

Esperamos que los informes de rendición de cuentas pendientes de presentar en el momento de finalización de este capítulo presenten de forma más sistemática y rigurosa, menos discrecional, las acciones realizadas. Especialmente porque con la entrada en vigor del Reglamento de servicios digitales, algunos de los compromisos suscritos en el Código han devenido en obligaciones.

Sobre el marco de actuación contra la desinformación

Teniendo en cuenta tanto los compromisos asumidos por los prestadores firmantes del Código de buenas prácticas en materia de desinformación como la progresiva aplicación del Reglamento de servicios digitales, este Grupo de Trabajo considera fundamental la puesta en funcionamiento en nuestro país, a la mayor brevedad de los sistemas de resolución extrajudicial de conflictos (ADR) previstos en el Reglamento, aprovechando la experiencia ya existente en la aplicación de este tipo de sistemas en materia de comunicaciones comerciales, como la que viene desarrollando la Asociación para la Autorregulación de la Comunicación Comercial (AUTOCONTROL). Ello con el fin de resolver las eventuales reclamaciones que puedan producirse contra las decisiones de supresión o no de contenidos adoptadas por las plataformas, redes y motores de búsqueda, garantizando los derechos de los usuarios ante la desinformación sin menoscabo de la actuación tuitiva de las autoridades administrativas y judiciales.

Igualmente, sería deseable la pronta constitución del coordinador de servicios digitales del Estado miembro, figura prevista también por el Reglamento. Una competencia que, atendiendo a lo señalado, podría asumir la Comisión Nacional de los Mercados y la Competencia (CNMC), como autoridad audiovisual que forma parte de ERGA en el marco europeo y que, por sus características, puede ejercer con eficacia y rigor esa función ante prestadores, usuarios y otros agentes vinculados.

La creación del coordinador de servicios digitales permitirá, a su vez, la designación de los alertadores fiables previstos en el reglamento, cuya aportación desde la sociedad civil consideramos esencial en el proceso colaborativo de lucha contra la desinformación.

En relación con el Código de buenas prácticas, consideramos asimismo fundamental la constitución de un grupo de seguimiento y evaluación a nivel español, que puede conocer de modo específico el cumplimiento de los compromisos de los firmantes, en especial de las plataformas, redes sociales y motores de búsqueda de mayor tamaño, en nuestro territorio.

Una opción sería dotar de continuidad y permanencia al Foro creado por la Dirección de Seguridad Nacional, con las adaptaciones necesarias, con la participación de todos los agentes implicados, ante el que pudieran comparecer los prestadores firmantes del código de conducta y aquellos que estén obligados al cumplimiento de lo previsto en el Reglamento. Una de las misiones de este Foro contra la desinformación, además de evaluar periódicamente las acciones de estos prestadores, sería la de proponer recomendaciones a las plataformas y a las Autoridades Nacionales y/o europeas.

En el marco de la lucha contra la desinformación, se valora como también como muy relevante la experiencia acumulada tras la elaboración y puesta en marcha de la Agenda Europea para la Seguridad, presentada por la Comisión Europea en el año 2015, que abordó como prioridad la lucha contra la difusión de contenidos propagandísticos en el marco de la actividad contraterrorista. Para ello se creó una estructura a nivel europeo

cuyo centro de coordinación radica en la agencia europea Europol y que da soporte a los Estados miembros en la labor de identificar y eliminar los contenidos y en cooperación con la industria privada.

En este sentido, desde el año 2015, se ha llevado a cabo la creación en diferentes Estados miembros de la UE, incluido España, de unidades especializadas en la monitorización, señalamiento y notificación de contenidos ilícitos a las empresas privadas de internet, denominadas *Internet Referral Unit* (IRU). Estas unidades, además, realizan habitualmente ejercicios coordinados en los diferentes países para el señalamiento y retirada de determinados contenidos ilícitos denominados *Referral Action Days* (RAD), que en ocasiones se focalizan en plataformas específicas de internet, fomentando así la coordinación, la optimización de los protocolos de comunicación y la interacción con las compañías privadas, y profundizando en el conocimiento común y en la acertada valoración de la tipología de contenidos ilícitos difundidos en internet.

A tenor de lo expuesto, se valora la utilidad de crear unidades o grupos permanentes de trabajo especializados en la monitorización, y el señalamiento y valoración del alcance de los contenidos desinformativos, que coordinen a nivel nacional, y se coordinen a nivel internacional, en este ámbito.

Sobre la participación de otros implicados en la lucha contra la desinformación

Este grupo de trabajo se pronuncia a favor de una participación más activa por parte de las organizaciones de la sociedad civil, en el marco de la garantía de los derechos, pero también del cumplimiento de su responsabilidad como difusores o redifusores de mensajes, en los procesos de seguimiento y evaluación de cumplimiento de las obligaciones de los operadores en línea en relación con la desinformación, la aplicación de algoritmos, la utilización de patrones oscuros, las comunicaciones comerciales desleales, los mensajes de odio y discriminación, los contenidos sexistas y la protección de los menores.

Formaría parte también de este modelo colaborativo la creación de canales de comunicación entre la sociedad civil, la Administración y los prestadores en línea con el modelo seguido por el Centro de Coordinación de Información sobre la Radicalización (CCIR).²⁶ Este Centro, además de recibir información ciudadana anonimizada sobre individuos que podrían estar inmersos en un proceso de radicalización violenta (en nuestro caso sobre

²⁶ Creado en diciembre de 2015 en el marco de la elaboración del Plan Estratégico Nacional de Lucha contra la Radicalización Violenta (PEN-LCRV). Recibe esa información a través de un teléfono 24/7 (900 822 066), de una página web (www.stop-radicalismo.es) y de un correo electrónico (stop-radicalismo@interior.es).

fuentes de desinformación) a través de un canal *online* sencillo y accesible que cuenta con aplicaciones para móvil, desarrolla una labor de información a la ciudadanía en momentos de crisis o estrés social y realiza funciones de elaboración contra narrativas ante la difusión de contenidos concretos, creando y difundiendo mensajes de tranquilidad dirigidos a que la ciudadanía perciba un apoyo institucional de forma objetiva, profesional y con un lenguaje comprensible.

Es fundamental, igualmente, potenciar, a nivel nacional, las campañas de sensibilización contra la desinformación por parte de los poderes públicos, como parte de las medidas de alfabetización mediática contempladas por la Ley General de la Comunicación Audiovisual y otras normas.

En este contexto, se considera clave contar con la participación de los propios prestadores de los servicios de comunicación, incluyendo programas y vídeos divulgativos sobre cómo se producen los procesos de información, que complementen las iniciativas ya en marcha sobre la verificación de bulos y noticias falsas. Esa divulgación debería orientarse preferentemente, aunque no exclusivamente, hacia la infancia y la juventud, y hacer hincapié en qué puede hacer/no hacer cada persona usuaria en la lucha contra la desinformación.

Y, por supuesto, reivindicar la figura del periodista profesional, como un agente acreditado para asumir una responsabilidad de atender el deber de informar a la ciudadanía desde la ética y el compromiso con la calidad, la veracidad y la diligencia profesional: contraste de las fuentes, separación entre hechos y opiniones, objetividad/imparcialidad, etc. No es ajeno al auge de la desinformación el hecho de haberse “diluido” la figura de ese mediador. Para lograr este objetivo, sería necesario reforzar la vinculación también con sus organizaciones (Colegios Profesionales) en el marco de sus propios sistemas de regulación voluntaria.

