

| GESTIONAR LA REPUTACIÓN EN LA ERA DE LAS *DEEPPFAKES*

I LA ERA DE LA DESCONFIANZA

En un mundo cada vez más digitalizado, las *deepfakes* representan un gran desafío para la integridad y la reputación de las marcas y sus directivos, quienes se han convertido en el objetivo principal de estos ataques en los últimos meses. La habilidad para crear un contenido falso tan convincente desafía nuestra percepción de la realidad y pone en jaque la confianza que tanto cuesta construir.

El futuro de la IA en la gestión de la reputación digital no es solo un desafío tecnológico, sino también una cuestión ética y social que requiere de una respuesta conjunta. Empresas, plataformas sociales, medios de comunicación y organismos reguladores deben colaborar para desarrollar procesos de gestión ágiles ante la desinformación.

¿De qué manera la IA puede ser una aliada en la detección y desmentido de *deepfakes*? ¿Qué medidas pueden tomar las empresas para anticiparse a las *deepfakes*? ¿En qué varía la gestión de estas crisis? ¿Qué papel juegan los medios de comunicación y las plataformas en la gestión de *deepfakes*? ¿Cómo agilizar el proceso de eliminación de las *deepfakes* en buscadores y plataformas? ¿Cómo puede fomentarse la colaboración público-privada para abordar este desafío (Incibe, Guardia Civil, fact checkers, Google, Meta...)

I LÍDERES EN COMUNICACIÓN CORPORATIVA

Expertos en comunicación reflexionan conjuntamente sobre cómo "Gestionar la reputación en la era de las *deepfakes*" y contribuyen a continuar enriqueciendo nuestro espacio de conocimiento.



Beatriz Escriña
Directora Global de
Comunicación en Prosegur



Ricard Fernández Rodríguez
Responsable de
Comunicación Digital en
Endesa



Borja Díaz-Merry
Responsable de
VerificaRTVE



Ana Ortas Pau
Directora de
Comunicación en Prisa



Eduardo Suárez
Jefe Editorial de Reuters Institute
for the Study of Journalism

| ¿AMENAZA O HERRAMIENTA?

En la era digital, la gestión de la reputación se ha convertido en un desafío crucial para las organizaciones. En particular, el ascenso de las *deepfakes* ha añadido una capa adicional de complejidad, llevando la desinformación a niveles sin precedentes. El volumen de contenido falso y manipulado está en constante aumento, lo que plantea serias amenazas para la integridad y la imagen pública de las empresas y sus directivos.

El impacto de la inteligencia artificial (IA) en este fenómeno es innegable. La misma tecnología que impulsa la creación de *deepfakes* también ofrece herramientas para detectar y combatir esta forma de engaño digital. Según Gartner, se estima que para el año 2023 al menos el 30% de las noticias y el video digital producido para el consumo masivo estarán generados por IA y deep learning, lo que subraya la creciente influencia de esta tecnología en la narrativa mediática y la percepción pública.

Esta dualidad de la inteligencia artificial como amenaza y clave para combatir las *deepfakes* representa un paradigma complejo para la gestión de la reputación y la comunicación corporativa. Por un lado, las empresas se enfrentan al desafío constante de proteger su imagen de la difusión de contenido falso y a su impacto reputacional. Este desafío se agrava debido a la proliferación de estos contenidos en los motores de búsqueda y plataformas digitales, donde su impacto se amplifica rápidamente ante una audiencia masiva antes de que se pueda tomar alguna medida correctiva. Por otro, debemos aprovechar las mismas herramientas tecnológicas para identificar y desacreditar eficazmente las manipulaciones que puedan surgir.

VERACIDAD EN TIEMPOS DE DESINFORMACIÓN

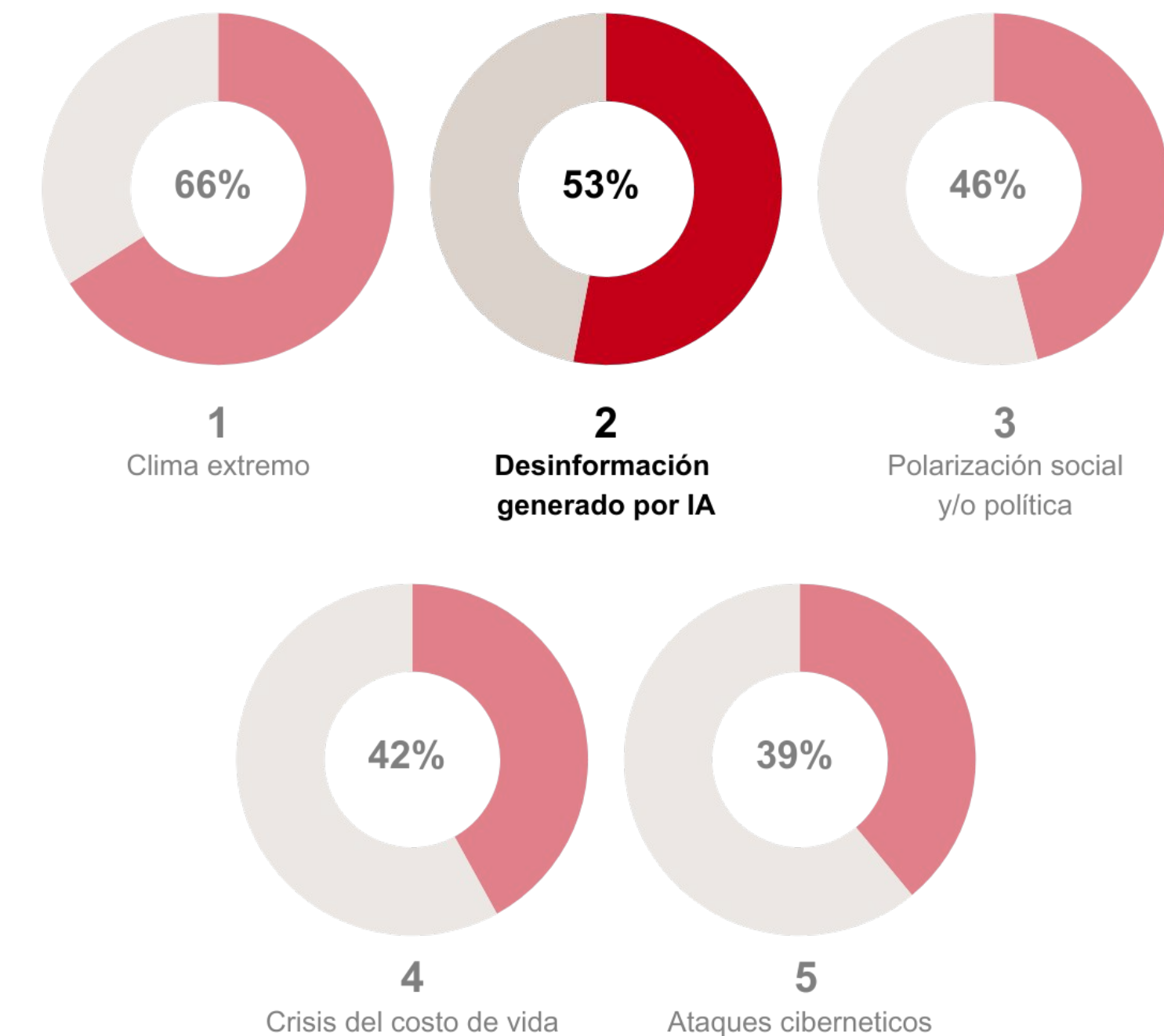
La desinformación y la información errónea emergen como uno de los riesgos globales más graves, según el [Foro Económico Mundial](#). Esta problemática se ha intensificado debido al rápido avance tecnológico, especialmente por las capacidades avanzadas de la inteligencia artificial generativa.

A su vez, no podemos obviar que **el error humano es clave en la propagación de desinformación**. A menudo las personas comparten estos contenidos sin verificar su autenticidad, impulsados por el impacto visual y emocional. Fomentar un enfoque crítico y validar la autenticidad antes de compartir, especialmente en plataformas digitales donde la rapidez supera la pausa para reflexionar, es vital para prevenir tanto la difusión de falsedades como el riesgo de ser engañados por estafas sofisticadas en las que la suplantación de identidad es difícil de identificar para la población general. La alfabetización digital será esencial para ganar la batalla.

Panorama de riesgos actual

Seleccione hasta cinco riesgos que crea que tienen más probabilidades de presentar una crisis material a escala global en 2024.

Fuente, Foro Económico Mundial, Global Risks Report 2024



“

Paradójicamente, la misma tecnología que acelera la creación de información falsa y manipulada se convierte en el antídoto que deberemos usar para combatir su propagación.

I LA INFORMACIÓN FALSA EN AUGE

Según el [Digital News Report 2023](#), publicado por Reuters Institute, en todos los mercados más de la mitad de los usuarios (56%) admiten su preocupación por distinguir qué es verdadero y qué es falso cuando se trata de noticias online. En España, el porcentaje de adultos que utilizan las redes sociales como su principal fuente de noticias si sitúa en el 50% ([Statista](#)) y un 64% afirman estar preocupados por la desinformación.

Este escenario nos obliga a un reflexionar sobre la relación de las personas con las fuentes de información y el papel crítico que cada uno de nosotros juega en discernir la verdad en la era digital. La intersección de una alta preocupación por la desinformación y la prevalencia del uso de redes sociales como principal fuente de noticias, resalta la necesidad urgente de fortalecer las competencias digitales de los ciudadanos, pero este reto no sólo recae en los individuos sino también en todos los actores que deben restaurar y mantener la confianza en el ecosistema informativo.

Proporción de adultos que utilizan las redes sociales como fuente de noticias

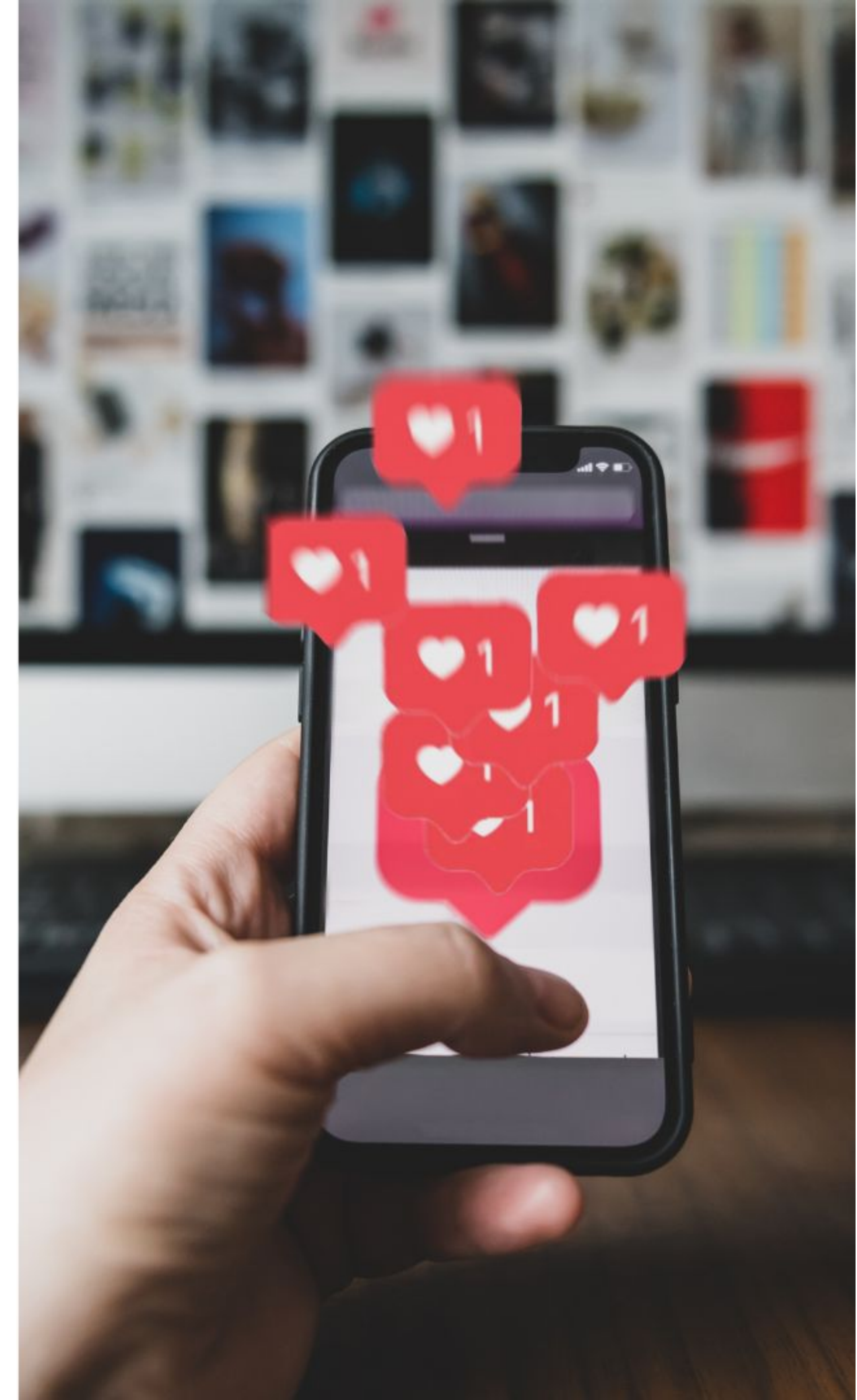
Fuente, Statista, Share of adults who use social media as a source of news in selected countries worldwide as of February 2023



| CONFIANZA FRACTURADA

“Observamos una significativa brecha de confianza entre lo que las personas perciben en plataformas digitales, como las redes sociales, y lo que consumen en los medios tradicionales. Los usuarios confían más en las noticias convencionales que en las provenientes de plataformas, donde la diversidad de emisores, incluyendo creadores de contenido humorístico y satírico, complica la distinción entre lo verdadero y lo falso. Este entorno ha fomentado un escepticismo saludable. Las personas están más alerta y conscientes de la posibilidad de encontrar información falsa.”

**Eduardo Suárez,
Jefe Editorial de Reuters Institute
for the Study of Journalism**



CRECIMIENTO IMPARABLE DE LAS *DEEPFAKES*

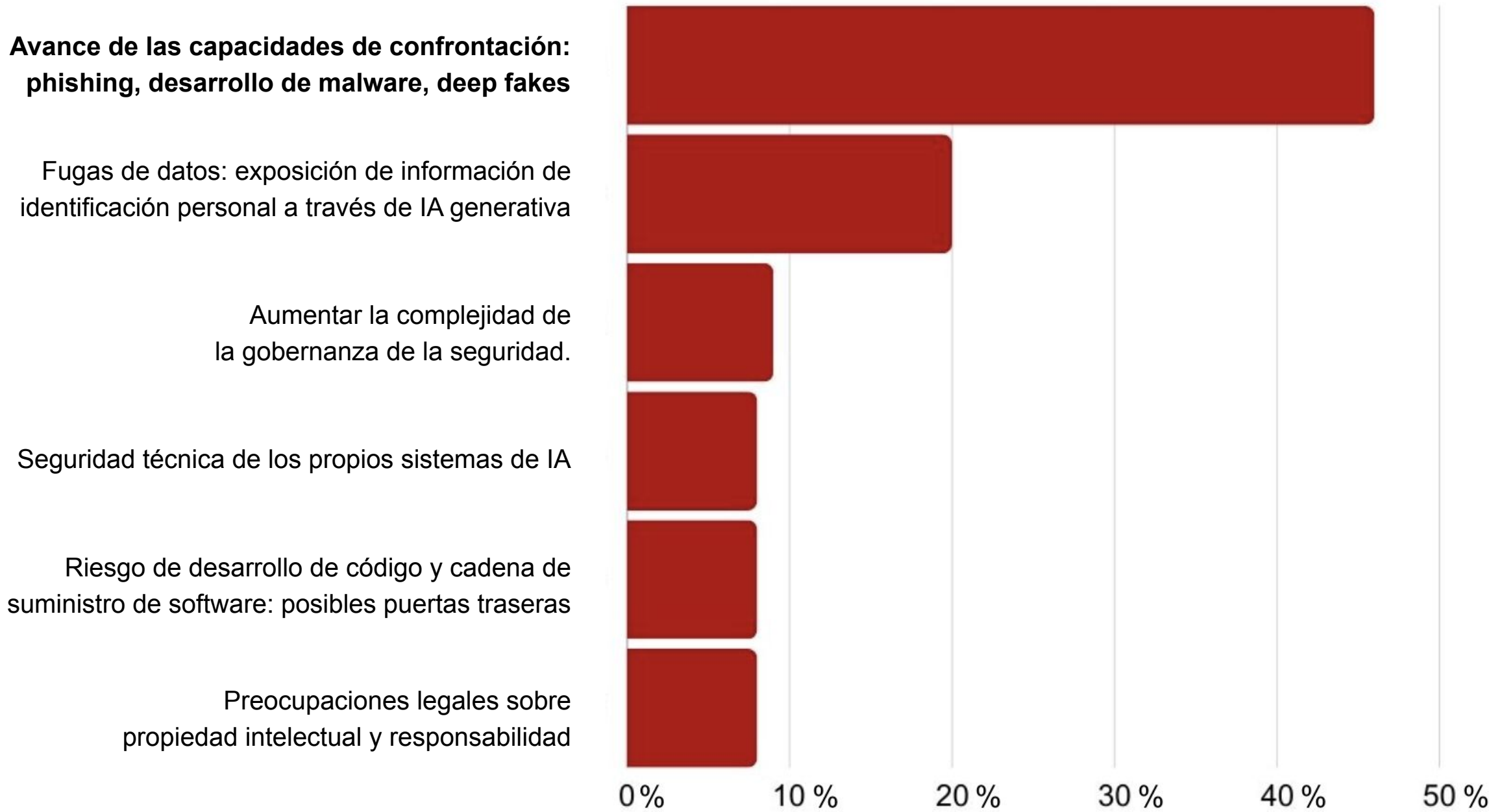
El avance de capacidades adversas, como el phishing, el desarrollo de malware y las *deepfakes*, nuevamente se sitúan como la mayor preocupación de los líderes empresariales con respecto al impacto de la IA generativa en el ámbito de la ciberseguridad, según el Foro Económico Mundial.

El aumento exponencial de las *deepfakes* a escala mundial, cuyo objetivo persigue el chantaje, la manipulación, la extorsión y el daño reputacional, entre otros, desafían la percepción de la realidad y socavan la confianza construida, obligando a las empresas y a sus dircoms a escalar este riesgo en la agenda del CEO.

Gestionar la reputación en este contexto requiere de respuestas tecnológicas, éticas y colaborativas. Para ello, las alianzas entre empresas, organismos reguladores, medios de comunicación y plataformas, será esencial para abordar esta problemática de manera integral y mitigar su impacto.

¿Qué es lo que más le preocupa con respecto al impacto de la IA generativa en la cibernética?

Fuente, World Economic Forum: Accenture, Statista 2024, What are you most concerned about in regards to generative AI's impact on cyber?



I LA FRAGMENTACIÓN DE AUDIENCIAS: UN CALDO DE CULTIVO PARA LA DESINFORMACIÓN

En un mundo donde la forma en que consumimos información se vuelve cada vez más segmentada, las audiencias fragmentadas se han convertido en un terreno muy fértil para la proliferación de la desinformación. Esta situación se agudiza por el hecho de que el **contenido desinformador a menudo se presenta a través de narrativas largamente construidas, atractivas, simples y alineadas con las predisposiciones de los usuarios**. La efectividad de estas narrativas radica en su capacidad para explotar sesgos sociales e ideológicos, permitiéndoles infiltrarse en comunidades específicas con contenido emocionalmente atractivo y que refuerza fuertemente sus creencias.

Este fenómeno se sostiene por unas comunidades altamente motivadas, usuarios afines, trolls y bots, que amplifican la desinformación. Sin duda, y más que nunca, el pensamiento crítico y la verificación son esenciales para combatir la desinformación.

“Sin narrativas previamente consolidadas en comunidades de nicho, las deepfakes aún podrían aparecer, pero su impacto y credibilidad serían diferentes.”

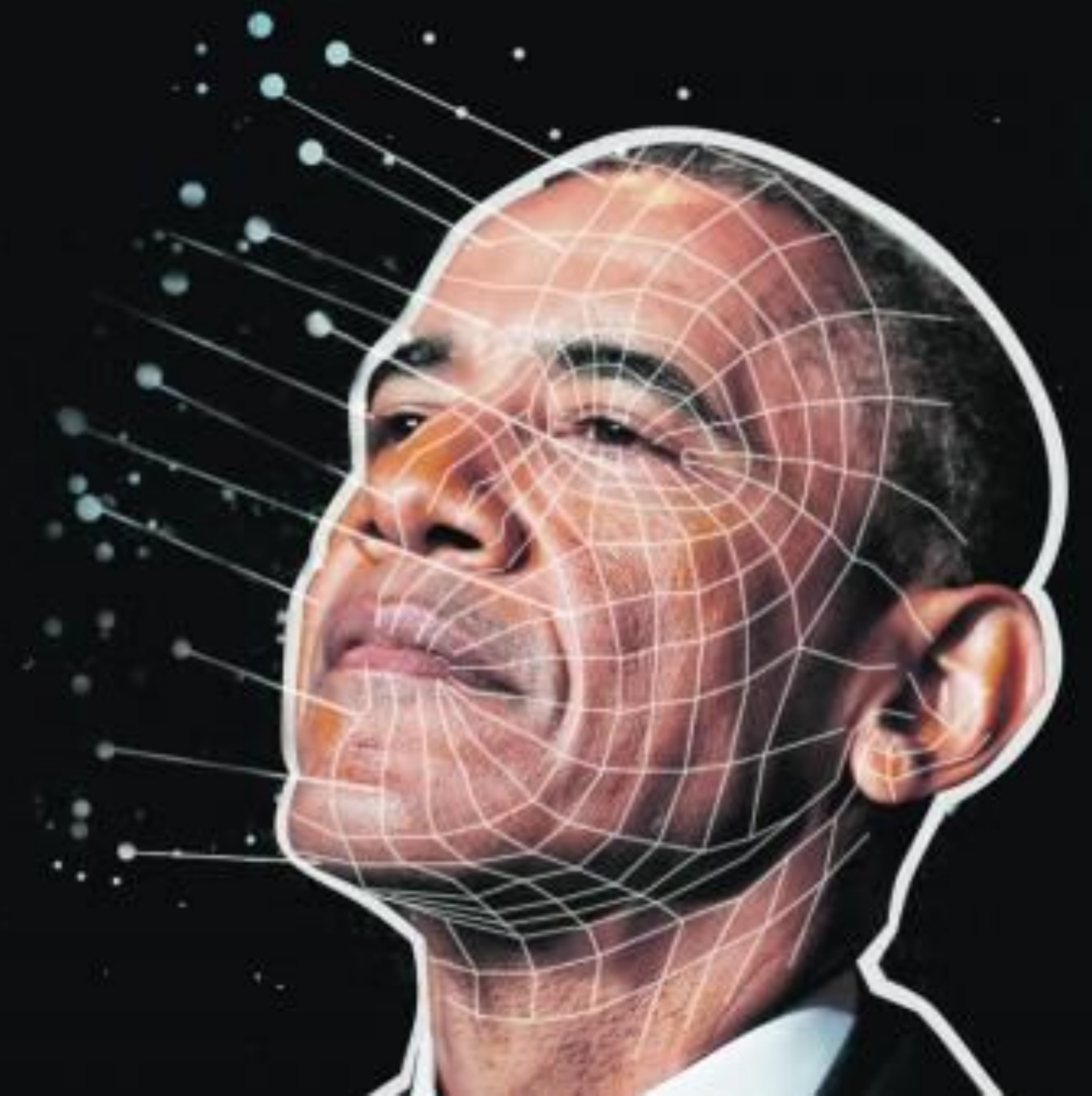
Ricard Fernández Rodríguez,
Responsable de Comunicación Digital
en Endesa

I ¿QUÉ SON LAS *DEEPFAKES*?

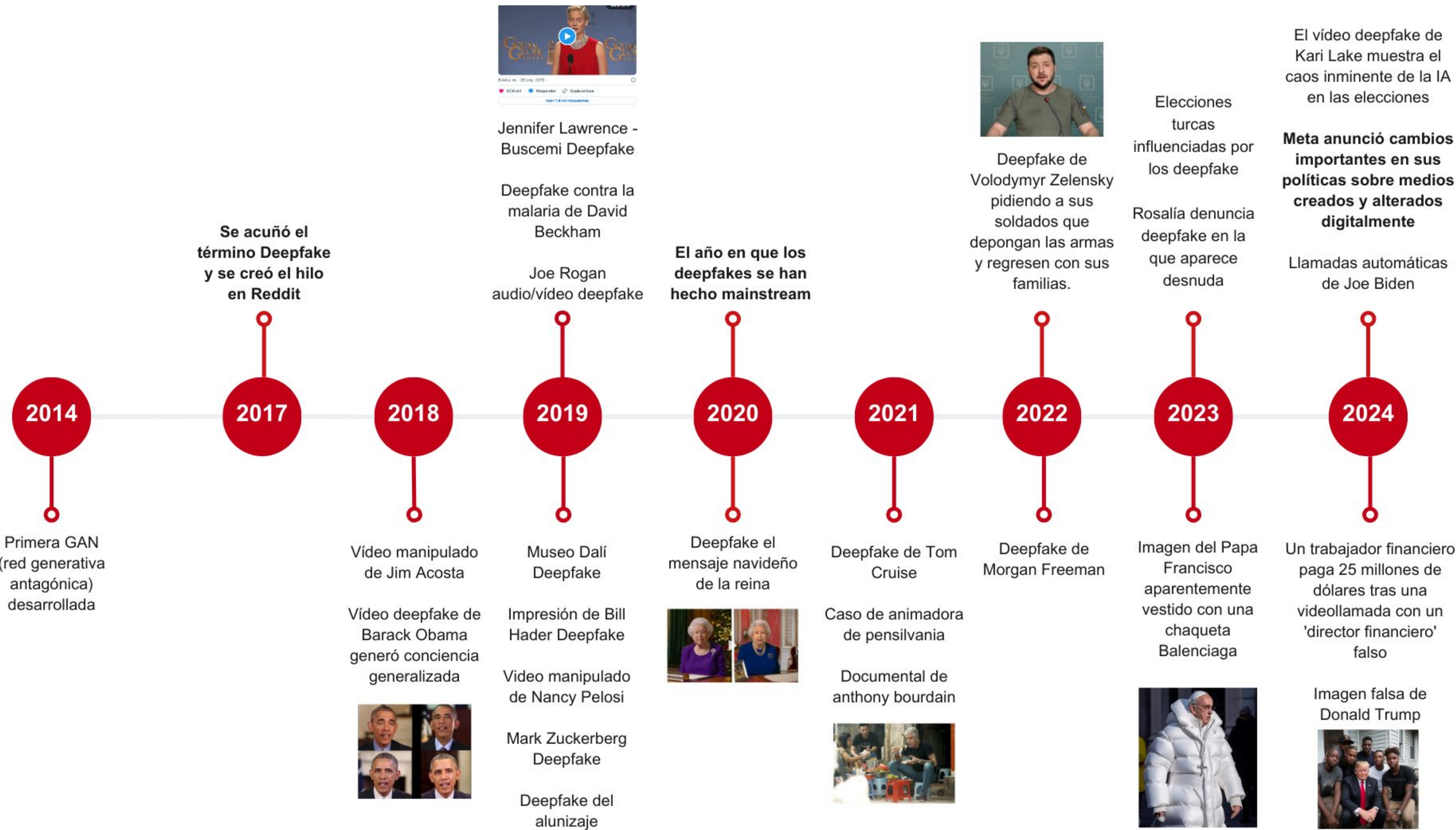
A pesar de que nos encontramos con el término *deepfake* en múltiples noticias, artículos o contenidos, según el estudio de iProov, el 71% de los encuestados a nivel global no sabe qué es un *deepfake* y el 43% admite que no podría detectar uno. Además, el 75% de la muestra en España confirmó que se sienten menos educados sobre los *deepfakes*.

Si tenemos en cuenta que el número total de videos *deepfake* online en 2023 ha aumentado un 550% (2023 State of *Deepfakes*) con respecto a 2019, la sensibilización y formación en la detección de *deepfakes* es una prioridad para la sociedad en general y un escudo reputacional para las compañías.

Pero, ¿qué son las *deepfakes*? Básicamente, son una forma de manipulación digital avanzada que utiliza inteligencia artificial para crear videos, imágenes o audios falsos que parecen auténticos y pueden resultar extremadamente convincentes. Estas tecnologías emplean algoritmos de aprendizaje profundo (de ahí el término "deep") para analizar, modificar y sintetizar contenido multimedia de manera realista.



EVOLUCIÓN DE LAS *DEEPPFAKES*



Fuente: Homeland Security, Increasing Threat of *Deepfake* Identities
Vogue Business, How *deepfakes* could change fashion advertising

UNA AMENAZA QUE NOS ALCANZA A TODOS

“La pregunta no es si nos pasará, sino cuándo ocurrirá. En un mundo donde tanto la información como la desinformación se han democratizado, cualquiera tiene la capacidad de publicar contenidos que impacten directamente en nuestra reputación y negocio. Estamos todos aprendiendo sobre la marcha, enfrentando desafíos inéditos y en tiempo real. La colaboración entre los sectores público y privado es crucial para asimilar las mejores prácticas y enfrentar conjuntamente estos retos. Lo que hoy le sucede a una empresa, mañana podría pasarle a otra.”

Beatriz Escriña,
Directora Global de
Comunicación en Prosegur



LA TECNOLOGÍA DETRÁS DE LAS *DEEPFAKES*

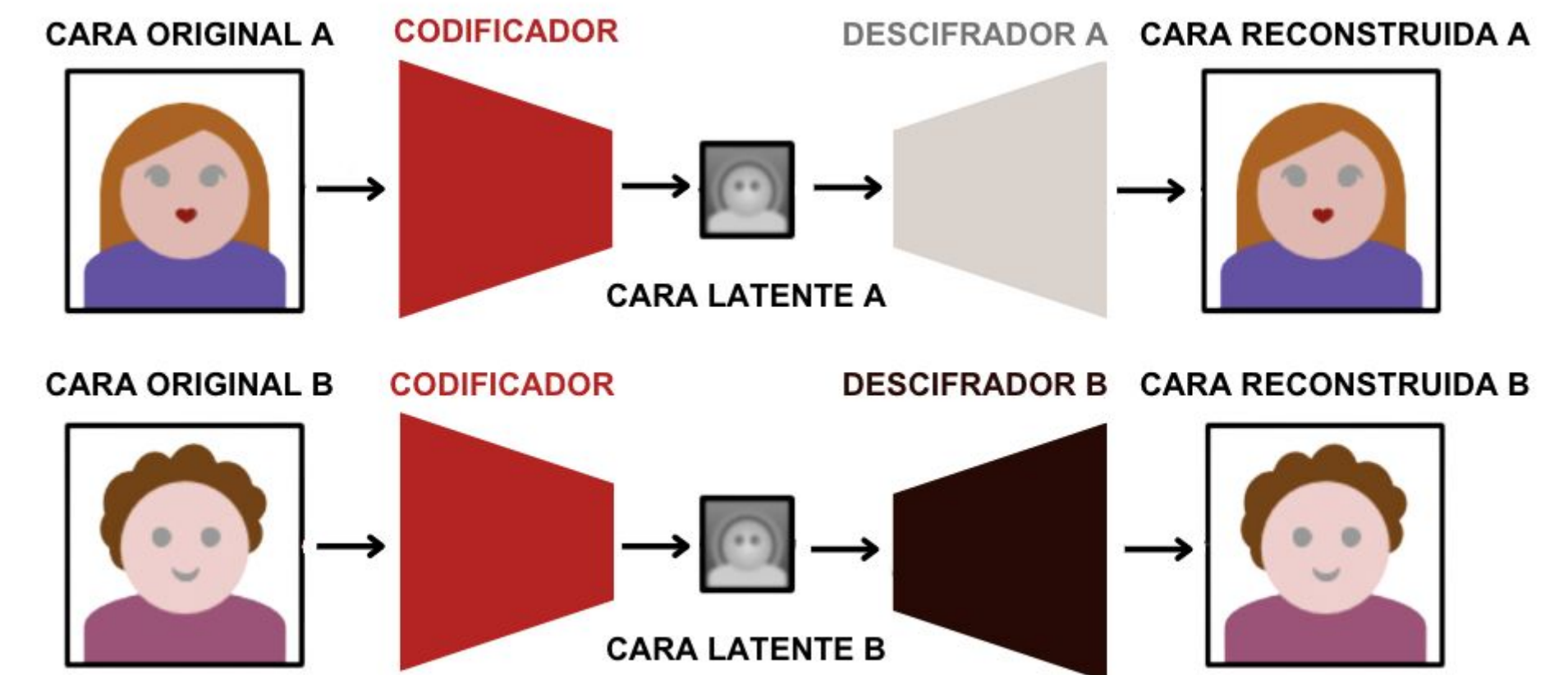
Las *deepfakes* se generan mediante inteligencia artificial (IA) y aprendizaje automático (machine learning), específicamente a través de una técnica conocida como **redes generativas antagónicas (GANs)**, por sus siglas en inglés).

Estas redes están compuestas por dos modelos:

- 1) Genera imágenes o videos falsos (el generador)
- 2) Intenta detectar estas falsificaciones (el discriminador).

A través de un proceso iterativo, el generador aprende a producir resultados cada vez más convincentes, mientras que el discriminador mejora su capacidad para identificar las falsificaciones.

El entrenamiento de estas redes requiere grandes cantidades de datos; por ejemplo, imágenes o clips de vídeo del sujeto que se quiere replicar. A medida que el modelo se entrena, aprende a imitar las expresiones faciales, la voz y los gestos del sujeto, permitiendo la creación de contenidos falsos pero visualmente realistas y difíciles de distinguir de los originales para un consumidor medio.



Fuente: Homeland Security, Increasing Threat of Deepfake Identities

TIPOS DE *DEEPPFAKES*



VÍDEO

Los algoritmos de *deepfake* pueden reemplazar el rostro de una persona (face swapping) en un video por el rostro de otro individuo, haciendo que parezca que la segunda persona está hablando o actuando en el video. Estos videos son conocidos por su alto grado de realismo, con expresiones faciales, movimientos y sincronización labial que se asemejan estrechamente al sujeto original.



IMAGEN

Se alteran o combinan fotos para crear imágenes realistas que parecen originales, pero son completamente ficticias. Estas imágenes pueden ser usadas para manipular percepciones o crear/recrear situaciones falsas.



TEXTO

Se genera o modifica texto, imitando estilos de escritura específicos. Se pueden crear desde emails falsificados hasta publicaciones en redes sociales, presentando riesgos de desinformación y suplantación de identidad.



VOZ

Es posible replicar la voz de una persona analizando sus patrones de habla y utilizando IA para generar nuevas grabaciones de audio con la voz de esa persona. Estas imitaciones tienen un alto grado de precisión y permiten crear grabaciones de audio falsas.



EN VIVO

A través de tecnología avanzada se modifican videos en tiempo real, alterando la apariencia o el discurso de una persona durante transmisiones en tiempo real.

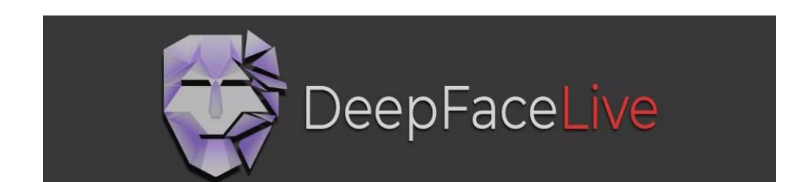
HERRAMIENTAS AL ALCANCE DE TODOS

Un factor clave en el avance de las *deepfakes* es la creciente accesibilidad de herramientas, software y comunidades online de fácil uso para el usuario. Hasta ahora, generar manipulaciones convincentes exigía conocimientos técnicos avanzados. Sin embargo, con la evolución de esta tecnología, han surgido aplicaciones y plataformas que han democratizado su creación.

Estas herramientas, que cuentan con interfaces gráficas intuitivas, han reducido las barreras de entrada para quienes desean experimentar. A su vez, el crecimiento de comunidades online dedicadas a la generación de *deepfakes* ha impulsado la colaboración, el intercambio de conocimientos y el desarrollo de proyectos de código abierto, lo que ha impulsado la propagación de este desafío.

Fuente, mpost.io, Top 10 AI Deepfake Generators for Photo and Video in 2023, 2024

Herramientas para creación de *deepfakes*



Deep Nostalgia

I LA DEMOCRATIZACIÓN DE LAS *DEEPFAKES*

“La evolución desde los primeros deepfake a los actuales es exponencial. Hoy en día cualquiera puede generarlos de manera más sencilla, económica y rápida, por ejemplo, modificando imágenes ligeramente y añadiendo voz sintética. Sin embargo, hay una clara diferencia entre un deepfake producido con IA de baja calidad y uno de alta calidad en diseño gráfico. Mientras que el primero puede ser creado por casi cualquier persona con herramientas accesibles, el segundo requiere muchas horas de estudio y elaboración técnica.”

Borja Díaz-Merry,
Responsable de VerificaRTVE



| SOLUCIONES TECNOLÓGICAS PARA DETECTAR *DEEPFAKES*

Pero, como ya avanzábamos, la IA es también el antídoto para identificar su uso indebido y contrarrestar eficazmente estas manipulaciones digitales. Es indispensable conocer los aliados tecnológicos que tenemos a nuestra disposición para poder integrarlas dentro de la estrategia de defensa de la reputación.

Fuente, Technopedia, 7 Best AI Deepfake Detector Tools For 2024

| Herramientas de detección de *deepfake* de IA



deepware



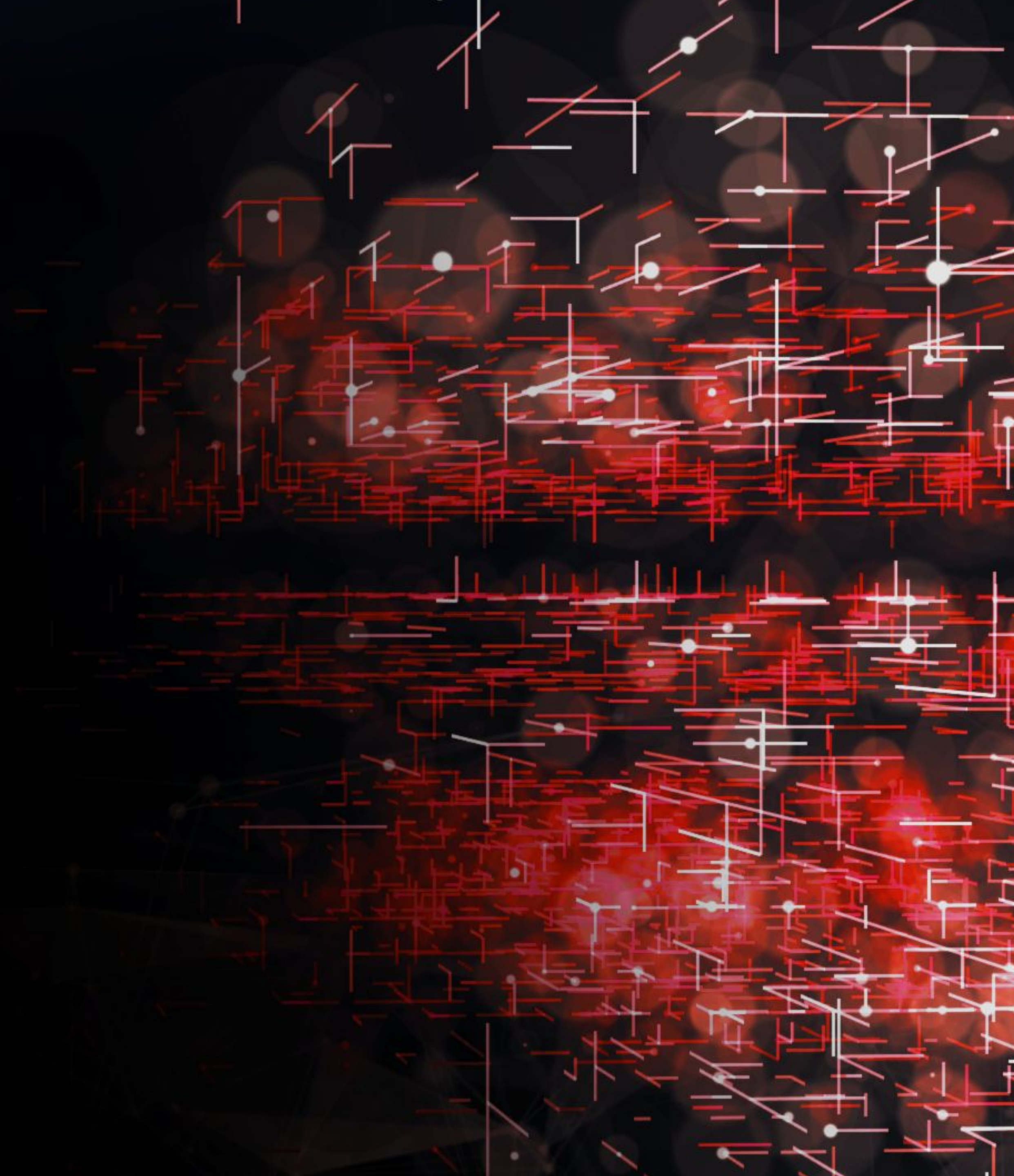
Intel's
FakeCatcher



Microsoft Video
AI Authenticator

Phoneme-Viseme
Mismatch Detector

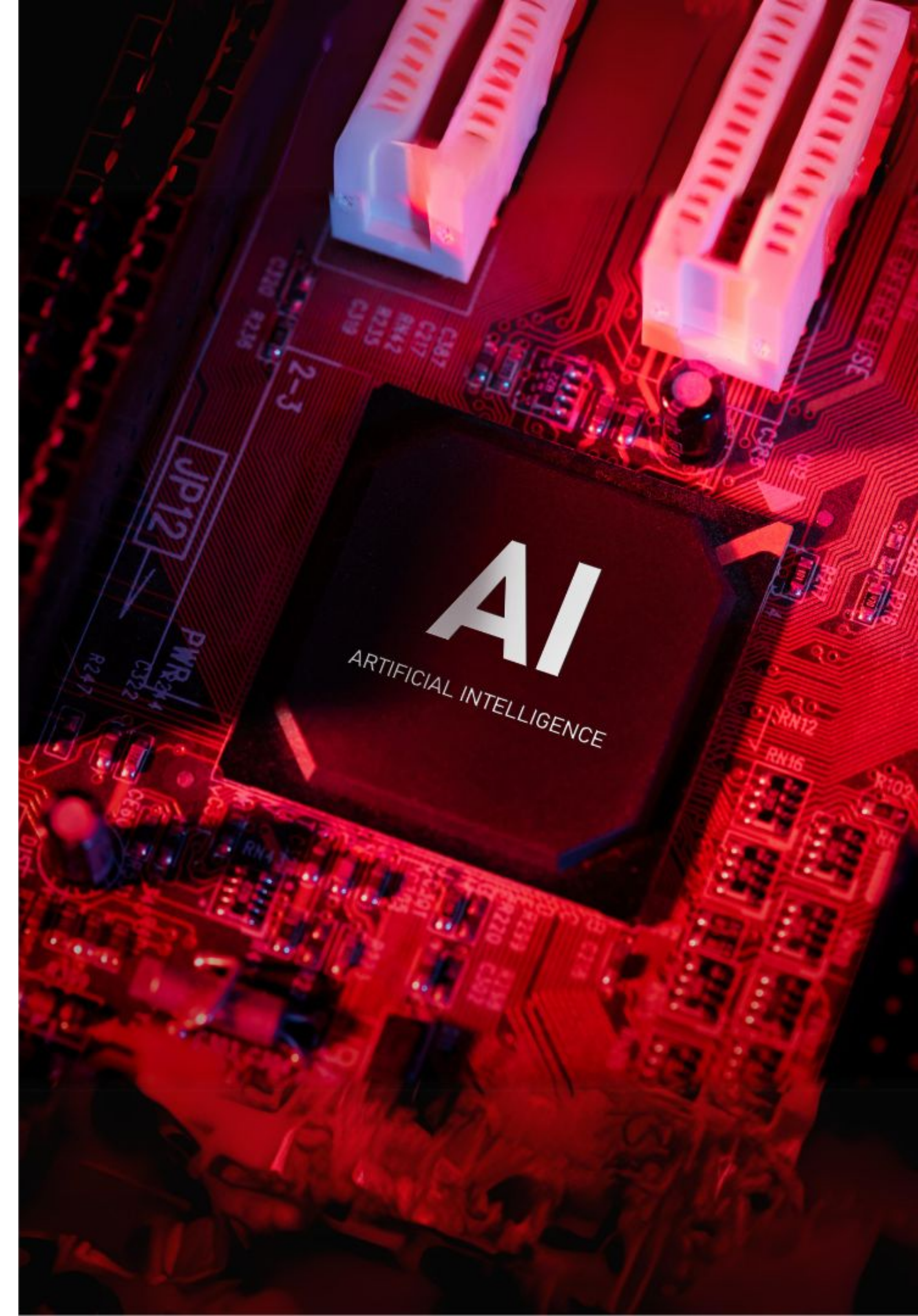
| EL IMPACTO DE LAS *DEEPPFAKES* EN LA REPUTACIÓN CORPORATIVA



IMPACTO DE LAS *DEEPPFAKES* EN LA REPUTACIÓN CORPORATIVA

En la actualidad, el 55% de los líderes encargados de la reputación de marca creen que la IA generativa (GenAI) representará un riesgo significativo para sus organizaciones, sin embargo, tan sólo el 21% de las organizaciones cuentan con planes y protocolos específicos sobre cómo evitar usos que podrían causar riesgos reputacionales, tal y como indica Gartner.

La reputación, la confianza del consumidor y la credibilidad de una marca son activos vitales, pero el impacto de las *deepfakes* amenazan con socavar estos fundamentos. En un contexto de disrupción tecnológica como el que nos encontramos, es vital redefinir el mapa de riesgos corporativo para mejorar la resiliencia organizativa y la posición competitiva. Si no invertimos en nuevas tecnologías, reforzamos nuestras capacidades en gestión de riesgos “cyber”, somos más ágiles y coordinamos esfuerzos internos para que la gestión de estas amenazas sean responsabilidad de todos, lo tenemos complicado para salir bien parados.



¿A QUÉ NOS ENFRENTAMOS?



Un empleado de una compañía financiera hongkonesa que transfirió 24 millones de euros a lo que él creía que era la filial de su empresa en el Reino Unido.

Los estafadores usaron la tecnología *deepfake* para hacerse pasar por el director financiero y otros directivos en una videoconferencia, resultando que eran todos falsos.



Un responsable de esta compañía realizó transferencias a una cuenta bancaria por más de nueve millones de euros creyendo que lo hacía por orden de su jefe, cuando en realidad se trataba de un ciberdelincuente que había suplantado su identidad.

“Un tuit falso consiguió que el Dow Jones cayese 143 puntos. Los *deepfakes* pueden multiplicarlo.”

¿A QUÉ NOS ENFRENTAMOS?



En este *deepfake* desmentido por [Verifica RTVE](#), se suplantán a las periodistas de la corporación Ana Blanco y Almudena Guerrero, al gobernador del Banco de España, Pablo Hernández de Cos, y a la presidenta del grupo Inditex, Marta Ortega, para promocionar de forma fraudulenta un proyecto financiero.



“Dio la noticia a todo el mundo sin pensar en las consecuencias. ¿Qué le espera a Carlos Herrera después de lo que ha dicho?”, dice un mensaje etiquetado como promocionado en la red social X. El [texto](#) adjunta un enlace a una web que se presenta como si fuera la página de RTVE que recoge una supuesta entrevista del presentador David Broncano al periodista Carlos Herrera en la que promueven una plataforma de criptomonedas llamada “Immediate Bitwave” que puede convertir “250 euros en un millón en tres o cuatro meses”.



Un deepfake que se presenta como un fragmento de una emisión del telediario de La 1, destaca a figuras públicas —como Marta Ortega o el gobernador del Banco de España— promocionando las excelencias de un presunto proyecto financiero revolucionario liderado por el presidente de Inditex y que ofrece enormes ganancias a todos sus usuarios.

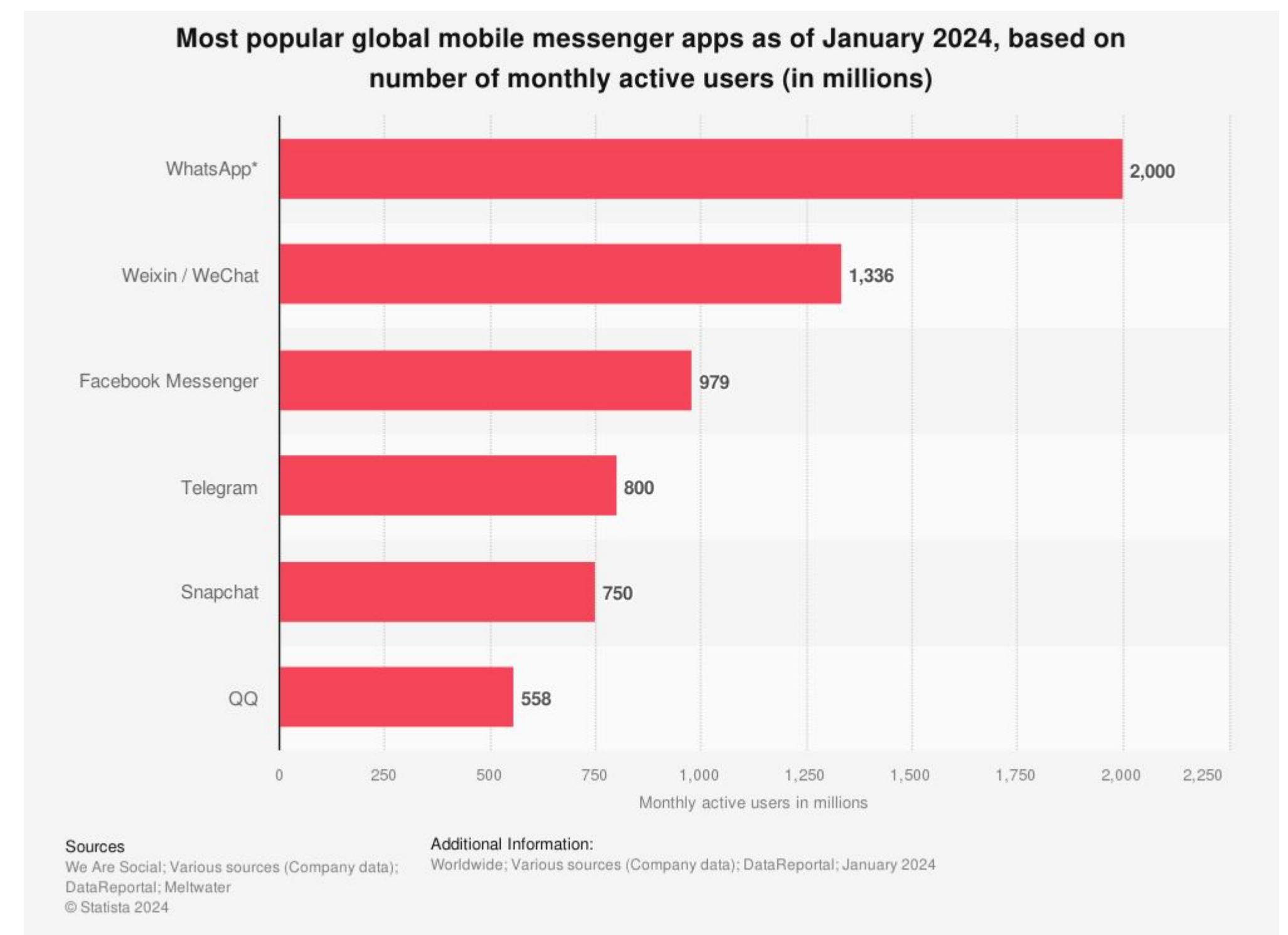
I LA PROPAGACIÓN EN “DARK”

El término "dark social" se refiere al contenido compartido a través de canales privados, como mensajes directos, correos electrónicos y aplicaciones de mensajería. El 84% de todos los contenidos de la web y de las redes sociales son compartidos por esta vía, así que más que nunca no podemos perderlo de vista.

Detrás de Facebook y YouTube, las plataformas con mayor número de usuarios mensuales activos son tres servicios de mensajería (WhatsApp, Facebook Messenger y WeChat), según el informe «Digital 2023 Global Overview Report» de We Are Social.

La distribución de *deepfakes* a través de dark social escapa a las herramientas de monitorización y verificación pública. En estas plataformas privadas, el contenido puede viralizarse rápidamente y sin un control efectivo, lo que hace más difícil rastrear su origen y contrarrestar la información falsa, amplificando así el potencial daño que pueden causar las *deepfakes*.

De nuevo, no hay soluciones mágicas, pero fomentar la educación digital entre los usuarios puede ayudar a mitigar el impacto de estos contenidos.



I DESAFÍOS EN LA ERA DIGITAL

“Hay tres tendencias que están magnificando el problema de la desinformación. Primero, la producción de contenido mediante inteligencia artificial y herramientas emergentes se ha vuelto más accesible, complicando aún más el control. Segundo, las redes sociales son cada vez menos transparentes, dificultando la detección de contenido viral. Por último, las plataformas tienen pocos incentivos legales para combatir la desinformación, lo que dificulta el cambio hacia un entorno más responsable.”

Eduardo Suárez,
Jefe Editorial de Reuters Institute
for the Study of Journalism



I LOS DIRECTIVOS EN EL OJO DEL HURACÁN

¿Te imaginas que te levantas un día y te encuentras con un *deepfake* en YouTube que involucra a tu CEO en la promoción de una estafa? Esta situación la vivió [Brad Garlinghouse](#), CEO de Ripple, una plataforma de código abierto que permite transacciones rápidas con un bajo coste, hace ya tres años. En el video, un suplantado Brad, invita a los espectadores a participar en un evento de XRP donde deben enviar tokens a una "dirección designada".

A cambio, el personaje del video promete devolver "el doble" como parte del falso sorteo.

¿Cuáles fueron las consecuencias?

- La asociación del CEO con una estafa tuvo un impacto directo en la confianza de sus clientes.
- A su vez, los usuarios de Reddit, descubridores de este contenido engañoso, confirmaron que Google no retiró inmediatamente el contenido porque no contradecía sus políticas, lo que permitió que consumidores siguieran siendo estafados y generó preocupaciones sobre la eficacia de las políticas de la plataforma para lidiar con *deepfakes* y contenidos fraudulentos.
- Transcurrieron 6 meses hasta la decisión judicial a favor de la empresa en su batalla con la Comisión de Valores y Bolsa (SEC).

Está claro que nos encontramos en un nuevo contexto de gestión de la reputación de nuestros líderes y nuestras compañías, y es vital que sepamos a qué nos enfrentamos. Vayamos paso a paso.

“

Es vital colaborar con plataformas de redes sociales y empresas de tecnología para mejorar las capacidades de rastreo y verificación de contenidos compartidos a través de dark social.



PRINCIPALES CARACTERÍSTICAS DE UNA CRISIS POR “*DEEPFAKE*”

Inmediatez_

La rapidez con la que se expande es muy alta. Gran visibilidad en redes y un coste operativo y reputacional muy alto, más cuando la credibilidad de los máximos directivos de la empresa está en juego.

Tecnicismo_

Nivel alto de especialización y expertise que exige la alfabetización de todos los equipos implicados en su resolución.

Disrupción_

Los ataques son cada vez más sofisticados y ágiles, respondiendo cada vez más a una estrategia clara de negocio. La IA utilizada como antídoto tiene que iterar al ritmo en el que avanza su mal uso.

Impacto transversal_

El impacto afecta de manera simultánea a un amplio número de stakeholders, algunos aún por identificar en función de cada tipología, y en canales que no podemos monitorizar (whatsapp, email, etc.)

Incertidumbre x1000_

Se gestiona la comunicación con un alto nivel de incertidumbre, sin disponer de todos los datos y sin poder dimensionar el alcance de la crisis. Y cuando se trata de una suplantación de identidad de un directivo, se añade la tensión adicional por el impacto de su marca personal.

Sin fin_

El alcance se multiplica y cuando parece que se ha controlado, salta a otra plataforma y revive la desinformación. El seguimiento continuo y la detección precoz son nuestras mejores armas, además del diálogo proactivo con todos los agentes para lograr una solución lo más rápida posible.

EL RIESGO DE DELEGAR DECISIONES A IA

“Si vamos dejando a la inteligencia artificial tomar decisiones cada vez más complejas, el problema puede ser muy grande si no tenemos un protocolo de actuación claro. Una inteligencia que reemplace a humanos puede parecer tentadora para las empresas, pero ¿qué hacemos si la empresa se equivoca? ¿Cómo gestionamos las decisiones erróneas? Si dejamos que esta inteligencia tome el control sin desarrollar un protocolo interno para supervisar y evitar errores, podríamos enfrentarnos a graves consecuencias tanto para la reputación de la empresa como para la toma de decisiones en general.”

Ana Ortas Pau,
Directora de Comunicación en Prisa



LOS RETOS DEL DIRCOM ANTE LA GESTIÓN DE *DEEPPFAKES*



Impulsar la estrategia gestión de la desinformación en la agenda del Dircom



Sensibilizar a la alta dirección e integrar los riesgos y oportunidades de la IA dentro de la gestión reputacional de la compañía.



Fomentar una gestión proactiva del riesgo reputacional



Dar solución al déficit de formación y competencias de los profesionales de comunicación



Educar, formar e impulsar un enfoque ético del uso de la IA como parte de la cultura corporativa

¿CÓMO GESTIONAR LAS *DEEPPFAKES*?

GESTIÓN PROACTIVA:

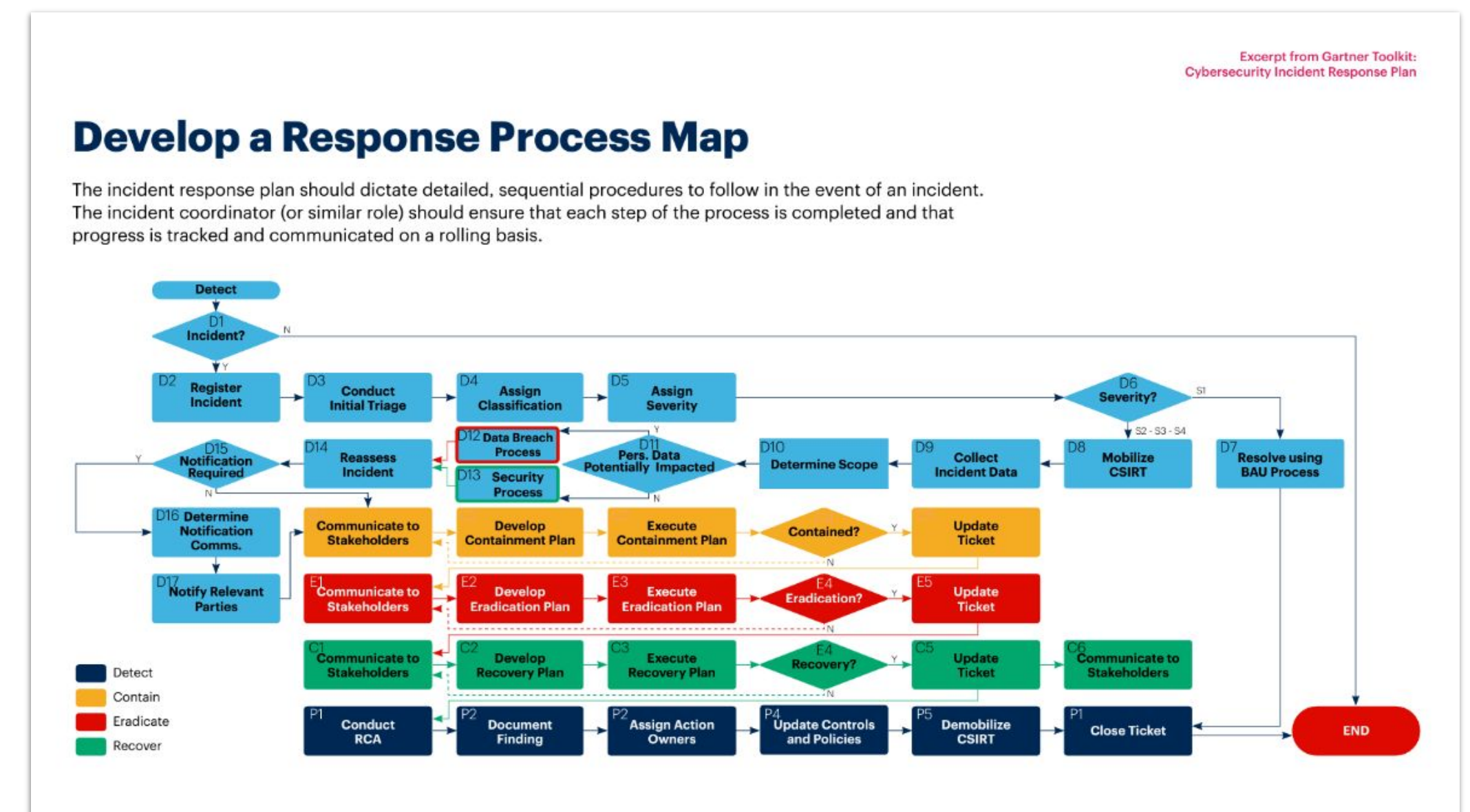
Monitorización Constante. Mantener una vigilancia constante sobre las plataformas digitales y los medios de comunicación para detectar rápidamente cualquier indicio de *deepfake*. Estar al día de las últimas tendencias tecnológicas para adaptar la estrategia de gestión de reputación ante futuras amenazas.

Evaluación preventiva del Impacto Evaluar de manera anticipada el potencial alcance de las *deepfakes*, identificando los públicos afectados.

Desarrollo de Políticas y Protocolos: Establecer guías específicos para prevenir y gestionar crisis relacionadas con *deepfakes*. Esta fase puede incluir la implementación de tecnologías de detección, actualización de políticas de seguridad digital y designación de un equipo de crisis especializado.

Desarrollo de simulacros realistas. Definir escenarios ficticios de *deepfakes* para probar la eficacia de los planes de respuesta y asegurarse de que los equipos están preparados para actuar rápidamente y de manera coordinada.

Activación del Plan de Contingencia. Ante la detección de un *deepfake*, es necesario activar de inmediato el plan de crisis específico para minimizar el daño reputacional. Respuesta en canales propios sobre el suceso para informar al público de manera transparente y precisa sobre la situación.



I ¿CÓMO GESTIONAR LAS *DEEPPFAKES*?

GESTIÓN CONTINUA:

Transparencia y Comunicación Clara. Mantener una comunicación transparente con el público y los stakeholders, proporcionando información verificada y asegurando que se están tomando medidas para abordar la crisis de manera efectiva.

Embajadores: Contar con embajadores internos y externos es fundamental gracias a su capacidad para fortalecer la comunicación y la credibilidad de la organización en un momento de crisis. Los embajadores pueden ayudar a diseminar información precisa y verificada, a desmentir falsedades y validar la información ante el público y los medios. Juntos, pueden contribuir a minimizar el daño y restaurar la confianza en la organización.

Educación y Sensibilización: Liderar iniciativas internas y externas para potenciar la educación y sensibilización sobre el fenómeno de las *deepfakes*, reforzando las capacidades de los stakeholders en la identificación de contenido falso y en las mejores prácticas para proteger la reputación de la empresa.

Alianzas externas: Desarrollar relaciones sólidas con expertos en seguridad cibernética, agencias gubernamentales, buscadores, plataformas, entidades de fact checking, organizaciones de consumidores y otras organizaciones relevantes para compartir conocimientos y generar acuerdos sobre cómo abordar esta problemática.

**| CONSTRUIR
ALIANZAS PARA
ACABAR CON LA
DESINFORMACIÓN**



ALIANZAS PÚBLICO-PRIVADAS

La colaboración entre empresas, plataformas, buscadores, organismos reguladores, medios de comunicación y otros agentes involucrados, es fundamental para combatir las desinformación de manera efectiva.

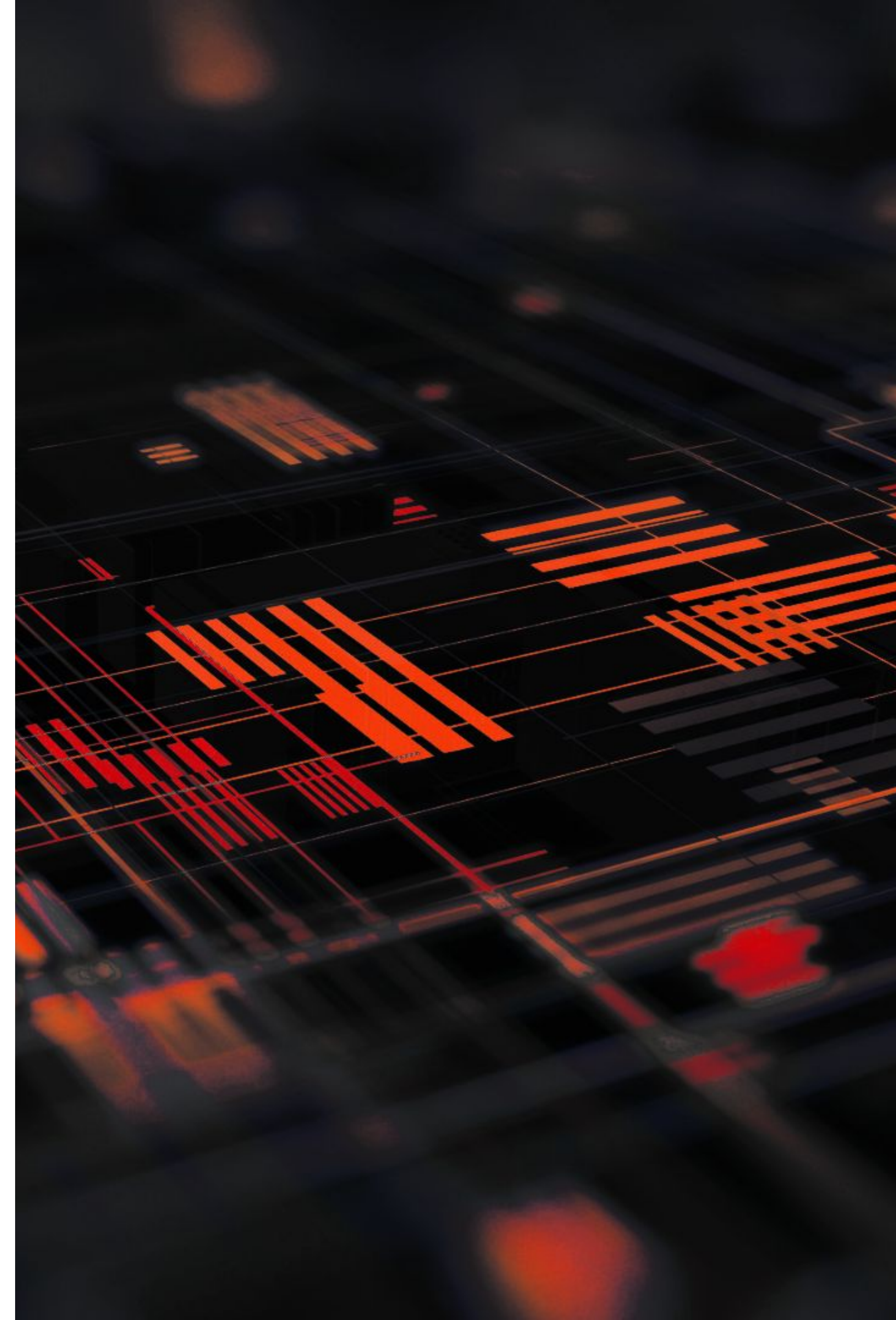
Estas alianzas nos deben ayudar a poder dar una respuesta más coordinada y potente desde todos los ámbitos, teniendo en cuenta que será clave encontrar el incentivo común para que todos estemos alineados y comprometidos para construir en una misma dirección. Además, trabajar juntos ayudará a establecer estándares y políticas más robustas que pueden aplicarse a nivel global, asegurando una lucha más uniforme y continua contra la desinformación. Este enfoque colaborativo no solo mejora la capacidad de respuesta ante incidentes de desinformación, sino que también fomenta la creación de una sociedad más informada y resiliente.



| SENSIBILIZAR Y CONCIENCIAR

“Es fundamental sensibilizar y concienciar a los comités de dirección y a los órganos de gobierno de las empresas sobre el verdadero impacto que las deepfakes pueden tener en su reputación. Aunque es un término ya conocido, su concepto sigue siendo abstracto para muchos, y existen quienes creen que no les afectará. Resulta esencial que quienes poseen capacidad de decisión dentro de las organizaciones comprendan el papel de las plataformas digitales, los medios de comunicación, los organismos de verificación y la propia empresa, para impulsar alianzas estratégicas entre el sector público y el privado.”

**Beatriz Escriña,
Directora Global de
Comunicación en Prosegur**



¿QUÉ DEMANDAN LOS DIRCOMS?

1 ELIMINACIÓN RÁPIDA DEL CONTENIDO FRAUDULENTO

Las redes sociales y los motores de búsqueda deberían eliminar de manera inmediata el contenido fraudulento relacionado con las *deepfakes*.

2 IMPLEMENTACIÓN DE POLÍTICAS CONTRA LA DESINFORMACIÓN

Se deben reforzar las políticas contra la desinformación y los *deepfakes*, estableciendo mecanismos más rigurosos para detectar y eliminar contenido falso. Estas medidas son esenciales para proteger la integridad de la información online y prevenir la propagación de estafas y fraudes.

3 AGILIDAD PARA IDENTIFICAR EL ORIGEN DEL CONTENIDO

Las plataformas deberían ayudar en la identificación de los responsables de crear y difundir el *deepfake* y prevenir futuros ataques.

4 COORDINACIÓN CON AUTORIDADES REGULADORAS

Es clave la coordinación y el diálogo entre los distintos agentes para abordar de manera efectiva el problema de los *deepfakes* y otras formas de desinformación online.

5 MAYOR TRANSPARENCIA Y RESPONSABILIDAD

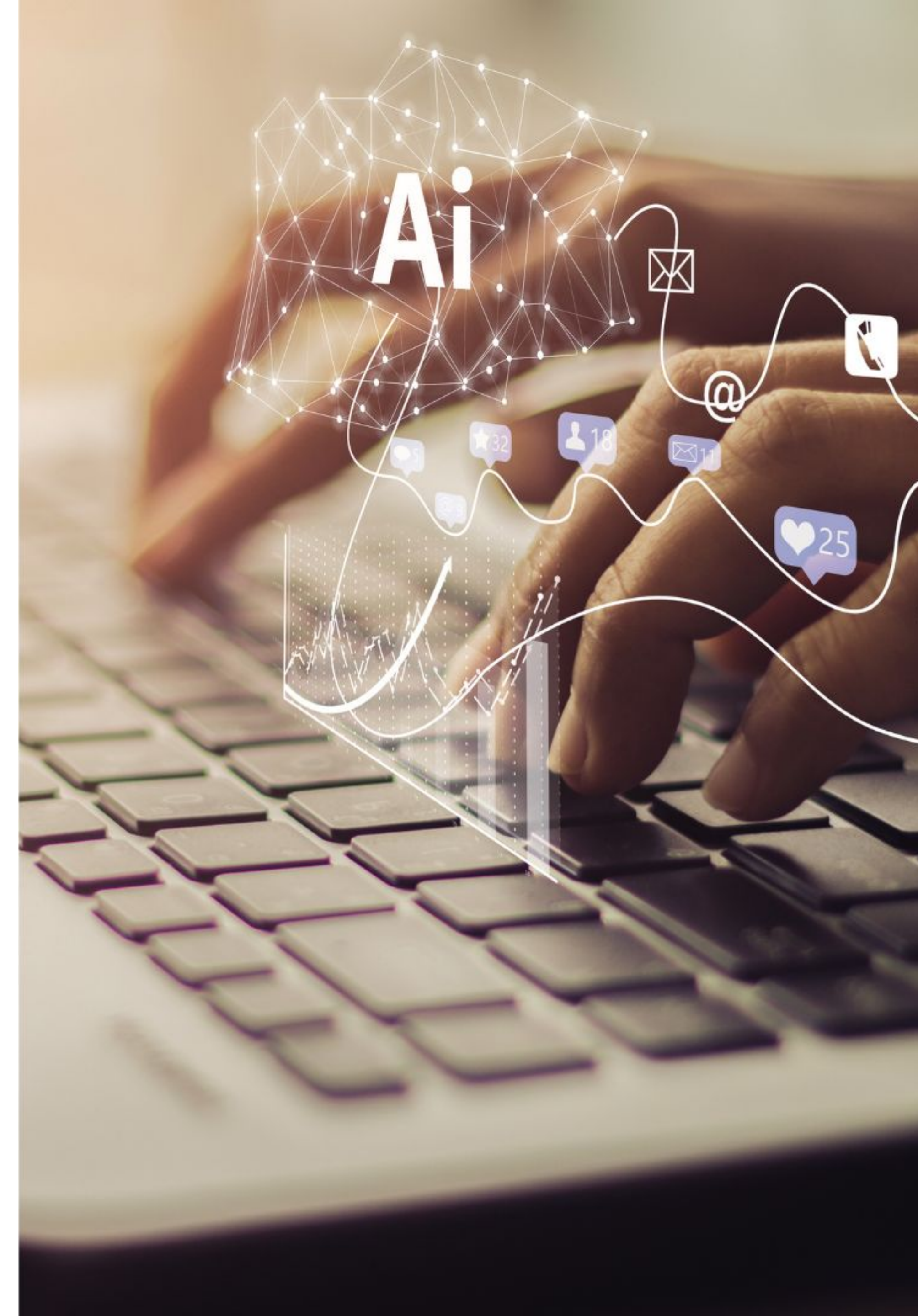
Es necesaria una mayor transparencia y responsabilidad por parte de las plataformas y buscadores en cuanto a sus políticas de moderación de contenido y acciones tomadas.

GESTIONAR LA REPUTACIÓN EN LA ERA DE LAS *DEEPPFAKES*

I NIVELANDO EL CAMPO DE JUEGO

“La regulación es fundamental, sobre todo porque ante un efecto positivo o dañino de una acción en la que está implicada la IA debería haber responsabilidad compartida y equitativa tanto de las empresas como de las plataformas de IA.”

Ana Ortas Pau,
Directora de Comunicación
en Prisa



¿QUÉ ARGUMENTAN LOS MEDIOS SOCIALES?

1

NORMAS Y POLÍTICAS INTERNAS

Las plataformas operan bajo normas y políticas internas establecidas para moderar el contenido. Estas normas están diseñadas para garantizar un equilibrio entre la libertad de expresión y la protección contra el discurso dañino, y que cualquier decisión de eliminación se basa en la aplicación coherente de estas

2

PROCESO DE REVISIÓN Y VERIFICACIÓN

Para eliminar cualquier contenido se sigue un proceso riguroso de revisión y verificación. Esto implica la necesidad de confirmar la veracidad de la demanda para eliminar un contenido y el crecimiento del volumen de implica crecer en su estructura para garantizar este factchecking.

3

IMPACTO EN LA CONFIANZA DEL USUARIO

Eliminar contenido de manera rápida puede generar preocupaciones sobre la censura y la manipulación de la información entre los usuarios. Las plataformas pueden argumentar que es importante mantener la confianza de los usuarios al permitir un proceso de revisión transparente y justo antes de tomar medidas drásticas como la eliminación de contenido.

4

NECESIDAD DE CONTEXTUALIZACIÓN

A veces, el contenido controvertido puede ser parte de un debate público más amplio o tener un contexto importante que no es evidente a primera vista.

5

MAYOR TRANSPARENCIA Y RESPONSABILIDAD

Plataformas y buscadores consideran que sus políticas de moderación de contenido y acciones tomadas son transparentes, pero usuarios y marcas reclaman mayor claridad y responsabilidad.

I UN INCENTIVO COMÚN

“Encontrar un incentivo común para todos es clave para hacer frente a las deepfakes. Así, cada uno (plataformas incluidas) tendría una razón para compartir buenas prácticas, intercambiar información y datos, garantizando la privacidad de los usuarios. La transparencia y colaboración son fundamentales para gestionar los contenidos maliciosos de manera efectiva.”

**Ricard Fernández Rodríguez,
Responsable de Comunicación Digital
en Endesa**



¿QUÉ ROL JUEGA EL FACT-CHECKING?

1

REFUERZO DE LA CREDIBILIDAD

En un entorno saturado de información falsa, los verificadores se convierten en los aliados de la defensa de la verdad y de la precisión informativa, ganando la confianza de las audiencias.

2

PREVENCIÓN DE LA DESINFORMACIÓN

El fact-checking actúa como un filtro esencial para prevenir la propagación de desinformación. Al verificar la autenticidad de las informaciones antes de compartirlas, se reduce significativamente el riesgo de que narrativas falsas ganen tracción y se conviertan en creencias erróneas ampliamente aceptadas.

3

EDUCACIÓN Y CONCIENCIA PÚBLICA

Al proporcionar herramientas y metodologías de fact-checking, se educa al público sobre cómo identificar información dudosa. Esta educación es crucial para desarrollar un pensamiento crítico sólido entre los ciudadanos, ayudándoles a cuestionar y analizar críticamente la información que reciben.

4

PROTECCIÓN “REAL TIME” CONTRA LA MANIPULACIÓN

Estar protegidos 24x7 será la norma. Gracias a la combinación de técnicas de verificación como el análisis de metadatos, el uso de la inteligencia artificial para reconocer patrones de falsificación o la colaboración con expertos técnicos, son capaces de ofrecer una respuesta integral.

5

FOMENTO DE UN DIÁLOGO CONSTRUCTIVO

El fact-checking fomenta un diálogo más informado y constructivo en la esfera pública. Al asegurar que las discusiones se basen en hechos verificados, se promueve un debate más sano y productivo, esencial para la toma de decisiones informadas en una sociedad democrática.

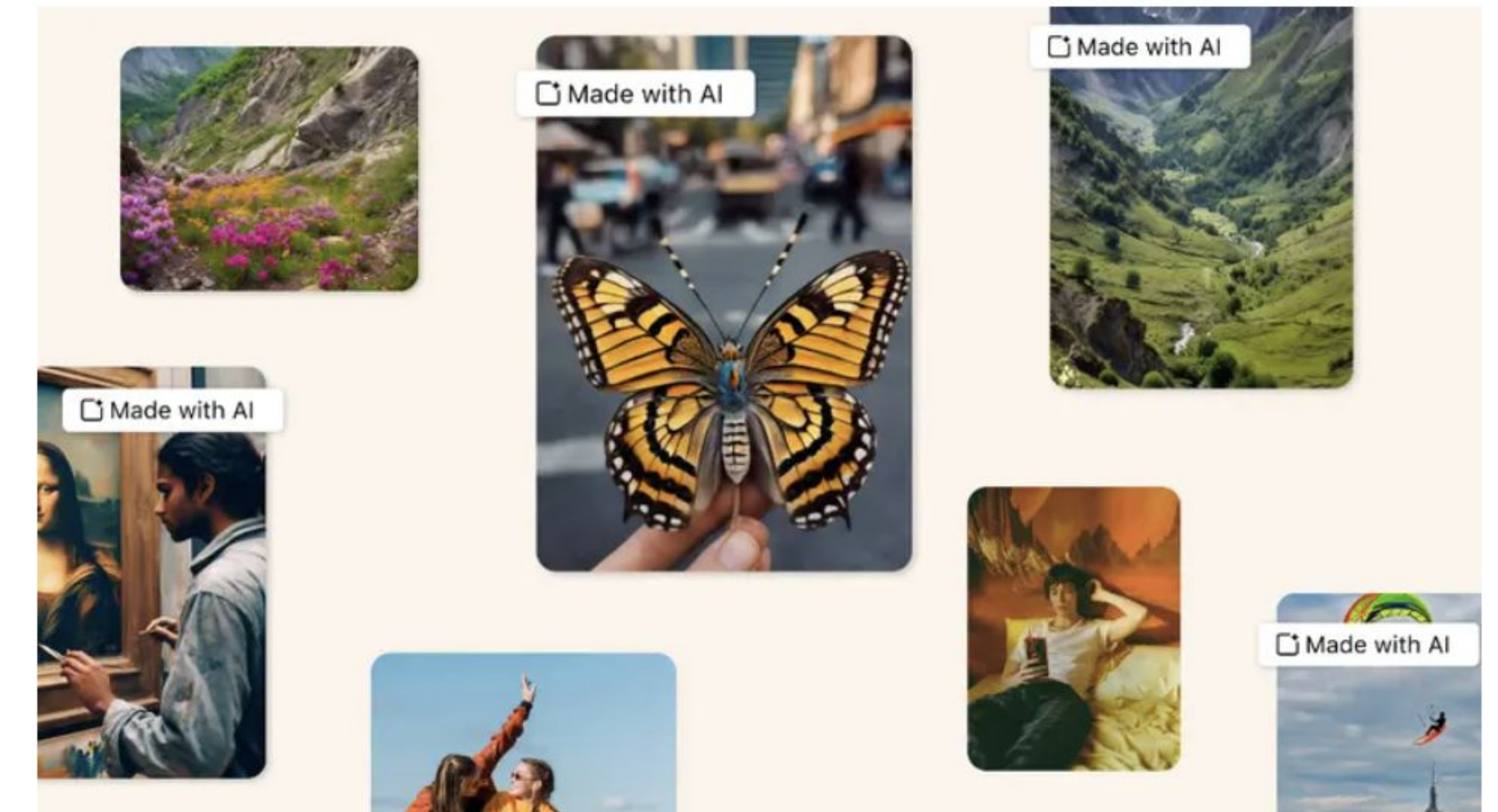
I ETIQUETAR CONTENIDO FALSO

“No se trata de censura, sino de garantizar que los vídeos falsos no circulen sin ningún tipo de advertencia en las plataformas. Es fundamental que se tomen medidas, cualesquiera que sean, para evitar la propagación de desinformación sin una notificación que señale claramente su falsedad.”

**Borja Díaz-Merry,
Responsable de VerificaRTVE**

Meta comenzará a etiquetar contenido creado con IA

Tanto en Facebook como Instagram y Threads se empezará a ver la marca tanto en imágenes como en vídeo y hasta en audios.



▲ Así se comenzará a ver el contenido en las redes de Meta Meta / Meta

Conclusiones

Identificación de riesgos y oportunidades de la IA

La IA generativa presenta oportunidades inmensas y riesgos significativos para la reputación corporativa. Es clave identificar los riesgos reputacionales asociados por grupos de interés y diseñar los planes de remediación adecuados. Un análisis bien fundamentado guiará a las empresas en la implementación de medidas de control adecuadas y en la maximización de los beneficios estratégicos de la IA, asegurando así la protección y el fortalecimiento de su marca en un mercado competitivo y en constante cambio.

Mapeo de narrativas de desinformación

En un contexto donde la fragmentación de las audiencias es cada vez más pronunciada, mapear narrativas de desinformación constituye una estrategia proactiva esencial para salvaguardar la reputación de las organizaciones. Esta práctica implica identificar y analizar las tendencias y patrones en la propagación de información falsa que podrían impactar negativamente en la imagen corporativa, y en diferentes segmentos del público. Al comprender cómo y dónde emergen estas narrativas en grupos específicos, las empresas pueden desarrollar respuestas informadas y específicas, minimizando así el riesgo antes de que se convierta en una crisis.

Nueva “normalidad” en la gestión de crisis

Los *deepfakes* demandan un enfoque proactivo y actualizado en la gestión de crisis. Las empresas deben mantener una monitorización constante de las plataformas digitales y estar preparadas para activar planes de contingencia específicos y prolongados en el tiempo, puesto que las crisis por desinformación desaparecen y reviven en otras plataformas. El nivel de incertidumbre de este tipo de crisis, y el impacto en las marcas personales de los directivos, requieren de esfuerzos coordinados y multiplicados.

Conclusiones

Antídoto tecnológico

La inteligencia artificial proporciona herramientas cruciales para la detección y desmentido de desinformación. Invertir en tecnología avanzada es esencial para el diseño de una estrategia proactiva de gestión de la reputación. Identificar y contrarrestar manipulaciones digitales ayudará a proteger la integridad informativa y la reputación corporativa.

La importancia de construir alianzas

La colaboración entre sectores públicos y privados es vital para enfrentar la desinformación de manera efectiva. Estas alianzas permiten compartir recursos y estrategias, estableciendo un frente unido contra las amenazas a la integridad informativa y la seguridad digital.

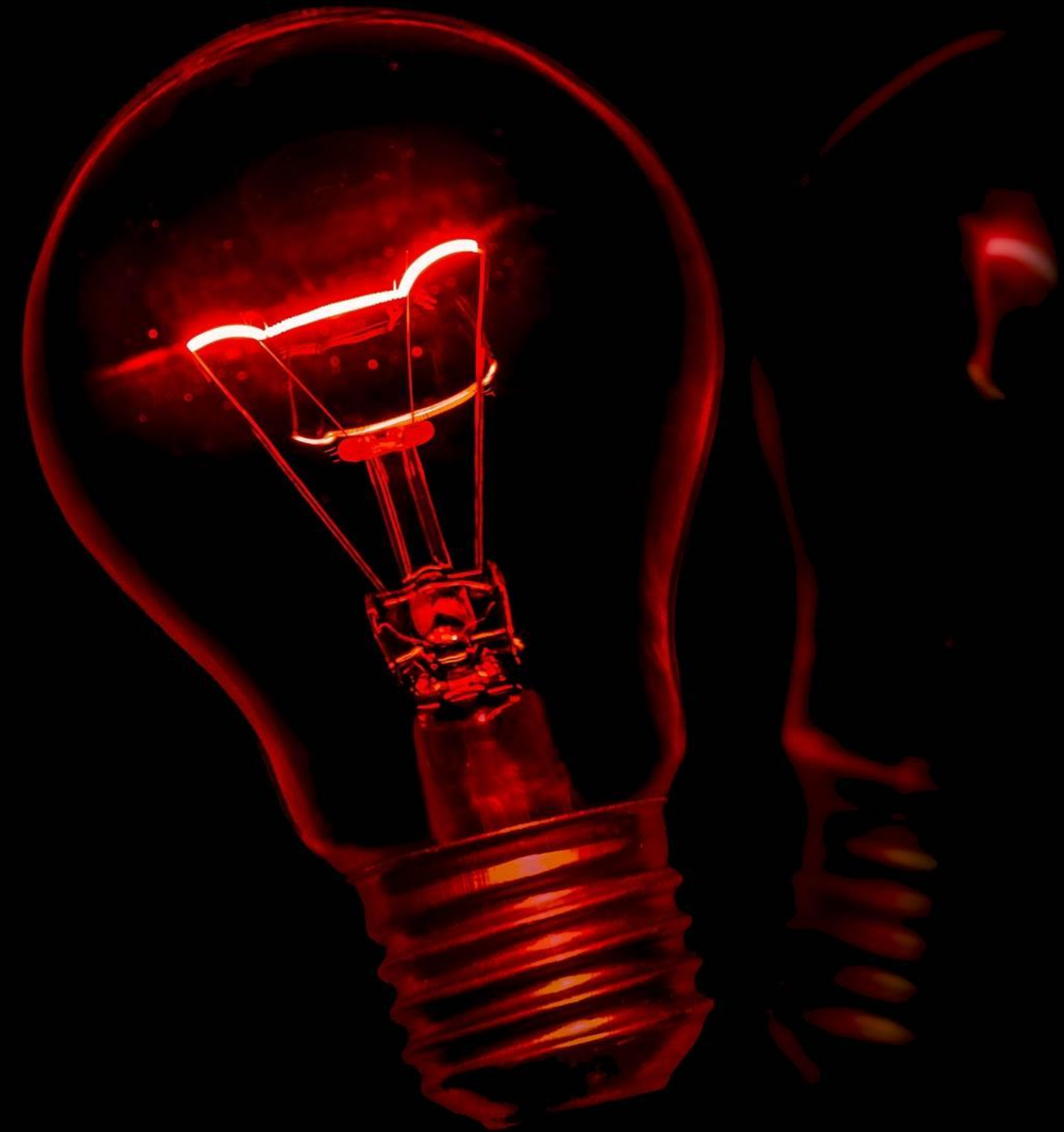
Educación como escudo

Fortalecer la alfabetización digital y fomentar el pensamiento crítico de los usuarios y la sociedad en general, son pasos fundamentales para combatir la desinformación. Las empresas deben liderar la educación de sus audiencias y empleados, mejorando su capacidad para reconocer y rechazar el contenido falso.

Sobre #Thinking&Co_

Thinking&Co_ es el ecosistema de generación de conocimiento de CommsTribe desarrollado a través de sesiones de reflexión co_creativa sobre tendencias en comunicación. Cuenta con la participación de los *tribers* de la comunidad CommsTribe y otros *thinkers* invitados.

Bienvenido a The power of Co_



COMMSTRIBE

THE POWER OF CO_