La mula de Napoleón en la era de la IA generativa: el nuevo código de prácticas europeo para la trazabilidad de contenidos

Resulta inquietante imaginar de qué serán capaces los **futuros modelos generativos**, alimentados en su entrenamiento con cantidades ingentes de datos. Una **evolución técnica imparable** que, de no ponerle remedio, acelerará la caída por el abismo de la desinformación. Esta es la cuestión que ha llevado a la Comisión Europea a impulsar un grupo de trabajo, en el que el autor de este artículo participa como uno de los representantes académicos, dedicado a desarrollar el **artículo 50 de la Ley de la IA**. El objetivo es definir un **"Código de Prácticas sobre el marcado y etiquetado de contenidos generados por IA".**

FERNANDO PÉREZ GONZÁLEZ

Napoleón, envuelto en una capa anaranjada, monta un corcel blanco que se alza encabritado sobre sus patas traseras. El viento despeina sus crines de oro mientras el general señala a sus tropas el camino en pos de los austriacos. En las rocas aparecen inscritos los nombres de Aníbal y Carlomagno, quienes también osaron franquear las montañas desde el norte. La pintura *Napoleón cruzando los Alpes* no deja de ser una de tantas obras de Jacques-Louis David caracterizadas por esa perturbadora rigidez teatral, pero ha pasado a la historia porque Napoleón no cruzó los Alpes a caballo, sino a lomos de una robusta mula.

El daguerrotipo, precursor de la fotografía, llegó prometiendo terminar con licencias artísticas como las de David. Por fin, los hechos podrían ser reflejados con "veracidad óptica". Se cuenta que el pintor Paul Delaroche, al ver uno

Fernando Pérez González es catedrático del Departamento de Teoría de la Señal y Comunicaciones de la Universidad de Vigo

de los primeros daguerrotipos en 1839, exclamó con amargura: "Desde hoy, la pintura está muerta" (aunque no debió decirlo muy convencido, porque no solo no cambió de oficio, sino que acabó pintando a un aterido Napoleón cruzando los Alpes sobre una sufrida mula). En todo caso, la inusitada credibilidad que trajo este revolucionario "espejo con memoria" pronto proporcionó un salvoconducto de legitimidad para quienes se apresuraron a fabricar realidades alternativas bajo el disfraz de la evidencia. Al principio, solo eran pequeños retoques del negativo con pinceles finísimos, para eliminar imperfecciones en los retratos. Esa técnica de retoque pronto evolucionó a métodos más sofisticados, como el célebre montaje de Ulysses S. Grant a caballo datado hacia 1864. Identificada en 2007 por la Biblioteca del Congreso de los Estados Unidos como un complejo *collage* de tres fotografías distintas, esta imagen fabricó una épica heroica inexistente, no muy diferente de la pretendida por David.

Durante el siglo XX, estas prácticas se convirtieron en un recurso sistemático de propaganda. En la Unión Soviética, la manipulación alcanzó una dimensión orwelliana: las personas desaparecían por completo de los registros visuales conforme eran purgadas por el régimen. Otros dictadores, como Hitler, Mussolini o Mao, también recurrieron de forma habitual a la alteración de sus fotos oficiales según su conveniencia. Lograr que tales engaños resultaran naturales y convincentes exigía la intervención de expertos retocadores, únicos capaces de manejar con precisión herramientas como el aerógrafo para intervenir la realidad de forma imperceptible.

La proliferación de bulos con evidencias visuales trucadas está devastando nuestra percepción de la realidad

La digitalización de la fotografía en la década de los 90 exacerbó el problema: enseguida surgieron programas de edición digital, capaces no solo de realzar la imagen, sino de producir con unos pocos clics de ratón lo que antes requería un laboratorio. Con el conocido Photoshop como exponente máximo, la manipulación de la realidad se convirtió en una tentación demasiado inmediata que dio pie a numerosos escándalos del fotoperiodismo, desde Brian Walski, despedido en 2003 de *Los Angeles Times* por combinar dos fotos de la Guerra de Irak para mejorar la composición, hasta el caso del reputado Narciso Contreras, a quien su pericia con la edición digital le costó el puesto en Associated Press en 2014 tras eliminar un inconveniente objeto en la imagen de un combatiente en la guerra de Siria.

Aunque la mayor usabilidad de estos programas fue simplificando la edición,

seguía siendo necesaria una cierta destreza para que el engaño fuese imperceptible. El verdadero salto cualitativo ocurrió en 2017, cuando el término *deepfake* irrumpió para describir falsificaciones de rostros generadas mediante redes neuronales. Sin embargo, lo que entonces requería miles de imágenes y horas de procesamiento, hoy se ha democratizado con la inteligencia artificial (IA) generativa. Ya no hace falta pericia técnica, sino apenas una indicación o *prompt*: basta con describir un deseo para que la máquina lo materialice. La diferencia es colosal: el dominio del programa de edición ha sido sustituido por el mero lenguaje.

Al principio, las alucinaciones de la IA eran toscas y dejaban costuras al aire: manos con seis dedos, pendientes que se fundían con el lóbulo o sombras que desafiaban las leyes de la física. Pero las grietas se están cerrando a una velocidad vertiginosa. Lo comprobamos en 2023, cuando la imagen de un papa Francisco enfundado en un inmaculado plumífero de Balenciaga dio la vuelta al mundo, logrando efímeramente que millones de personas aceptaran como real una estampa puramente sintética. La generación de vídeo parecía ser un reto mucho más complejo: justo entonces se popularizaba, por horripilante, un vídeo generado por IA en el que el rostro de Will Smith se transformaba grotescamente mientras engullía un plato de espaguetis que parecía fundirse con su boca. Pocos apostaban entonces a que esa falta de consistencia temporal tardaría tan poco en resolverse. Así que resulta inquietante imaginar de qué serán capaces los futuros modelos generativos, alimentados en su entrenamiento con cantidades ingentes de datos. Una evolución técnica imparable que, de no ponerle remedio, acelerará nuestra caída por el abismo de la desinformación.

Si las redes sociales habían allanado el camino a la posverdad, la IA generativa ha tendido la alfombra dorada: ya no se trata solo de parodias inofensivas, la proliferación de bulos con evidencias visuales trucadas está devastando nuestra percepción de la realidad. El riesgo se agrava cuando estas piezas permean hasta los medios de comunicación tradicionales que, víctimas de la inmediatez, a veces relajan sus protocolos de verificación, otorgando carta de naturaleza a lo que es inventado y socavando así su propia reputación. Para describir este naufragio, cito las proféticas palabras de Hannah Arendt en *Los orígenes del totalitarismo*: "Las masas habían llegado al punto de que, al mismo tiempo, creían todo y nada, pensaban que todo era posible y que nada era verdad".

Ante este escenario, cabe preguntarse si la tecnología que con tanta fuerza nos empuja al nihilismo puede ofrecernos un asidero. Esta es, precisamente, la cuestión que ha llevado a la Comisión Europea a impulsar un grupo de trabajo, en el que participo como uno de los re-

presentantes de la academia, dedicado a desarrollar el artículo 50 de la Ley de la IA. El objetivo es definir un "Código de Prácticas sobre el marcado y etiquetado de contenidos generados por IA" que establezca estándares técnicos de transparencia y detección. Con ello, buscamos ofrecer un marco de soluciones al que se adhieran los proveedores de sistemas de IA generativa, garantizando así que la procedencia de cualquier contenido multimedia sea siempre verificable. Comoquiera que no existe una solución universal e infalible, la aproximación propuesta combina varias capas tecnológicas redundantes y complementarias que revisaremos a continuación.

La primera de estas capas consiste en la integración de metadatos firmados digitalmente. El borrador del Código establece que los proveedores deben incrustar información criptográfica con un sello temporal que certifique si el contenido ha sido generado o manipulado por una IA. Esta idea se inspira y apoya en el estándar industrial C2PA. El consorcio C2PA es una alianza de empresas tecnológicas y de medios (como Microsoft, Adobe, Sony o la BBC) dedicada a crear estándares abiertos para la certificación de la procedencia y la autenticidad de los contenidos. Aunque este sistema nació originalmente para certificar contenidos "naturales" y proteger la veracidad de la imagen capturada por una lente, la Ley de IA le otorga ahora la función inversa: señalizar de forma inequívoca lo puramente sintético.

Técnicamente, el mecanismo se basa en la generación de un resumen matemático único (*hash*) que vincula los datos del archivo con un "manifiesto digital". Este certificado se firma con una clave privada que identifica al proveedor, asegurando la integridad del registro; la alteración de cualquier bit del contenido provocará que el *hash* deje de coincidir y el vínculo de confianza se romperá. Para garantizar la interoperabilidad, el sistema utiliza estándares abiertos que permiten a cualquier plataforma verificar la información mediante la consulta de este registro criptográfico.

Las mayores debilidades de esta capa son la facilidad con la que los metadatos se pueden eliminar (de hecho, muchas redes sociales borran los metadatos de los contenidos compartidos en ellas) y el que las alteraciones benignas (que no afectan al contenido, como compresiones para ahorrar espacio) dejan de tener un *hash* válido. Existen *hashes* "robustos" a este tipo de alteraciones, también llamados *hashes* "perceptuales", como los que usa YouTube para detectar contenidos duplicados, pero su implementación no es obligatoria según el borrador actual del Código.

La segunda capa consiste en la inserción de marcas de agua digitales. Las marcas de agua pueden ser visibles (un icono común), como el "diamante" que embebe la IA generativa de Google (Nano Banana) en la esquina inferior

derecha de las imágenes que produce. Como es lógico intuir, resulta muy sencillo eliminar estas marcas, bien empleando *software* para reconstruir la zona ocupada por ellas, bien simplemente recortando la imagen. Por esta razón, el Código hace hincapié en el empleo de marcas de agua imperceptibles. Estas marcas son invisibles para el ojo humano, pero es relativamente fácil verificar su presencia con un elemento (de *software*) denominado "detector". La condición de imperceptibilidad es esencial para que las marcas resulten difíciles de eliminar. El objetivo es que la marca sirva de "testigo" de que el contenido ha sido sintetizado por una IA. Una marca de agua bien diseñada será robusta, esto es, sobrevivirá a alteraciones leves del contenido, ya sean debidas a la propia cadena de procesado (por ejemplo, las redes sociales suelen recomprimir los contenidos multimedia) o producidas con intención de destruirla.

Un elemento crítico en el marcado de agua es el propio detector. Investigaciones realizadas por nuestro grupo y otros expertos han demostrado que, si no existe una limitación en el número de veces que se puede invocar al detector, es posible borrar la marca mediante lo que denominamos "ataques de oráculo" (sistemas que logran eliminar la marca a base de ensayo y error). Por ello, es muy desaconsejable que el detector se distribuya como un programa local en el ordenador del usuario o que permita consultas ilimitadas. Esta necesidad de restringir el acceso por motivos de seguridad explica por qué los distintos proveedores son reacios a desarrollar un estándar común, optando por mantener sus detectores cerrados. Sin embargo, esta fragmentación penaliza al usuario, a quien no le resulta práctico tener que probar suerte con diferentes detectores para averiguar si una imagen es sintética. La solución a este dilema podría radicar en la inserción de una "marca universal" que actúe como índice; es decir, una señal estandarizada que contenga únicamente la información sobre qué detector específico se debe emplear. No obstante, para que esta arquitectura funcione, la extrema robustez de dicha marca universal resulta crucial.

Una marca de agua bien diseñada debe ser robusta, esto es, sobrevivirá a alteraciones leves del contenido

Por último, existe una barrera que, aunque no es obligatoria en el documento de trabajo del Código, resulta recomendable y tiene la ventaja de no requerir la colaboración de los proveedores de IA generativa: el uso de detectores forenses. Estos detectores buscan huellas imperceptibles que los generadores dejan involuntariamente en los conteni-

dos multimedia que producen. Aunque algunas de estas huellas están bien caracterizadas matemáticamente, los mejores detectores forenses son, a su vez, otras IA entrenadas para distinguir entre imágenes reales y sintéticas. Si bien su capacidad para identificar falsificaciones es hoy muy notable, adolecen de una tasa excesiva de falsos positivos (es decir, imágenes naturales que el detector clasifica como sintéticas), un problema que se ve agravado por el hecho de que sus decisiones no son explicables.

Gran parte del éxito de la iniciativa dependerá de la voluntad de adherirse por parte de las grandes empresas proveedoras de IA generativa

Como vemos, no hay una solución mágica, pero sí motivos para la esperanza. Gran parte del éxito de la iniciativa dependerá de la voluntad de adherirse por parte de las grandes empresas proveedoras de IA generativa, en su mayoría estadounidenses y casi todas participantes en el grupo de trabajo. Esperemos que sus habituales quejas hacia lo que consideran un exceso de reglamentismo europeo no supongan un obstáculo para la publicación, prevista para agosto de este año, de un Código que ofrezca soluciones concretas a un problema cada día más acuciante.

Las medidas que propone este documento buscan convertir en norma algo que Jacques-Louis David omitió en su lienzo: el aviso explícito de la licencia artística. Hoy es la IA la que fabrica la ficción y el Código pretende establecer mecanismos para que la obra digital delate sus artificios. Si la industria tecnológica asume este compromiso, los majestuosos corceles blancos generados a golpe de *prompt* no dejarán de asombrarnos. Simplemente contaremos con la certeza de que, oculto entre los píxeles, un testigo digital nos estará recordando que, en el mundo real, Napoleón cruzó aterido los Alpes a lomos de una mula. ■