

Twitter contra los deep fakes: la red social etiquetará los videos falsos



La red social anunció nuevas reglas para combatir la difusión de videos manipulados, con el fin de viralizar engaños. La red social dijo que cuando identifique este tipo de contenido falso le añadirá una etiqueta y mostrará una advertencia a los usuarios antes de que le den un like o retuit.

A su vez, se reducirá la visibilidad del tuit en Twitter y se ofrecerán explicaciones o aclaraciones adicionales, siempre y cuando estén disponibles, como redireccionar a otra página que ofrezca más contexto.

“No puedes compartir de manera engañosa contenido sintético o manipulado que pueda causar daño. Además, podríamos etiquetar los tuits que presenten contenido sintético y manipulado, a fin de ayudar a que las personas comprendan la autenticidad del contenido, así como ofrecer un contexto adicional”, informó la compañía en un comunicado.



La tecnología de deep fakes permite insertar la cara de una persona para crear un video falso. Esta es una captura de un video porno que fue alterado para poner el rostro de Gal Gadot en el de la protagonista.

Cabe recordar que esto impactará, entre otras cosas, en los **deep fakes**, que son aquellos videos que son alterados de modo tal que parezca que una persona está diciendo algo que no dijo. También se emplea esta tecnología para fusionar personas digitalmente o crear rostros 100% digitales.

La red social explicó el criterio que tendrán en cuenta para considerar contenidos audiovisuales engañosos. En este sentido, dijeron que evaluarán si el contenido fue editado al punto de alterar, de forma sustancial, la composición, secuencia, tiempo, o número de cuadros en el video.

A su vez, se evaluará si alguna información visual o auditiva fue agregada o eliminada o si el contenido fue fabricado con el fin de generar un engaño o un mensaje manipulado.



Un caso de un video manipulado donde se cruza el cuerpo de Jennifer Lawrence con Steve Buscemi. Con este contenido se buscó hacer una humorada sobre este tipo de tecnología.

“También consideraremos si el contexto en el que se comparten los contenidos podría resultar en una confusión o falta de entendimiento, o si sugiere un intento deliberado para engañar a la gente sobre la naturaleza o el origen del contenido; por ejemplo, si declara falsamente que muestra la realidad”, se menciona en el comunicado.

Se tendrá en cuenta el contexto que se proporciona junto a los contenidos, es decir: el texto que acompaña el tuit, los metadatos asociados al contenido compartido, la información en el perfil de la persona que comparte el material y los sitios web vinculados en el perfil de la persona que comparte el contenido, o en el tuit.

A su vez, si el contenido manipulado pudiera impactar en la seguridad pública o causar un daño serio podrían ser eliminados por completo. En esta clasificación entran los contenidos que impliquen amenazas a la seguridad física de un grupo o persona; el riesgo de una violencia masiva o de disturbios civiles generalizados; así como las amenazas a la privacidad o a la capacidad de una persona o grupo, de expresarse libremente o participar en eventos cívicos.

En octubre del año pasado se había anticipado ya que Twitter tomaría medidas para combatir los deep fakes aunque en aquella oportunidad no se había mencionado cuáles serían esas nuevas reglas. Ahora, por medio de la difusión de esta iniciativa se tiene más claro cómo se llevará adelante este desafío, cada vez mayor, para luchar contra el contenido falso.