Democracia deepfake: la tecnología de inteligencia artificial complica la seguridad electoral

Si bien los riesgos de ciberseguridad para el proceso democrático han sido omnipresentes desde hace muchos años, la prevalencia de la IA representa ahora nuevas amenazas.

Imagen de Nathan Eddy, escritor colaborador

Nathan Eddy, escritor colaborador

9 de febrero de 2024

Lectura de 5 minutos

Una mano poniendo una papeleta en una urna sobre un fondo naranja

FUENTE: SAPHIENS A TRAVÉS DE UNA FOTO DE STOCK DE ALAMY

Los acontecimientos recientes, incluida una llamada automática falsa generada por inteligencia artificial (IA) que se hace pasar por el presidente Biden e insta a los votantes de New Hampshire a abstenerse de las primarias, sirven como un claro recordatorio de que los actores maliciosos ven cada vez más a las plataformas modernas de IA generativa (GenAI) como un arma potente para atacar. Elecciones estadounidenses.

Plataformas como ChatGPT, Gemini de Google (anteriormente Bard) o cualquier número de modelos de lenguaje grande (LLM) de la Dark Web especialmente diseñados podrían desempeñar un papel en la interrupción del proceso democrático, con ataques que abarcan campañas de influencia masiva, trolling automatizado y la proliferación de contenido falso.

De hecho, el director del FBI, Christopher Wray, expresó recientemente su preocupación por la guerra de información en curso mediante el uso de deepfakes que podrían sembrar desinformación durante la próxima campaña presidencial, mientras los actores respaldados por el estado intentan influir en los equilibrios geopolíticos.

GenAI también podría automatizar el surgimiento de redes de "comportamiento coordinado no auténtico" que intentan desarrollar audiencias para sus campañas de desinformación a través de medios de comunicación falsos, perfiles convincentes en las redes sociales y otras vías, con el objetivo de sembrar discordia y socavar la confianza pública en el proceso electoral. .

Influencia electoral: riesgos sustanciales y escenarios de pesadilla

Desde la perspectiva de Padraic O'Reilly, director de innovación de CyberSaint, el riesgo es "sustancial" porque la tecnología está evolucionando muy rápidamente.

"Promete ser interesante y quizás también un poco alarmante, a medida que veamos nuevas variantes de desinformación que aprovechan la tecnología deepfake", afirma.

Específicamente, dice O'Reilly, el "escenario de pesadilla" es que la microtargeting con contenido generado por IA proliferará en las plataformas de redes sociales. Esa es una táctica familiar del escándalo de Cambridge Analytica, donde la compañía acumuló datos de perfiles psicológicos de 230 millones de votantes estadounidenses, para enviar mensajes altamente personalizados a través de Facebook a individuos en un intento de influir en sus creencias y votos. Pero GenAI podría automatizar ese proceso a escala y crear contenido altamente convincente que tendría pocas, o ninguna, características de "robot" que pudieran desanimar a las personas.

"El robo de datos de objetivos [instantáneas de la personalidad de un usuario y sus intereses] combinados con contenido generado por IA es un riesgo real", explica. "Las campañas de desinformación rusas de 2013-2017 sugieren qué más podría ocurrir y ocurrirá, y conocemos de deepfakes generados por ciudadanos estadounidenses [como el que] presenta a Biden y Elizabeth Warren".

La combinación de redes sociales y tecnología deepfake fácilmente disponible podría ser un arma apocalíptica para la polarización de los ciudadanos estadounidenses en un país que ya está profundamente dividido, añade.

"La democracia se basa en ciertas tradiciones e información compartidas, y el peligro aquí es una mayor balcanización entre los ciudadanos, lo que lleva a lo que la investigadora de Stanford Renée DiResta llamó 'realidades a medida'", dice O'Reilly, también conocido como gente que cree en "hechos alternativos".

Las plataformas que los actores de amenazas utilizan para sembrar división probablemente serán de poca ayuda: agrega que, por ejemplo, la plataforma de redes sociales X, antes conocida como Twitter, ha destruido su control de calidad (QA) del contenido.

"Las otras plataformas han dado garantías estándar de que abordarán la desinformación, pero las protecciones de la libertad de expresión y la falta de regulación aún dejan el campo abierto para los malos actores", advierte.

La IA amplifica los TTP de phishing existentes

GenAI ya se está utilizando para crear campañas de phishing dirigidas y más creíbles a escala, pero en el contexto de la seguridad electoral ese fenómeno es aún más preocupante, según Scott Small, director de inteligencia de amenazas cibernéticas de Tidal Cyber.

"Esperamos ver a los ciberadversarios adoptando IA generativa para hacer que los ataques de phishing y de ingeniería social (las principales formas de ataques relacionados con las elecciones en términos de volumen constante durante muchos años) sean más convincentes, lo que hace más probable que los objetivos interactúen con contenido malicioso. " el explica.

Small dice que la adopción de la IA también reduce la barrera de entrada para lanzar este tipo de ataques, un factor que probablemente aumentará el volumen de campañas este año que intentan infiltrarse en campañas o hacerse cargo de las cuentas de los candidatos con fines de suplantación, entre otras posibilidades.

"Los adversarios criminales y Estados-nación adaptan regularmente señuelos de phishing e ingeniería social a eventos actuales y temas populares, y es casi seguro que estos actores intentarán capitalizar el auge del contenido digital relacionado con las elecciones que se distribuye en general este año, para tratar de entregar contenido malicioso. contenido a usuarios desprevenidos", afirma s.

Defensa contra las amenazas electorales de la IA

Para defenderse de estas amenazas, los funcionarios electorales y las campañas deben ser conscientes de los riesgos impulsados por GenAI y cómo defenderse de ellos.

"Los funcionarios electorales y los candidatos dan constantemente entrevistas y conferencias de prensa de las que los actores de amenazas pueden obtener fragmentos de deepfakes basados en IA", dice James Turgal, vicepresidente de riesgo cibernético de Optiv. "Por lo tanto, les corresponde asegurarse de tener una persona o equipo responsable de garantizar el control sobre el contenido".

También deben asegurarse de que los voluntarios y trabajadores estén capacitados sobre amenazas impulsadas por IA, como ingeniería social mejorada, los actores de amenazas detrás de ellas y cómo responder a actividades sospechosas.

Con ese fin, el personal debe participar en capacitación en ingeniería social y videos deepfake que incluya información sobre todas las formas y vectores de ataque, incluidos los intentos electrónicos (correo electrónico, mensajes de texto y plataformas de redes sociales), en persona y por teléfono.

"Esto es muy importante, especialmente en el caso de los voluntarios, porque no todo el mundo tiene una buena higiene cibernética", afirma Turgal.

Además, los voluntarios de campañas y elecciones deben recibir capacitación sobre cómo proporcionar información de manera segura en línea y a entidades externas, incluidas publicaciones en las redes sociales, y deben tener cuidado al hacerlo.

"Los actores de amenazas cibernéticas pueden recopilar esta información para adaptar señuelos diseñados socialmente a objetivos específicos", advierte.

O'Reilly dice que a largo plazo, la regulación que incluye marcas de agua para deepfakes de audio y video será fundamental, y señala que el gobierno federal está trabajando con los propietarios de LLM para implementar protecciones.

De hecho, la Comisión Federal de Comunicaciones (FCC) acaba de declarar "artificiales" las llamadas de voz generadas por IA en virtud de la Ley de Protección al Consumidor Telefónico (TCPA), ilegalizando el uso de la tecnología de clonación de voz y proporcionando a los fiscales generales estatales de todo el país nuevas herramientas para combatir tales llamadas. actividades fraudulentas.

"La IA se está moviendo tan rápido que existe un peligro inherente de que cualquier regla propuesta se vuelva ineficaz a medida que avanza la tecnología, potencialmente perdiendo el objetivo", dice O'Reilly. "En cierto modo, es el Salvaje Oeste, y la IA está llegando al mercado con muy pocas salvaguardias".