IA para todas as idades

13 febreiro

Qué é un *fake* audiovisual?

- Contido de vídeo, imaxe ou audio **alterado ou xerado artificialmente** mediante IA.
- Pode ter como obxectivo suplantar identidades, manipular información ou crear escenas que nunca ocorreron.
- Inclúe *deepfakes* (rostros/voz), vídeos sintéticos e audios clonados.

Qué é un *deepfake*?

- Medios sintéticos onde a cara, a voz ou os xestos dunha persoa se substitúen polos doutra mediante redes neuronais.
- O termo xorde de *deep learning* + *fake*.
- Baséanse principalmente en modelos xerativos como as GANs.

Raíces históricas

- **Anos 1990 – CGI avanzado:** primeiros intentos de xerar rostros realistas por ordenador. Non ten a calidade suficiente para resultar confuso. (https://www.youtube.com/watch?v=UvLQMMaVmzU)
- **1997 – Video Rewrite (Stanford):** permitía alterar vídeo para que unha persoa “pronunciase” un audio diferente; un precedente claro. (https://www.youtube.com/watch?v=5ymelRjHIfI)

Últimos anos

- 2014 - Aparición das GANS
- Permiten xerar imaxes e vídeo extremadamente realistas enfrontando dúas redes: xerador vs. discriminador.
- Base da maioría dos deepfakes modernos
- 2017 marca a chegada do deepfake ao público xeral e, con iso, tamén o seu uso indebido.
- 2018**:** expertos alertan da rapidez do avance; plataformas comezan a regular.
- 2019 en diante: xa presentan calidade moi alta e aparecen leis e competicións de detección.

Casos famosos

- En 2020 difundíuse un deepfake da primeira ministra belga Sophie Wilmès no que se lle atribuían declaracións falsas sobre a COVID-19.
- O vídeo, que se fixo viral, converteuse nun dos primeiros exemplos europeos de manipulación política mediante deepfakes e puxo en alerta gobernos e medios sobre os riscos da desinformación audiovisual.”

Caso Baltimore

En xaneiro de 2024 circulou un **audio falso xerado con IA** no que supostamente o director do Pikesville High School (Maryland) facía **comentarios racistas e antisemitas**.
O contido fíxose viral en redes sociais e provocou:

- A súa retirada temporal do cargo
- A recepción de mensaxes de odio
- Medidas de seguridade para protexer á súa familia

Tras unha investigación da policía, o FBI e expertos forenses, confirmouse que o audio era **fabricado con IA**.
O autor resultou ser o **director deportivo da escola**, Dazhon Darien, que o fixo para retaliar por problemas laborais. Foi detido e acusado de múltiples delitos.

Caso Lionel Messi

En marzo de 2024 circularon en Instagram **anuncios manipulados** nos que Lionel Messi supostamente promocionaba a app **“Wildcat Dive”** como unha das súas principais fontes de ingresos.
O vídeo era un **deepfake** que empregaba fragmentos dunha entrevista real que Messi deu ao programa arxentino **“Olga”**.
A voz do futbolista e do entrevistador foron **modificadas artificialmente** para parecer auténticas.
O deepfake estaba deseñado para **dar confianza**:

- Messi mencionaba ingresos reais doutras marcas.
- Inseríanse clips publicitarios durante a súa “fala” para ocultar discrepancias entre labios e voz.

A app, analizada por ESET e medios especializados, resultou ser **fraudulenta**, con reseñas sospeitosas e relatos de usuarios incapaces de retirar o diñeiro investido.

Caso Elon Musk

- Deepfakes de Elon Musk foron usados en **YouTube, Instagram, TikTok, Facebook e X** para promocionar falsas inversións en criptomoedas.
- Os vídeos imitaban a súa voz e aparencia, prometendo **altas ganancias aseguradas** e dirixindo ás vítimas a webs fraudulentas.
- A FTC rexistrou **case 7.000 vítimas** que perderon **80 millóns de dólares** en estafas relacionadas con criptomoedas, incluíndo esquemas usando a imaxe de Musk.
- Estes deepfakes aproveitan a enorme presenza mediática de Musk e o seu vínculo público co mundo cripto para dar **credibilidade falsa** ás estafas.

Caso Ferrari

- En xullo de 2024, un executivo de Ferrari recibiu mensaxes e unha chamada dun número descoñecido que **imitaba á perfección a voz e o acento do CEO Benedetto Vigna**, mediante IA.
- O falso CEO solicitou **colaboración nunha operación financeira e unha transferencia relacionada cun hedge cambiario**, alegando confidencialidade e urxencia.
- A suplantación foi tan convincente que só se descubriu cando o executivo lle fixo ao impostor unha **pregunta persoal sobre un libro recomendado días antes**, algo que a IA non puido responder. O estafador colgou inmediatamente.
- Ferrari abriu unha investigación interna; o ataque evitouse grazas á **verificación humana baseada en información privada**.

Caso Arup

- En xaneiro de 2024, un empregado financeiro da empresa de enxeñaría Arup foi vítima dun dos fraudes con IA máis sofisticados rexistrados.
- O ataque comezou cun **correo de phishing** que parecía proceder do CFO do Reino Unido pedindo un "*transacción confidencial*".
- Para reforzar a credibilidade, os estafadores organizaron unha **videoconferencia con varios directivos… todos eles deepfakes de vídeo e audio**.
- O empregado, convencido pola presenza dos supostos superiores, realizou **15 transferencias** que sumaron **200 millóns de HKD (~25,6 millóns de dólares)** cara a contas en Hong Kong.
- O fraude descubriuse cando o empregado contactou máis tarde coa sede real de Arup, que negou ter solicitado tal operación.