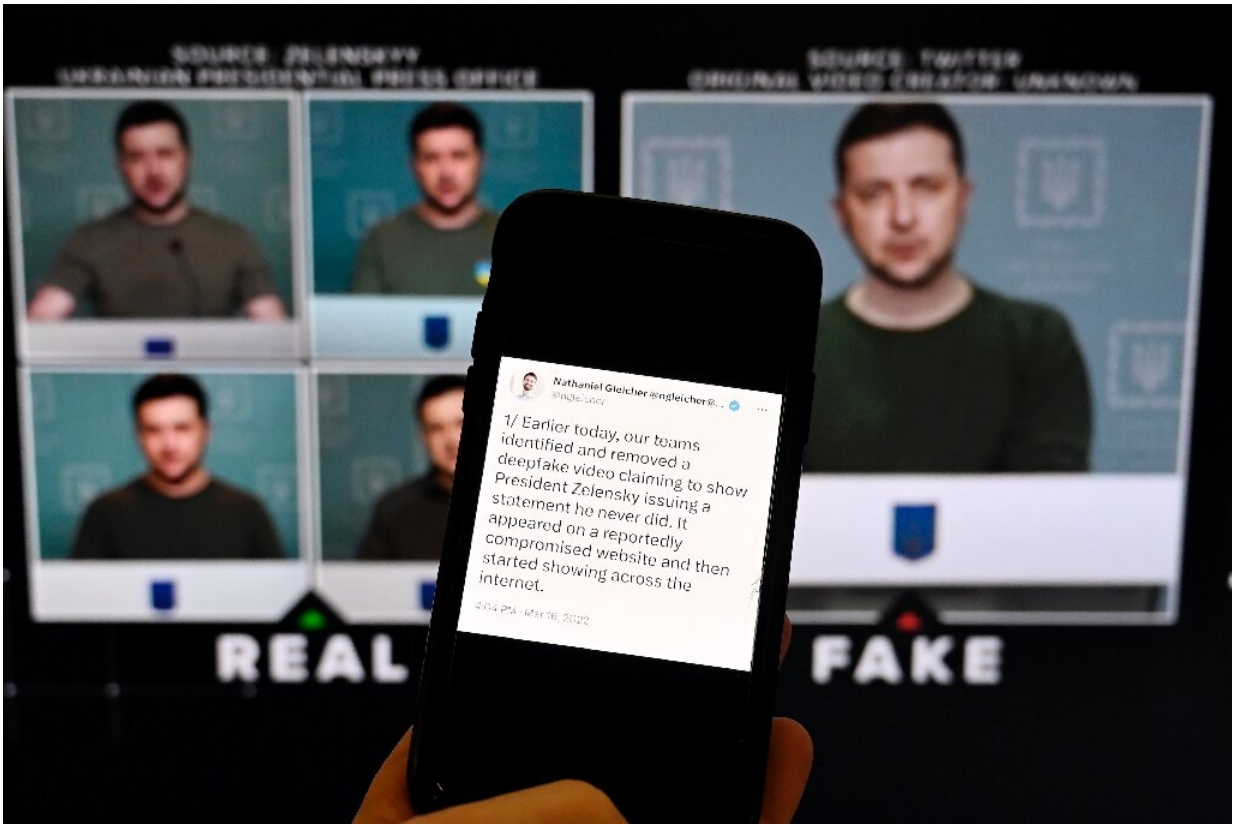
Seeing is believing? Global scramble to tackle deepfakes

February 2 2023, by Anuj Chopra with Saladin Salem in Berlin



A phone shows a statement from the head of security policy at Meta in front of a fake video of Ukrainian President Volodymyr Zelensky calling on his soldiers to lay down their weapons.

Chatbots spouting falsehoods, face-swapping apps crafting porn videos

and cloned voices defrauding companies of millions—the scramble is on to rein in AI deepfakes that have become a misinformation super spreader.

Artificial Intelligence is redefining the proverb "seeing is believing," with a deluge of images created out of thin air and people shown mouthing things they never said in real-looking deepfakes that have eroded online trust.

"Yikes. (Definitely) not me," tweeted billionaire Elon Musk last year in one vivid example of a deepfake video that showed him promoting a crypto currency scam.

China recently adopted expansive rules to regulate deepfakes but most countries appear to be struggling to keep up with the fast-evolving technology amid concerns that regulation could stymie innovation or be misused to curtail free speech.

Experts warn that deepfake detectors are vastly outpaced by creators, who are hard to catch as they operate anonymously using AI-based software that was once touted as a specialized skill but is now widely available at low cost.

Facebook owner Meta last year said it took down a deepfake video of Ukrainian President Volodymyr Zelensky urging citizens to lay down their weapons and surrender to Russia.

And British campaigner Kate Isaacs, 30, said her "heart sank" when her face appeared in a deepfake porn video that unleashed a barrage of online abuse after an unknown user posted it on Twitter.

"I remember just feeling like this video was going to go everywhere—it was horrendous," Isaacs, who campaigns against non-consensual porn,

was quoted as saying by the BBC in October.

The following month, the British government voiced concern about deepfakes and warned of a popular website that "virtually strips women naked."

'Information apocalypse'

With no barriers to creating AI-synthesized text, audio and video, the potential for misuse in identity theft, financial fraud and tarnish reputations has sparked global alarm.

The Eurasia group called the AI tools "weapons of mass disruption."

"Technological advances in artificial intelligence will erode social trust, empower demagogues and authoritarians, and disrupt businesses and markets," the group warned in a report.

"Advances in deepfakes, facial recognition, and voice synthesis software will render control over one's likeness a relic of the past."

This week AI startup ElevenLabs admitted that its voice cloning tool could be misused for "malicious purposes" after users posted a deepfake audio purporting to be actor Emma Watson reading Adolf Hitler's biography "Mein Kampf."

Elon Musk tweeted "Yikes. Def not me" about a deepfake video of him supposedly promoting a new cryptocurrency scam.

The growing volume of deepfakes may lead to what the European law enforcement agency Europol described as an "information apocalypse," a scenario where many people are unable to distinguish fact from fiction.

"Experts fear this may lead to a situation where citizens no longer have a shared reality or could create societal confusion about which information sources are reliable," Europol said in a report.

That was demonstrated last weekend when NFL player Damar Hamlin spoke to his fans in a video for the first time since he suffered a cardiac arrest during a match.

Hamlin thanked medical professionals responsible for his recovery, but many who believed conspiracy theories that the COVID-19 vaccine was behind his on-field collapse baselessly labelled his video a deepfake.

'Super spreader'

China enforced new rules last month that will require businesses offering deepfake services to obtain the real identities of their users. They also require deepfake content to be appropriately tagged to avoid "any confusion."

The rules came after the Chinese government warned that deepfakes present a "danger to national security and social stability."

In the United States, where lawmakers have pushed for a task force to police deepfakes, digital rights activists caution against legislative overreach that could kill innovation or target legitimate content.

The European Union, meanwhile, is locked in heated discussions over its proposed "AI Act."

The law, which the EU is racing to pass this year, will require users to disclose deepfakes but many fear the legislation could prove toothless if it does not cover creative or satirical content.

"How do you reinstate digital trust with transparency? That is the real question right now," Jason Davis, a research professor at Syracuse University, told AFP.

"The (detection) tools are coming and they're coming relatively quickly. But the technology is moving perhaps even quicker. So like cyber security, we will never solve this, we will only hope to keep up."

Many are already struggling to comprehend advances such as ChatGPT, a chatbot created by the US-based OpenAI that is capable of generating strikingly cogent texts on almost any topic.

In a study, media watchdog NewsGuard, which called it the "next great misinformation super spreader," said most of the chatbot's responses to prompts related to topics such as COVID-19 and school shootings were "eloquent, false and misleading."

"The results confirm fears... about how the tool can be weaponized in the wrong hands," NewsGuard said.

Citation: Seeing is believing? Global scramble to tackle deepfakes (2023, February 2) retrieved 26 April 2024 from https://techxplore.com/news/2023-02-believing-global-scramble-tackle-deepfakes.html