AI Explained · Deepfakes, Part 1

How Deepfakes Are Made and Why We Believe Them

By Dr Peter Fields

THE 2-MINUTE SUMMARY

AI can imitate a voice, change what a person appears to say on video, or generate a scene that never happened. A genuine clip with a misleading caption can cause much the same confusion.

A familiar voice, a plausible story and a request that seems to come from someone in authority can all affect how we respond. The research distinguishes recognising a person, believing a claim and acting on it.

For an unusual request, call back through a route you already trust. For a public claim, find the original and look for independent evidence. The same checks apply when someone says a real image or recording is fake.

The meeting that looked real

In January 2024, a finance worker at the engineering firm Arup received an email that appeared to come from the chief financial officer. It invited the worker to a video meeting about a confidential transaction. Hong Kong's official account says the meeting was pre-recorded. The fraudster followed up by instant message. The worker authorised transfers to five accounts and the reported loss was about HK$200 million.[1]

ITV News reported that the employee initially suspected phishing. On the call, people who looked and sounded like colleagues appeared to confirm the request. But a video call arranged by the sender of a suspicious email is part of the same chain of contact.[1]

How a deepfake is made

A deepfake is an AI-generated or altered image, video or recording that makes a person appear to say or do something they did not, or presents an invented scene as real. In a voice clone, a recording gives the system a reference for how someone sounds. It can then produce speech in that voice from new text. A video tool can swap a face, change mouth movements to fit a different soundtrack, or generate a whole scene.[2]

A language model that generates text one piece at a time predicts the next token, a small unit of text. Some speech models use a similar idea, but their tokens represent pieces of sound. Video generation has to keep faces, objects and movement coherent from one moment to the next. The same broad techniques have legitimate uses. Someone losing their speech, for instance, can bank their voice for a communication aid.[2,3]

Why a plausible story gets an easier hearing

Confirmation bias is the tendency to give more weight to evidence that fits what we already believe. When a clip appears to catch a politician doing exactly what we suspected, it is worth asking what proof we would want if it showed someone we supported.

There is also the question of what we expect the speaker to say. Hameleers, van der Meer and Dobber tested several fake speeches by Nancy Pelosi with participants in the United States and the Netherlands. The version closest to her known positions was rated more credible than the two more implausible versions. The most extreme fake still led to more negative views of Pelosi than the genuine clip did. The two results measure different things. A clip can score poorly for credibility and still leave its audience with a worse impression of the person.[4]

Familiar words and familiar voices

In experiments with news headlines, Pennycook, Cannon and Rand found that seeing a false headline before made it seem a little more accurate later, even when it ran against the reader's politics. That may help explain why a claim starts to feel more plausible after it turns up in several feeds. The study tested headlines, so it leaves open how strong the effect is for video.[5]

Knowing the person can help. In an experiment by Rosi, Soopramanien and McGettigan, listeners familiar with a voice were better at spotting clones. Participants knew they were listening for a clone and had recordings to compare. On an unexpected call you may have just one voice to go on. In a separate task, familiar listeners rated clones that sounded more like the original voice as more trustworthy.[6]

Why we say yes

In Influence, Robert Cialdini describes the cues we use when deciding whether to go along with a request. A finance director appears to have authority. A familiar colleague may benefit from what Cialdini calls liking. If others in a meeting seem to accept what is happening, that gives us social proof, the sense that they have already judged it reasonable. A deadline can suggest we are about to lose an opportunity, a form of scarcity. These cues help explain why the set-up might work. We know little about what went through the Arup employee's mind.[1,7]

Does a deadline make us easier to fool? In an email experiment, Auton and Sturman found that people given seven seconds per message were worse at distinguishing phishing from genuine emails than those given 15. Black and Sarno set a deadline for the whole task. People decided faster, with no clear change in accuracy. In their study, emails invoking authority took the longest to assess and were the hardest to classify. Both studies used email, so they leave open what happens under pressure on a call.[8]

A voice that told voters to stay home

Days before the 2024 New Hampshire primary, automated calls using an imitation of Joe Biden's voice urged people to stay away from the polls. The US Federal Communications Commission imposed a $6 million penalty on the organiser for the illegal caller-ID spoofing. The order documents the calls and the penalty. It gives no count of voters who stayed home because of them.[9]

What if someone says the real thing is fake?

In 2024, Donald Trump claimed that an image of Kamala Harris's airport crowd in Michigan had been fabricated with AI. Other photos and footage showed the crowd. Chesney and Citron call the advantage a person gains by dismissing genuine evidence as fake the liar's dividend. What independent evidence supports the denial?[10]

Sometimes the effect is uncertainty. Vaccari and Chadwick showed different versions of an Obama deepfake created with Jordan Peele. About 35–37 per cent of viewers who saw versions without the reveal were unsure whether Obama had actually said what the video showed, compared with about 28 per cent of those who saw the full video. The study found no significant increase in belief that Obama had made the statement.[11]

Geek Box · What can a label tell you?

A Content Credential can give you a signed record of where a file came from or which edits were recorded. Even a file with a valid credential can show a staged scene, and a screenshot may lose the record. The companion article, Can You Tell It's AI?, explains what credentials, watermarks and detectors can tell you.[12]

Before you act or share

- **Call back before acting.** A LastPass employee reported a suspected impersonation of the chief executive after receiving urgent messages over WhatsApp, outside normal work channels. If someone asks you to transfer money, change bank details or send sensitive material, contact them on a number you already hold. At work, follow the normal approval process.[13]

- **Trace a dramatic clip to its source.** Find the earliest version you can and check the date and location. Look for a separate recording or independent reporting. Ten accounts reposting one clip still give you one piece of evidence.

- **Give yourself time before sharing.** A clip can feel convincing when it fits what you already believe. If you would want more evidence before sharing a similar claim about someone else, wait for it here too.

What if the person being impersonated is you? Part 2 starts there, with what to do about a fake account, a threatening message or an image already being passed around.

Sources and notes

1. Hong Kong SAR Government (26 June 2024), LCQ9: Combating frauds involving deepfake, reply (1). ITV News (17 May 2024), reporting Arup's confirmation. Official reply · Arup confirmation

2. Chengyi Wang and colleagues (2023), Neural Codec Language Models are Zero-Shot Text to Speech Synthesizers, arXiv:2301.02111. University of Washington (2017), on Suwajanakorn, Seitz and Kemelmacher-Shlizerman's lip-sync study. OpenAI (2024), Video generation models as world simulators. Speech · Lip sync · Video

3. MND Association, Voice banking and communication aids. Guidance

4. Michael Hameleers, Toni G. L. A. van der Meer and Tom Dobber (2024), Distorting the truth versus blatant lies, Computers in Human Behavior 152, 108096, §§5.3–5.4 and 6.1–6.4. Paper

5. Gordon Pennycook, Tyrone D. Cannon and David G. Rand (2018), Prior Exposure Increases Perceived Accuracy of Fake News, Journal of Experimental Psychology: General 147(12), 1865–1880, Studies 2–3 and supplement. Paper and supplement

6. Victor Rosi, Emma Soopramanien and Carolyn McGettigan (2025), Perception and social evaluation of cloned and recorded voices, Computers in Human Behavior: Artificial Humans 4, 100143, §§2.3.6, 3.1–3.3 and 4.2–4.6. Paper

7. Robert B. Cialdini (2021), Influence: The Psychology of Persuasion, new and expanded ed., Harper Business. See also Cialdini's public explanation of the principles. Book information · Principles

8. Jaime Claire Auton and Daniel Sturman (2025), Persuasion under pressure, Information & Computer Security 33(5), pp. 845–859, especially pp. 849 and 853–855. Jeffrey Black and Dawn M. Sarno (2023), The Influence of Time Pressure and Persuasion Principles on Phishing Detection, HFES Proceedings 67(1), pp. 1977–1982, especially pp. 1979–1982 and Table 1. Auton and Sturman · Black and Sarno

9. US Federal Communications Commission (2024), Steve Kramer, Forfeiture Order FCC 24-104, para. 1. FCC order

10. Robert Chesney and Danielle Keats Citron (2019), Deep Fakes: A Looming Challenge for Privacy, Democracy, and National Security, California Law Review 107(6), pp. 1753–1820, section 'The Liar's Dividend: Beware the Cry of Deep-Fake News'. PolitiFact (12 August 2024), corroboration of the Michigan crowd. Chesney and Citron · PolitiFact

11. Cristian Vaccari and Andrew Chadwick (2020), Deepfakes and Disinformation, Social Media + Society 6(1), pp. 4 and 6–9, Fig. 1. Paper

12. Coalition for Content Provenance and Authenticity, C2PA and Content Credentials Explainer and C2PA Harms Modelling, §6.2. Explainer · Limits

13. LastPass (2024), Attempted Audio Deepfake Call Targets LastPass Employee. Stop Scams UK, Get help now. Incident account · Callback guidance