

Deepfakes: How They're Made and How They're Caught

FREE EDUCATIONAL BRIEF · LEARN AI. UNCOVER THE DATA. PROTECT THE TRUTH.

A deepfake is synthetic media that impersonates a real person — a face swapped into video, a voice cloned from seconds of audio, a photorealistic scene that never happened. The tools are cheap, fast, and consumer-grade. The defense is a moving combination of forensic detection, provenance records, and verification habits — with the balance shifting toward provenance.

THE MAKER'S TOOLKIT

» Face swap & reenactment

One person's face mapped onto another's performance, with expression and lighting matched frame by frame.

» Voice cloning

A model learns the timbre and cadence of a target voice from short samples, then speaks any script.

» Full generation

Diffusion models create faces, scenes, and increasingly video from a text description — no source victim footage required.

THE CATCHER'S TOOLKIT

Forensic detectors hunt for artifacts — inconsistent lighting, boundary glitches, unnatural blinking, spectral fingerprints in audio. They work, imperfectly, and each generator improvement erodes them; every detector score is a probability, never a verdict. That is why the field is converging on provenance: cryptographic content credentials attached at capture or creation, recording what device or tool produced a file and what edited it. A signed origin can be checked in seconds; forensic guesswork cannot keep that pace.

THE PRACTICAL PROTOCOL

For any consequential clip: check the source account and date, look for the original upload, check provenance credentials, and corroborate through an independent channel — call the person back on a known number. For creators, registering work with hashes and timestamps (as the SMI Registry does) builds the evidence trail before a dispute exists.

INVESTIGATOR'S TAKEAWAYS

- Seconds of audio or a handful of photos suffice for a credible fake.
- Detection outputs probabilities; treat scores as signals, not proof.
- Provenance credentials beat forensics: verify origin, not artifacts.
- Out-of-band confirmation defeats most real-world voice scams.
- Creators should anchor provenance before they need it, not after.