# FAKING IT

**Richard Ford** *investigates the rise of deepfakes and how best to defend against them*

**Deepfake technology represents one of the most unnerving developments in the cyber threat landscape over the past decade. What began as a curious and humorous novelty involving videos of celebrities saying unlikely things they never actually said has swiftly evolved into a powerful and disruptive force with the potential to undermine trust, damage reputations and destabilise institutions and even nation states. We now live in an era where seeing is no longer believing.**

This isn't a theoretical threat. Deepfakes have already been weaponised for fraud, blackmail, political misinformation and corporate sabotage. The technology is advancing at pace, fuelled by generative artificial intelligence and vast libraries of publicly available training data. As the barriers to entry fall and capabilities improve, deepfakes are moving from the fringes of cyber crime into the mainstream. Organisations can no longer afford to dismiss them as rare or exotic. They are a clear and present danger that demands a proactive, well-considered response.

At the heart of the deepfake phenomenon lies a convergence of several powerful trends. First is the rapid advancement in AI, particularly in the area of generative adversarial networks (GANs). These systems pit two neural networks against each other – one generating synthetic content, the other trying to detect it – creating a feedback loop that produces ever more realistic results. Initially used to enhance images or produce synthetic voices, GANs can now fabricate entire audio-visual experiences that are almost indistinguishable from the real thing.

**Tools that were once the preserve of elite researchers are now freely available online**

Second is the sheer volume of personal data available online. Social media platforms, video-sharing sites and even corporate websites offer a treasure trove of content that can be scraped and fed into training models. Public figures are especially vulnerable due to the amount of footage and audio recordings of them available in the wild. But the average employee is not immune. A few minutes of video, a handful of voice clips, and a social engineering narrative are often all that's required to craft a convincing deepfake tailored to a specific context.

The final piece of the puzzle is accessibility. Tools that were once the preserve of elite researchers are now freely available online. Open-source code, deepfake-as-a-service platforms and low-cost computing power have democratised the creation of synthetic media. This ease of access has led to a proliferation of use cases – some entertaining or artistic, but many more malicious and manipulative.

Deepfakes are now being deployed in phishing attacks, business email compromise (BEC) scams, and disinformation campaigns. Consider the case of a finance employee receiving what appears to be a video call from their CEO, authorising an urgent transfer of funds. Or a journalist tricked into publishing a fake video of a politician admitting to corruption. Or a boardroom manipulated into believing a whistleblower has made damning statements that were never actually spoken. The psychological impact of such attacks is significant. Deepfakes exploit our natural instinct to trust our senses and, in doing so, they introduce a chilling uncertainty into every digital interaction.

The US Department of Homeland Security named deepfakes as a rising threat to national security, citing their potential to disrupt democratic processes, incite unrest and conduct influence operations. In the UK, the National Cyber Security Centre (NCSC) has warned of the potential for deepfakes to erode public trust in institutions and amplify polarisation through synthetic propaganda.

While geopolitical manipulation is one vector, the private sector has its own set of risks to contend with. Fraudsters are already combining deepfakes with traditional social engineering techniques to create highly persuasive lures. One particularly alarming case saw an AI-generated voice used to impersonate a CEO, tricking a subordinate into transferring £200,000 to a criminal account. In another, a technology company discovered that its senior leadership's faces were being cloned and used in fake product endorsements online.

This is not simply a matter of reputational risk. The financial and operational consequences of a successful deepfake attack can be severe. Imagine the fallout if a synthetic video falsely depicted a CEO making racist remarks. Or if a manipulated audio recording suggested a data breach cover-up. Share prices could plummet, customers might flee and regulatory fines could follow. The challenge lies not only in preventing the initial deception, but in recovering trust once doubt has been cast. Even after a deepfake is debunked, the damage to credibility may linger.

In response to this growing threat, a cottage industry of detection tools has emerged. Many of these rely on forensic techniques that analyse media for inconsistencies, for example, subtle lighting mismatches, irregular eye movements, unnatural speech cadences or compression artefacts. Others use machine learning to distinguish between real and synthetic patterns.

While these tools can be effective, they are inherently reactive. The arms race between creators and detectors means that every advance in detection is swiftly followed by a new evasion technique. As with many areas of cyber security, there is no magic bullet. Defensive strategies must go beyond simply identifying deepfakes after the fact. They must encompass prevention, education and resilience.

Training staff to spot the signs of manipulated media is vital, but it must be done in a way that doesn't lead to complete scepticism or paralysis. We do not want a workforce that automatically distrusts every video call or voice message. Instead, we need a healthy scepticism, combined with clear protocols for verifying unusual requests or instructions.

## ULTIMATELY, THE DEEPFAKE PROBLEM IS NOT JUST ABOUT TECHNOLOGY, IT'S ABOUT TRUST

This is especially important for roles with access to sensitive data or financial controls. If a senior executive calls with an urgent demand, there should be an established back channel – perhaps a secondary verification via secure messaging or a face-to-face confirmation – before action is taken. The principle of 'trust, but verify' is more relevant than ever in a world of synthetic personas.

Equally important is fostering a culture where employees feel safe reporting suspected deepfake encounters. Too often, individuals hesitate to raise the alarm for fear of embarrassment or being wrong. Organisations should actively encourage vigilance, making it clear that false positives are preferable to overlooked threats.

To effectively counter deepfakes, we also need to rethink how digital identity is established and authenticated. Traditional methods – such as usernames, passwords or even multi-factor authentication – do little to protect against someone impersonating your face or voice. In the age of deepfakes, identity becomes fluid and the signals we once relied on to verify someone's presence are no longer sufficient.

This has led to growing interest in so-called 'zero-trust' architectures, where no user or device is automatically trusted, regardless of location or appearance. Continuous authentication, behavioural biometrics and cryptographic proofs of origin are all becoming more important as part of a layered defence strategy.

There are also promising developments in digital watermarking and provenance tracking. Initiatives such as the Content Authenticity Initiative (CAI), backed by Adobe and others, aim to embed metadata in images and videos to verify their origin and integrity. While not foolproof, these technologies offer a path forward for verifying legitimate media

and distinguishing it from manipulated content. However, adoption remains patchy. Until provenance tools are widely integrated across devices, platforms and content creation workflows, they will only provide partial coverage. We must also grapple with the possibility that malicious actors will find ways to strip or spoof such metadata, further complicating the trust equation.

## DEEPFAKES HAVE THE POTENTIAL TO ERODE PUBLIC TRUST AND AMPLIFY POLARISATION

The rise of deepfakes also calls for a thoughtful regulatory response. Some jurisdictions have introduced laws criminalising the malicious use of synthetic media, particularly in the context of non-consensual pornography or election interference. Yet legislation alone cannot fully solve the problem. Any successful regulatory framework must strike a careful balance between preventing harm and preserving freedom of expression, artistic experimentation and legitimate research.

It is also vitally important that we hold developers and distributors of generative AI technologies to account. Transparency around training data, model capabilities and intended use is essential. Ethical AI design must be a cornerstone of responsible innovation. That includes building in safeguards to prevent misuse, limiting access to potentially harmful capabilities, and investing in robust oversight mechanisms.

Public awareness is also crucial. Just as we've educated people to spot phishing emails and question dubious phone calls, we must now help them navigate the challenges of synthetic media. This includes not just employees and customers, but the general public, schools and civic institutions. The more people understand how deepfakes work and what they're capable of, the less likely they are to be fooled.

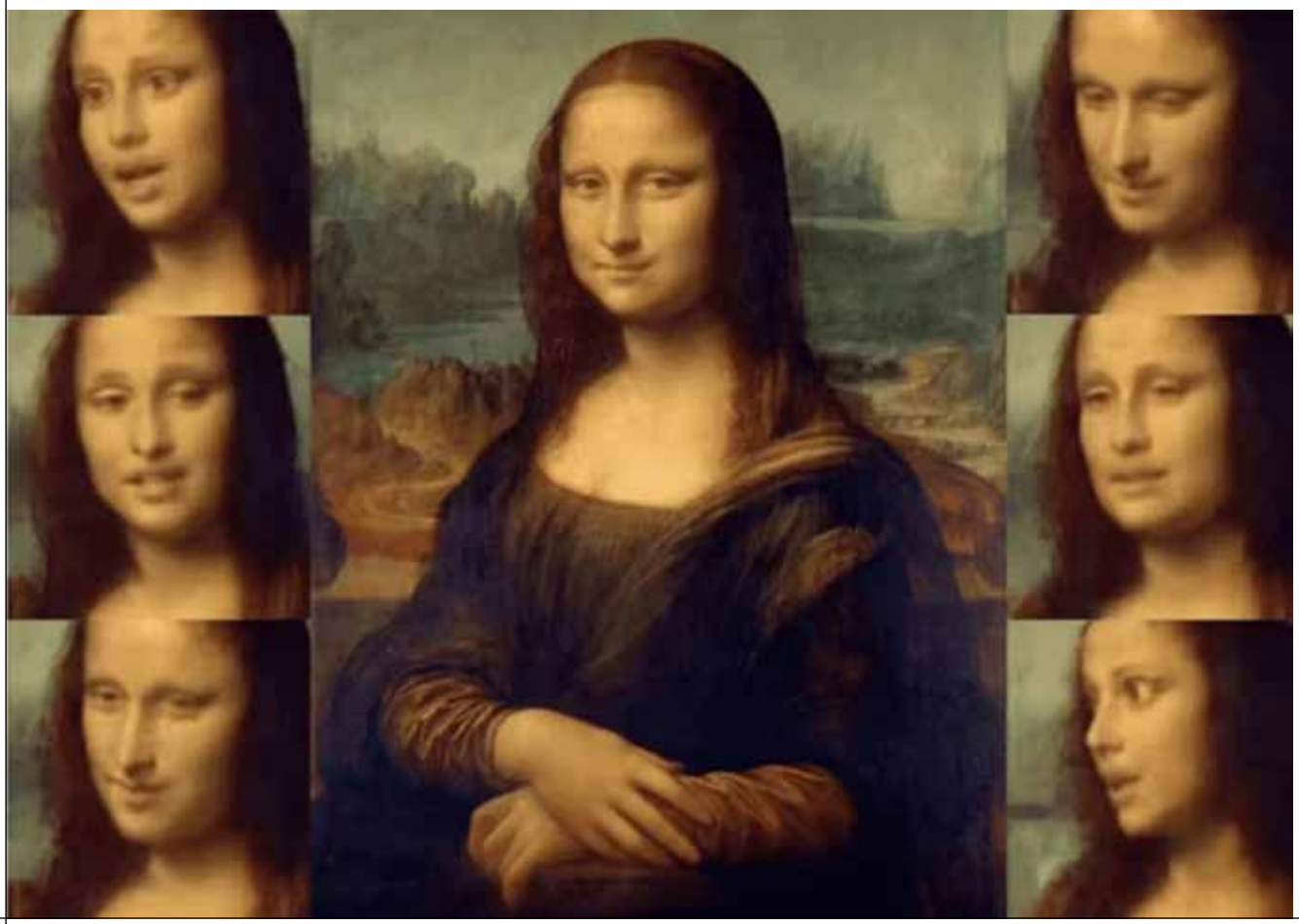
Ultimately, the deepfake problem is not just about technology, it's about trust. Trust in what we see and hear. Trust in our institutions. Trust in each other. As synthetic media grows more sophisticated, we must cultivate a collective resilience that allows us to navigate a world where appearances can deceive.

For businesses, this means adopting a holistic approach to deepfake defence: combining detection tools with policy, education, cultural awareness and layered identity protections. It means engaging with emerging standards, embracing transparency and staying informed about new threats and techniques.

It also means accepting a difficult truth: we cannot eliminate the risk of deepfakes entirely. But we can reduce their impact, limit their effectiveness and respond swiftly when they occur. In doing so, we reinforce the integrity of our communications, security of our operations and credibility of our people.

The rise of deepfakes is one of the defining challenges of our digital age. By facing it head-on – with clarity, caution, and commitment – we stand a much better chance of preserving trust in the midst of technological uncertainty ●

**Richard Ford,** Group Chief Technical Officer at Integrity360, brings over 15 years' experience of the IT sector with the majority of his career spent directly in the IT security channel.



**GANs can now fabricate entire audio-visual experiences that are almost indistinguishable from the real thing**