The Progression of Social Engineering in the Generative AI Era

Aidan Ingram
aidan.ingram@ku.edu
University of Kansas
Lawrence, Kansas, USA

Abstract

Social engineering has long depended on a structural trade-off between the scale of an attack and the personalization required to make it succeed, with defenders relying on observable artifacts like grammatical errors, mismatched URLs, or an unnatural voice/image to identify malicious content. Generative AI has dissolved this trade-off and eliminated those artifacts, producing synthetic media that defeats both human verification and automated detection under real-world conditions. This paper presents a survey-style technical analysis tracing the evolution of social engineering up to GenAI-enabled synthetic presence, examining image and voice deepfake generation, documented incidents, and the current state of detection. The central finding is structural rather than quantitative: the assumption that visual and auditory presence proves identity no longer holds. The paper concludes that no single defensive layer is sufficient, and that effective mitigation requires combining technical detection, content provenance frameworks, and process-level verification controls.

Keywords

social engineering, generative AI, deepfakes, voice cloning, phishing, large language models, content provenance, C2PA, cybersecurity

1 Introduction

For many tech-based professionals in the industry, the first encounter with a cyberattack is not a sophisticated intrusion, but rather a simple phishing email. This is because social engineering issues are not abstract threats in the modern business landscape. A single malicious link, clicked by an individual low-level employee in a conglomerate, can disrupt operations across an organization without exploiting a single line of code. LLMs and AI are now making these links, and methods of intrusion, even harder to recognize and defend against. With the new tools contributing to the forced evolution of cyberattacks, security professionals need to consider the constantly changing threat model being presented to employees who might be unaware of the greater risks posed by "just clicking a link."

At the heart of the problem, social engineering is not like a typical cyberattack in the way a defender would try to mitigate the issue. It is all about the psychology of persuasion[8], ensuring that vulnerable users in any environment trust some malicious entity enough to allow for a security breakdown. To defend against a buffer overflow, programmers can utilize secure coding practices or compiler-level protections. Social engineering, on the other hand, requires every employee with access to outside resources in a company to have some degree of training regarding malicious actors and how they might target them individually. In the security landscape, the biggest issue is always the people. Whether it be that security is too expensive or incurs too large an overhead, humans with free choices are always the reason that severe vulnerabilities are created in business environments. And when these security approaches begin to require attention to detail, or supplementary action on behalf of all involved parties, the risk that a malicious actor could manipulate the situation grows even more.

Generative AI did not invent social engineering or the concept of psychological manipulation, but rather has industrialized the manufacture of trust, transforming three classical attack dimensions (realism, personalization, automation) and creating modalities (real-time deepfakes, voice cloning) that defeat the human verification characteristics on which prior defenses have relied. With the increasing desire and usage of generative AI in "safe" business atmospheres, the set of information these models train on grows larger by the day, increasing the risk posed to users. Without an updated approach to security in these times, eventually, social engineering could become such an issue that certain threats will require public acknowledgement and education from a young age, not dissimilar to how media literacy has begun entering formal education frameworks in response to the broader synthetic media crisis[32]. To assist in the mitigation of these emerging threats, this paper aims to contribute a historical taxonomy of social engineering from manual phishing through GenAI-enabled synthetic presence, a consolidated technical analysis of image/facial and voice deepfake generation and detection, and an evaluation of current defenses against current attack capabilities. This survey-style technical analysis aims to collect and combine what the research community knows, organizing it into a coherent narrative that helps to identify where the gaps are in current solutions.

The remainder of this paper is organized as follows. Section 2 establishes the foundations of traditional social engineering, covering its taxonomy and psychological mechanisms utilized by attackers. Section 3 examines how generative AI has transformed the security landscape along three dimensions: realism, personalization, and automation. Sections 4 and 5 analyze the two primary synthetic media threats in depth — image and facial deepfakes, and voice cloning, respectively — including generation techniques, documented incidents, and the current state of detection. Section 6 briefly evaluates current defense strategies, and Section 7 concludes with a summary of contributions and key takeaways.

2 Background: Traditional Social Engineering

Before examining how generative AI has reshaped social engineering, it is necessary to establish what social engineering was before that inflection point. This section traces the foundational pieces of

Technical Report, University of Kansas
2026.

attack types, the psychological principles that make them effective, and the pre-AI attack patterns that defined the threat landscape for decades. This includes the observable "tells" that traditional defenses relied upon and that modern AI has systematically eliminated.

2.1 Definition and taxonomy

Social engineering encompasses a broad family of attack techniques, each exploiting human psychology through a different vector. Establishing this classification is important to the analysis that follows, as the AI-driven threats examined in later sections represent direct evolutions of these classical forms. Table 1 summarizes the primary attack types, organized by their delivery mechanism and the psychological principle they exploit. While these categories are not mutually exclusive, and a single series of attacks may chain several techniques, the distinctions are meaningful because AI amplifies each modality differently. Phishing and spear phishing benefit most from personalization and linguistic realism; vishing and smishing are transformed by voice cloning and automated message generation; pretexting gains from AI's ability to construct and sustain false identities at scale. These attacks are considered in the context of Mouton et al.'s[29] framework, with the idea that AI amplification enables these attacks to be layered in a manner previously difficult to accomplish.

The framework referenced, which expands on Mitnick's Art of Deception[28], decomposes a social engineering attack into six sequential phases: attack formulation, in which the attacker defines the goal and selects a target; information gathering, where data about the target is collected from open sources; preparation, where the attacker synthesizes that information into an attack vector; develop relationship, in which initial communication is established and rapport is built; exploit relationship, where the target is primed emotionally and the desired information or action is elicited; and debrief, where the attacker maintains the target's emotional state to avoid suspicion before either satisfying the goal or transitioning back to information gathering for a follow-up phase. This six-phase structure persists in AI-altered attacks, but generative AI dramatically lowers the cost and effort of specific phases, particularly information gathering, preparation, and the realism achievable during the develop and exploit relationship phases, which enables attack patterns that were previously impractical at scale.

2.2 The psychological medium

Robert Cialdini's Influence: The Psychology of Persuasion[7] establishes six foundational principles of influence that explain how compliance can be obtained in everyday human interaction. These principles form the psychological substrate that social engineering attacks exploit, and as this paper will explore, generative AI has allowed what was once a manual, time-intensive practice to be executed at scale and with greater efficiency. The principles relevant to social engineering are as follows: reciprocity, in which a person given a gift or service feels obliged to repay this debt; liking, where a relationship established with a victim instills a false sense of trust or familiarity; consensus (sometimes called social proof), which convinces a victim that to belong to a greater group or society at large they must believe or act in a certain way; authority, in which victims are led to believe that the individual they are talking to holds some type of power or influence over them; scarcity, where an implied worry of missing out on something or wasting time can cause stress; and consistency, in which a victim's prior commitment, however small, is leveraged to extract compliance with progressively larger requests, exploiting the human desire to act in alignment with one's past behavior. Understanding all of these principles leads to a comprehension of the vulnerabilities people inherently possess and how they can be mitigated.

2.3 The pre-AI attack pattern

The evolution of phishing, which this paper considers the dominant digital vector for social engineering, illustrates what generative AI has actually changed. Liu et al. characterized this evolution as an arms race driven by a single tension they term the Scale–Personalization Conflict: the trade-off between attacks that reach many targets cheaply and attacks tailored enough to succeed against any one of them[24]. They identify four stages of phishing generation, three of which preceded generative AI. The first stage, from the mid-1990s through the early 2000s, relied on manual and templated emails. Phishing kits made this approach simpler, allowing unskilled attackers to launch broad campaigns at success rates below one percent. The defining flaws of this era were grammatical errors, generic greetings, and mismatched URLs, which then became the foundation of decades of user security training. The second stage inverted this strategy. As awareness rose, attackers shifted to spear phishing built on Open-Source Intelligence, achieving success rates of thirty to seventy percent but requiring an average of thirty-four minutes of human labor per email[24]. Only high-value targets justified the investment. The third stage attempted to bridge this divide through algorithmic generation using Markov chains, and later RNN and LSTM-based models, but generative outputs remained too inconsistent to reliably deceive readers, leaving stage-three phishing at five to fifteen percent success. Across all three stages, an attack could be scalable or personalized, but not both. Section 3 examines how that constraint dissolved and led to the fourth stage, which can be characterized as an LLM-driven revolution.

2.4 The "tells" defenders relied upon

As established in the previous subsections, pre-AI social engineering produced characteristic surface artifacts that defenders learned to recognize. First-stage attacks were filled with grammatical errors, awkward phrasing, and generic salutations — all flaws that arose from non-native authorship and the cost constraints of mass production. Second-stage attacks added detectable signatures of their own: suspicious URLs, mismatched anchor text, and poor domain spoofing that did not survive closer inspection. User-facing security education evolved around precisely these tells, training employees to examine sender addresses, hover over links, and treat poorly written messages as presumptively malicious. This training rests on an assumption that generative AI invalidates: that personalized, contextually plausible attacks would be too costly to produce at scale. Sections 4 and 5 examine the modalities through which this assumption breaks down. The stakes of that breakdown are substantial, as even before the capabilities discussed in this paper became widespread, social engineering was responsible for the majority

Table 1: Taxonomy of Common Social Engineering Attack Types

Attack	Delivery	Psychological Principle	Description
Phishing	Email/SMS	Authority, Fear	Mass-distributed fraudulent messages impersonating legitimate entities to harvest credentials or deliver malware.
Spear Phishing	Email	Authority, Liking	Targeted phishing using victim-specific details to increase perceived legitimacy.
Whaling	Email	Authority	Spear phishing directed at senior executives or high-value targets, often intended to get funds or information directly.
Vishing	Voice call	Authority, Urgency	Phone-based impersonation of trusted figures such as banks or IT support.
Smishing	SMS	Urgency, Fear	Phishing conducted via text message, often containing malicious links.
Pretexting	Multi-channel	Liking, Trust	Fabrication of a scenario or false identity to manipulate a victim into disclosing information.
Baiting	Physical/Digital	Curiosity	Luring victims with an appealing offer, such as a free software download or a branded/infected USB drive.

of corporate breaches, with the UK Cyber Security Breaches Survey 2024 attributing over 84% of incidents to socially engineered vectors[13].

3 The Generative AI Inflection Point

At a certain point following the third stage referenced in Section 2.3, algorithmic generation was determined to be insufficient for the tasks it was needed for. There was not widespread adoption of larger models, and the training materials available to adversaries were not advanced enough to further the progress of social engineering attacks effectively; then, in late 2022, a turning point was reached.

3.1 The capability shift

ChatGPT released version 1.0 on November 30, 2022, and in the months that followed, researchers were able to observe a spike in unique social engineering attacks with more than a 135% increase from January to February 2023, per Nageab et al[30]. This increase marked the beginning of a shift in the cybersecurity landscape as a whole. In the first week, ChatGPT boasted 1 million users, but by late 2024, it was able to report 300 million users, with that number climbing. In that same timeframe, nearly 40% of the entire United States adult population reported using generative AI, according to the National Bureau of Economic Research[4]. This many users introducing such a large amount of data means that attackers now have means previously never considered to exploit the very models that have grown successful by aiding working individuals.

3.2 The three amplification dimensions

Generative AI's impact on social engineering can be identified along three dimensions within a framework created by Gonzaga et al[19]: realism, personalization, and automation. Realism through the use of Generative Adversarial Networks (GANs) is becoming more of a problem through applications like DALL-E and Stable Diffusion, which create fake images that are both realistic and personalized; to do this, a GAN utilizes multiple neural networks to generate an image, then discriminates against the counterfeit to make the image as real as possible. This work with GANs means that in some cases, generated content is indistinguishable from human-made text, images, audio, or video. The personalization dimension's greatest advancements stem from the union of Open Source Intelligence gathering (OSINT) and generative AI. Where stage-two spear phishing required an average of twenty-three minutes of manual intelligence gathering per target[24], LLM-driven systems can now intake a target's digital footprint and produce an attack in seconds, collapsing the scale-to-personalization trade-off that defined the earlier evolutionary stages of social engineering. The automation dimension completes the framework. Generative AI enables attack infrastructures requiring minimal human intervention, from LLM-generated content delivered through automated pipelines to chatbots sustaining real-time conversations with victims[19]. Mishra et al. demonstrate the accessibility of such automation directly, showing that novice users with no prior offensive experience can execute a complete phishing campaign — from generating a fraudulent login page to harvesting credentials — using only ChatGPT-4o Mini and basic jailbreaking techniques[27]. Gonzaga et al. further note the emergence of LLMs specifically architected for offensive purposes, such as WormGPT, which removes even the jailbreaking step and provides what they call "Jailbreaking-As-A-Service"[19]. Together, the three dimensions constitute a shift in what social engineering can accomplish. Table 2, as reported in Liu et al., illustrates this convergence empirically. LLM-generated phishing emails match or exceed human expert performance across open, click, and credential-input rates while reducing per-email generation time by 92%, from thirty-four minutes to under three[24]. The scale versus personalization trade-off that constrained social engineering for three decades has effectively been resolved.

3.3 Economic framing

The progressing dissolution of the technical barrier previously making social engineering a more time-consuming task has reshaped the threat landscape economically as well as operationally. The FBI's 2024 Internet Crime Report recorded $16.6 billion in losses across 859,532 complaints, a 33% increase from the prior year, with phishing representing the most frequently reported crime type[17].

Table 2: Comparison of LLM-Generated vs. Human Expert-Written Phishing Emails. Bold indicates LLM superiority. Cred = Credential Input Rate. Adapted from Roy et al. (2024) and Bethany et al. (2025), as reported in Liu et al.[24]

Scenario	Type	Open (%)	Click (%)	Cred (%)	Time (min)
Manager to Subordinate	Human	61.97	21.67	10.64	34
	LLM	61.68	21.34	**11.07**	**2.7**
Timely Phishing	Human	36.08	6.66	2.51	34
	LLM	**40.37**	**10.34**	**4.19**	**2.7**
Avg. Improvement		+2.7%	+17.1%	+19.3%	-92%

The 2025 report found losses had climbed further to $20.87 billion across 1,008,597 complaints, a 26% increase, with phishing retaining its position as the most reported crime[18]. The year-over-year trajectory validates the concern raised by Gonzaga et al., who project losses from generative AI-driven fraud to quadruple by 2027, reaching an expected $40 billion at an annual growth rate exceeding 30%[19]. These figures reflect not only the increasing complexity of individual attacks but the industrialization of social engineering as a practice. The losses accumulated thus far illustrate how generative AI's impact extends beyond text-based attacks into synthetic media tools that defeat human defenses more fundamentally — through image and facial deepfakes, and voice cloning. The following sections examine each in turn.

4 Image and Facial Deepfakes

When considering the synthetic media modalities enabled by generative AI, image and facial deepfakes represent one of the most documented and rapidly accelerating threat vectors. Unlike text-based attacks, which exploit grammatical plausibility, visual deepfakes target a more basic human assumption: that seeing is believing. This section examines the technical mechanisms behind image and facial deepfake generation, analyzes a landmark real-world incident that demonstrates their operational impact, and evaluates the current state of detection, including both its successes and its limitations.

4.1 Generation: GANs to diffusion

Early deepfakes relied predominantly on Generative Adversarial Networks, first introduced by Goodfellow et al. in 2014, in which a generator and discriminator network compete iteratively, with the generator producing synthetic images and the discriminator learning to distinguish them from real ones. This is repeated until the generator's outputs become indistinguishable[20]. While effective, GAN-generated material suffered from training instability, mode collapse, and limited output diversity. These outputs were produced with identifiers like irregular eye blinking, boundary inconsistencies, or unnatural lighting — all things that detection systems learned to exploit[19, 33]. Diffusion models, introduced in 2015 and refined through architectures like Stable Diffusion and Imagen, addressed these limitations by learning to reverse a noise-addition process rather than relying on competition[33]. Quantitative comparisons confirm the improvement, as shown in Table 3. Fréchet Inception Distance (FID), which measures how closely generated images match the statistical distribution of real photographs, drops from 68.1 for early GAN architectures to 7.27 for Stable Diffusion, a nearly tenfold improvement in realism. Inception Score (IS), which captures both individual image quality and output diversity, climbs from 6.2 to 38.0 across the same progression. Importantly, Stable Diffusion achieves these results at generation speeds of approximately 0.5 seconds per image, making high-quality synthetic face generation accessible without specialized hardware[33]. For social engineering, this shift matters because diffusion-generated synthetic images carry a different statistical signature than GAN outputs — one that detectors trained on GAN artifacts do not reliably recognize. The accessibility of these tools has lowered the barrier to producing convincing fake identities to the point where a functional deepfake can be constructed from available footage with no specialized hardware or expertise. The consequences of this accessibility are illustrated by a documented incident from Hong Kong in January 2024.

4.2 Case study: the Arup incident

The practical consequences of advancements in social engineering via deepfakes are no longer theoretical. One of the most costly intrusions occurred in January 2024 when a finance employee at the UK-based engineering firm Arup was deceived into authorizing fifteen wire transfers totaling approximately $25.6 million after attending a video conference in which every other participant — including the company's CFO and several colleagues — was an AI-generated deepfake constructed from publicly available footage including LinkedIn videos and recordings of virtual meetings[16, 25]. The employee had initially suspected the meeting to be a phishing attempt when they received an email requesting a "secret transaction," but the visual confirmation of familiar faces and voices overrode their skepticism entirely. Arup's CIO Rob Greig later characterized the incident as "technology-enhanced social engineering," noting that none of the company's systems were compromised, but the attack still succeeded entirely through the impersonation of human identity. Greig also noted that "It wasn't even a cyberattack in the purest sense"[16]. This characterization reflects a critical misunderstanding of the new generation of social engineering: it was a cyberattack in the purest sense, and it did precisely what all effective attacks do — exploit the human variable. Whether through a line of code enabling intrusion or simply insufficient training, the attack was made possible by technology that was considered science fiction two decades ago. The Arup incident illustrates the limitations of human detection capabilities when faced with high-quality synthetic

Table 3: Performance comparison of GAN-based and diffusion-based image generation models. FID (lower is better), IS (higher is better). Adapted from Xiao (2025)[33]

Model	Type	FID↓	IS↑
DCGAN	GAN	68.1	6.2
StyleGAN	GAN	18.3	9.1
DDPM	Diffusion	7.9	15.0
Guided Diffusion	Diffusion	7.3	28.0
Stable Diffusion	Diffusion	7.27	—
Imagen	Diffusion	7.27	38.0

media. The following section examines the technical approaches developed to detect such content automatically, and the degree to which defenders have succeeded in closing this gap.

4.3 Detection: state of the art

The technical response to deepfake advancements has produced a rapidly evolving research component of social engineering defense. Damodar and Manek (2025) surveyed the field across five major architectural categories, covering convolutional neural networks, recurrent and transformer-based models, multimodal fusion frameworks, frequency-domain analyses, and hybrid approaches. They reported detection accuracies ranging from 70% on the most challenging benchmarks to 99.81% on controlled datasets[12]. The size of this range reflects two truths: detection performance under favorable conditions has become impressive, but said performance is sensitive to the conditions under which detectors are trained and evaluated.

The current state of the defense against visual deepfakes is best represented by hybrid architectures that combine complementary feature types. Aribe (2025) proposes a hybrid framework combining forensic features — Photo Response Non-Uniformity noise residuals, JPEG compression traces, and Discrete Cosine Transform frequency descriptors — with deep learning representations from ResNet-50 and Vision Transformer. Evaluated on three benchmark datasets, this hybrid model achieved F1 scores of 0.96 on FaceForensics++, 0.82 on Celeb-DF v2, and 0.77 on the DeepFake Detection Challenge (DFDC) dataset, outperforming forensic-only, CNN-only, and transformer-only baselines across all conditions[22]. Bhattacharjee et al. take a parallel approach with their proposed CAE-Net, a weighted ensemble combining EfficientNet-B0 for local feature extraction, a Data-Efficient Image Transformer (DeiT) for global features, and ConvNeXt with wavelet-based preprocessing to detect hidden frequency patterns. Validated with the IEEE Signal Processing Cup 2025 dataset, which aggregates eight benchmark deepfake datasets to test generalization, CAE-Net achieved 94.46% accuracy and 97.60% AUC, outperforming prior ensemble approaches[3]. The shared finding across both papers is that no single feature type — spatial, temporal, frequency, or forensic — is sufficient on its own; effective detection requires combining these signals.

These benchmark numbers, however, are obtained under conditions that are increasingly separated from the threat landscape attackers exploit. Abdullah et al. evaluate eight state-of-the-art detectors against sixteen user-customized variants of Stable Diffusion, all models created by internet users through low-cost fine-tuning techniques such as Low-Rank Adaptation (LoRA), and freely distributed on platforms like CivitAI and HuggingFace[1]. All eight detectors exhibited significant performance degradation, with average recall drops from 19.69% to 53.92% across the sixteen variants. CNN-based defenses generalized worst, and frequency-domain features generalized best, but even the strongest detector lost nearly a fifth of its recall when applied to generators it was not trained on. Abdullah et al. additionally demonstrate that an attacker leveraging publicly available vision foundation models can create deepfakes that degrade detector performance by as much as 88.35% without adding visible noise to the image[1].

4.4 Limits of detection alone

Social engineering in the past has been proactively defended against through cybersecurity training and employee awareness programs. Deepfake detection is, by nature, a reactive defense — an arms race that detection will continue to lose while adversaries further develop malicious image generation models. Chandra et al. demonstrate that state-of-the-art open-source detectors experienced AUC reductions of 50% for video, 48% for audio, and 45% for image-based models when tested against current deepfakes rather than older academic datasets. The authors further observe that many existing benchmarks no longer reflect real-world threats due to the rapid advancement of generative models[6]. Similarly, Richings et al. show that detectors trained on contemporary deepfake datasets experience significant performance decay when evaluated on newer content, with recall dropping by over 30% when models are tested on deepfakes generated only six months after their training data[31]. These findings emphasize that deepfake detection is not a single-solution problem and requires continuous retraining to maintain effectiveness as synthetic media evolves.

This problem has motivated interest in provenance-based approaches. Post-hoc detection refers to identifying fake media after it has been created and distributed, based solely on observable content. Provenance, by contrast, refers to verifiable records of a media piece's origin and modification history, enabling an authenticity check at the time of capture. Standards such as the Coalition for Content Provenance and Authenticity (C2PA) framework aim to embed such verifiable metadata into digital media, shifting trust from content analysis to documented origin[9]. However, provenance systems are primarily designed for controlled capture and editing environments, and thus carry limitations further addressed in Section 6. As a result, vulnerabilities remain across other modalities — in particular, audio manipulation represents a parallel attack

surface with similar challenges in detection. This motivates Section 5, which examines deepfake generation and detection in the audio domain.

5 Audio and Voice Cloning Deepfakes

While similar to synthetic visual media, audio and voice cloning deepfakes represent an entirely different threat as it pertains to social engineering. Audio deepfakes target the same rational assumption that visual deepfakes do: that a human will trust what they hear if they recognize the voice. These synthetic audios operate through a channel that can carry additional context. A familiar voice conveys not just identity but authority, urgency, and relationship context in ways that text and even video cannot replicate with the same immediacy. Where a suspicious email can be ignored or a deepfake video can prompt a second look, a phone call from what sounds like a known employer or family member causes a different response. Since people often believe they would be able to tell the difference between someone they know and an impersonation of their voice, an element of overconfidence occurs. Ebert et al., shown in Table 4, identify this bias through detection measurements in their controlled voice phishing experiment where 91% of participants rated a synthesized message as having "high credibility" even when informed in advance that it was artificially generated[14]. This shows that when presented with a voice alone, people often lack the perception required to establish appropriate skepticism, and are thus more likely to make rash decisions when presented with a deepfake call utilizing a sense of urgency. The following subsections examine how voice cloning technology has developed, the documented record of its use in fraud, and the state of current detection methods.

Table 4: Human Detection Performance Against AI-Generated Audiovisual Content [14]

Metric	Result
Voice Deepfake (n=56, ElevenLabs, prior warning)	
Found message credible	91% (51/56)
Perceived as untrustworthy	9% (5/56)
Video Deepfake (n=34, DeepLiveCam, no warning)	
Failed to detect manipulation	74% (25/34)
Detected manipulation	26% (9/34)
Mean authenticity rating	4.3 / 5

Note: Voice study used binary credibility; video study used a 1–5 authenticity scale.

5.1 Generation: TTS to neural cloning

Early voice generation systems relied on concatenative TTS (text-to-speech), splicing pre-recorded phoneme segments to produce recognizable robotic output that was easily distinguishable from human speech. The shift to neural architectures changed this fundamentally. Autoregressive models such as WaveNet and Tacotron introduced learned speech generation, converting text to mel-spectrograms and then to audio waveforms through neural vocoders, eliminating the phoneme-splicing artifacts of earlier systems[23]. GAN-based vocoders such as HiFi-GAN further improved upon synthetic audio. Voice conversion (VC) approaches allowed attackers to take existing speech and remap its acoustic characteristics to match a target speaker's voice, preserving the content of the speech while substituting the target identity[23]. The most recent generation of systems has collapsed this pipeline entirely. VALL-E, Microsoft's neural codec language model, treats TTS as a conditional language modeling task rather than a signal regression problem, generating discrete audio codec tokens directly from phoneme input and a short reference sample. VALL-E is trained using a language modeling objective comparable to large language models, enabling prompt-based voice generation from minimal reference audio[2]. This evolution has produced convincing voice cloning from as little as three seconds of source audio, requiring no fine-tuning on the target speaker and no specialized hardware. For defenders, auditory social engineering presents a unique challenge: unlike text or image deepfakes, a convincing voice clone requires no visual inspection, no link to click, and no document to examine, meaning verification is entirely dependent on the channel itself.

5.2 Why voice proves harder to verify than text

The verification problem with voice deepfakes is structural rather than perception-based. With text-based phishing, recipients have access to inspectable signals such as sender domains, embedded URLs, or formatting inconsistencies — all things that security training can teach them to examine. With image deepfakes, boundary artifacts, lighting inconsistencies, and unnatural eye movement provide at least some observable indicators. Voice provides none of these. A phone call offers no equivalent of "hover over the link before clicking," since the channel itself carries no inspectable data visible to the recipient, and the interaction is brief by design. This structural gap extends to automated systems as well. Speaker verification systems deployed by banks and corporate phone systems were designed to confirm identity through vocal characteristics. Voice cloning defeats them directly: modern generation systems replicate not just pitch and tone but the acoustic features that speaker verification models treat as identity markers, meaning a cloned voice can pass both human and automated authentication simultaneously. A large-scale study of over 1,200 participants found that while people apply intuition and verbal cues when evaluating audio, overall detection performance remains poor in complex scenarios, with existing automated detection systems not consistently outperforming human judgment[35]. The verification layer organizations have built around voice as an authentication factor has thus been invalidated by the same technology that makes the attack possible. The only reliable verification approach that remains is out-of-band confirmation: ending the call and contacting the caller through a separately verified channel such as a known number or in-person confirmation. The practical obstacle is that the same Cialdini principles the call uses to pressure victims — urgency, authority, and the social pressure of a live interaction — make hanging up to verify feel inappropriate or unnecessary.

5.3 Documented incidents

The real-world record of voice deepfake fraud reflects both the maturity of the technology and its accelerating deployment as a social engineering tool. Table 5 documents a selection of confirmed incidents spanning 2019 to 2025. Several patterns emerge from this record. First, the financial losses per incident have grown substantially, from $243,000 in the earliest documented case to $25.6 million in the 2024 Arup attack, reflecting improvements in generation quality that allow attackers to sustain more elaborate deceptions over longer interactions. Second, the two failed attacks in 2024, against LastPass and WPP, demonstrate that skepticism still works as a defense, but both failures were attributed to the target's prior channel awareness rather than any technical detection capability. Third, the 2025 Florida incident reflects a shift in the threat surface. That attack required no deepfake video, no coordinated multi-participant call, and no corporate intelligence gathering — just a brief voice clone constructed from social media audio and a phone call lasting minutes. This trajectory mirrors the technical evolution described in Section 5.1: as generation barriers fall, the threat does not simply scale upward, but outward toward the general public, where victims have fewer resources, less awareness, and no incident response training.

5.4 Detection: spectral and behavioral signals

The detection of synthetic audio relies primarily on spectral feature analysis. Classical approaches extract Mel-Frequency Cepstral Coefficients (MFCCs) or Linear Frequency Cepstral Coefficients (LFCCs) from audio waveforms and classify them using convolutional or recurrent architectures, with the ASVspoof challenge series serving as the primary benchmark for evaluating countermeasure performance[23, 35]. The limitations of this detection methodology were documented as early as 2023, when Mai et al. found that humans correctly identified speech deepfakes only 73% of the time in controlled conditions, and that out-of-domain automated detectors performed worse than human listeners, suggesting that neither perception-based nor computational defenses were reliable against generation methods that were not yet state-of-the-art[26]. State-of-the-art systems combining self-supervised frontends such as WavLM with graph attention network backends have since achieved Equal Error Rates (EER) below 3% on controlled ASVspoof benchmark datasets[35]. However, when evaluated on the In-the-Wild dataset, which contains audio recorded outside controlled conditions, the same leading models exhibit EERs ranging from approximately 9% to over 27% depending on training configurations[35]. The problem is identical to the one documented for image deepfake detectors in Section 4.3: detectors trained on known synthesis methods generalize poorly when attackers use newer generation methods, with each new synthesis architecture effectively resetting the detection baseline.

6 Toward Defense: Emerging Approaches and Limitations

The limitations documented in Sections 4 and 5 suggest that no single defensive layer is sufficient against the current threat landscape of social engineering. Technical detection faces the generalization problem identified by Abdullah et al. and the time-based decay established across both the visual and audio domains[1]. Behavioral training, which can be seen as the primary defense against social engineering, faces an equally central problem. Mai et al.'s controlled experiment found that familiarizing participants with examples of speech deepfakes improved detection accuracy by less than 4%, a gain so small that it suggests awareness training cannot compensate for the human limitations these attacks exploit[26]. Eberl et al. identify three cognitive biases that explain this mechanism precisely: authenticity bias, the human tendency to accept visual and auditory information as genuine unless it is obviously fabricated; overconfidence in detection, where individuals overestimate their ability to identify manipulated media; and confirmation bias, where deepfakes that align with a victim's expectations are accepted without scrutiny[14]. These biases are not fixable through awareness alone because they operate as cognitive shortcuts rather than conscious decisions. Generative AI does not just make attacks harder to spot — it produces content engineered to manipulate the human sense of trust. Training remains a necessary first layer, given that human error contributes to the vast majority of security incidents[19], but it cannot be the only layer.

Different types of deepfakes require different solutions. For synthetic image generation, the inadequacy of post-hoc detection has motivated provenance-based approaches as a complementary layer. The C2PA framework, briefly touched on in Section 4.4, provides this by embedding cryptographically signed metadata into media at the moment of capture, creating a tamper-evident chain from sensor to display that allows recipients to verify not whether content appears authentic, but where it actually originated[9]. This conceptual shift from "is this fake?" to "where did this come from?" is illustrated in Figure 1, which contrasts post-hoc inspection with provenance verification. Adoption has grown among hardware and platform stakeholders, in large part thanks to the Content Authenticity Initiative. This collective includes camera manufacturers like Leica, Sony, and Canon embedding Content Credentials at capture, and platforms including Adobe, Microsoft, LinkedIn, and Cloudflare implementing verification support at the display layer. Additionally, Samsung launched the Galaxy S25 with native C2PA support, the first smartphone to do so[10]. However, provenance systems carry limitations that prevent them from serving as a complete solution. C2PA is strongest in controlled media environments where content moves directly from a capture device through verified editing tools to a compliant display platform. In open environments — where media is passively recorded, re-shared across platforms that strip metadata, or captured in real time — the provenance chain is absent by design. Solutions like watermarking face similar constraints, as embedded signals can be stripped through compression and re-encoding.

For the audio realm, no equivalent provenance infrastructure currently exists at scale. The structural verification problem identified in Section 5.2 — that voice carries no inspectable metadata visible to the recipient — means that audio deepfake defense cannot follow the same provenance model as image defense. Proactive approaches such as AntiFake attempt to disrupt voice cloning before synthesis occurs by embedding imperceptible adversarial perturbations into source audio that degrade speaker identity extraction, achieving over 95% protection rates against modern synthesizers including commercial black-box models[34]. These systems show promise in

Table 5: Selected Documented Voice Deepfake Fraud Incidents

Year	Target	Method	Outcome
2019	UK energy firm	Voice clone of parent company CEO via phone call	$243,000 lost [11]
2020	Japanese firm (HK)	AI voice synthesis + email spoofing impersonating director of parent company	$35M lost [5]
2024	Arup (HK)	Multimodal deepfake video call with synthesized voices	$25.6M lost [16, 25]
2024	LastPass	Audio deepfake of CEO via WhatsApp	Attack failed [36]
2024	WPP	Voice clone of CEO + real YouTube footage used in fake meeting	Attack failed [21]
2025	US parent	AI voice clone of daughter in distress	$15,000 lost [15]

controlled pipelines but share C2PA's core limitation: they require defenders to control the original audio before it is exposed, which is impossible for voice data already available through social media, public recordings, or prior phone calls. For audio social engineering specifically, the absence of a provenance layer means the defense burden falls entirely on the process layer — the organizational protocols governing how identity is verified during interactions. The out-of-band callback that would have stopped the Arup attack, the channel awareness that stopped the LastPass intrusion, and the family verification codes recommended by cybersecurity experts in response to personal scams such as the Florida parent case all reflect the same underlying principle: when the channel itself cannot be trusted, verification must occur through one that can. This is the lesson that synthetic media social engineering forces onto defenders across both modalities. The requirement is not simply that detection must improve, but that verification cannot depend on the integrity of any single channel.

7 Conclusion

This paper set out to examine what generative AI has actually changed about social engineering — specifically the concrete transformation of the threat that security professionals, organizations, and individuals face today. The evidence assembled across Sections 2 through 6 supports a single conclusion: generative AI has resolved the scale-to-personalization conflict that hindered social engineering for decades, producing attack capabilities that are simultaneously broad, targeted, and convincing in ways that prior defenses were not designed to address. With the FBI's Internet Crime Report recording increasing losses across more complaints — a trajectory that Gonzaga et al. project will reach $40 billion by 2027[18, 19] — these figures are not just the product of more attacks. They reflect attacks that are harder to recognize, harder to detect automatically, and harder to defend against institutionally, because the friction that made them costly to produce is now irrelevant.

The deeper finding of this paper is structural, not quantitative. The premise on which human verification has always rested — that visual and auditory presence proves identity — no longer holds. Diffusion models generate synthetic faces that defeat detectors trained on GAN artifacts. Voice cloning systems replicate the acoustic features that speaker verification models treat as identity markers. And the recipients of these attacks are not failing because they are

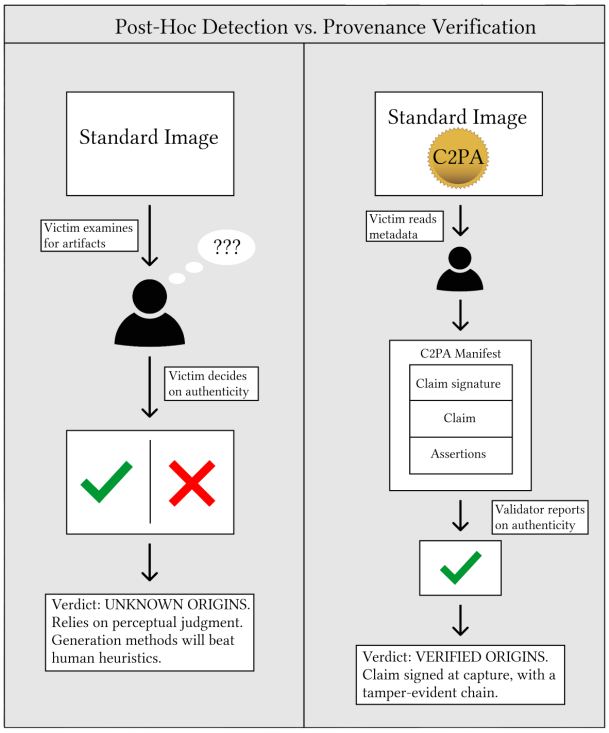

Figure 1: Post-hoc detection requires the recipient to identify generative artifacts after the fact. Provenance verification via C2PA alters this by attaching signed metadata at the moment of capture (Author-created via Figma).

untrained; they are failing because the attacks are engineered to exploit the fundamental mechanisms of trust. Eberl et al.'s finding that 91% of participants rated a synthesized voice as credible even when explicitly warned it was artificial demonstrates that authenticity bias, overconfidence in detection, and confirmation bias operate almost imperceptibly[14]. Training is therefore necessary, but it is not sufficient.

The layered defense posture described in Section 6 — combining technical detection, provenance verification through frameworks like C2PA, and process controls that do not depend on any single channel — represents the most realistic path available given current capabilities. None of these layers alone closes the gap. Detection generalizes poorly across synthesis methods. Provenance verification requires controlled media pipelines that open communication environments lack. Process controls require infrastructure that individuals do not have. The 2025 Florida case[15] illustrates where the threat surface has arrived: attacks requiring no specialized hardware, no corporate intelligence, and no technical expertise, targeting people with no training and no incident response capabilities. Simple phishing email attacks have not disappeared, but they are no longer representative of the frontier. Attacks now do not announce themselves with suspicious links or awkward phrases; they arrive as a familiar voice, a known face, or a request that feels entirely plausible.

References

[1] Sifat Muhammad Abdullah, Aravind Cheruvu, Shravya Kanchi, Taejoong Chung, Peng Gao, Murtuza Jadliwala, and Bimal Viswanath. 2024. An Analysis of Recent Advances in Deepfake Image Detection in an Evolving Threat Landscape. In *2024 IEEE Symposium on Security and Privacy (SP)*. 91–109. doi:10.1109/SP54263.2024.00194

[2] Hussam Azzuni and Abdulmotaleb El Saddik. 2025. Voice Cloning: Comprehensive Survey. arXiv:2505.00579 [cs.SD] https://arxiv.org/abs/2505.00579

[3] Anindya Bhattacharjee, Kaidul Islam, Kafi Anan, Ashir Intesher, Abrar Assaeem Fuad, Utsab Saha, and Hafiz Imtiaz. 2026. CAE-Net: Generalized deepfake image detection using convolution and attention mechanisms with spatial and frequency domain features. *Journal of Visual Communication and Image Representation* 115 (Jan. 2026), 104679. doi:10.1016/j.jvcir.2025.104679

[4] Alexander Bick, Adam Blandin, and David J. Deming. 2024. *The Rapid Adoption of Generative AI.* NBER Working Paper 32966. National Bureau of Economic Research. https://www.nber.org/system/files/working_papers/w32966/w32966.pdf

[5] Thomas Brewster. 2021. *Fraudsters Cloned Company Director's Voice In $35 Million Heist, Police Find.* Forbes. https://www.forbes.com/sites/thomasbrewster/2021/10/14/huge-bank-fraud-uses-deep-fake-voice-tech-to-steal-millions/

[6] Nuria Alina Chandra, Ryan Murtfeldt, Lin Qiu, Arnab Karmakar, Hannah Lee, Emmanuel Tanumihardja, Kevin Farhat, Ben Caffee, Sejin Paik, Changyeon Lee, Jongwook Choi, Aerin Kim, and Oren Etzioni. 2025. Deepfake-Eval-2024: A Multi-Modal In-the-Wild Benchmark of Deepfakes Circulated in 2024. arXiv:2503.02857 [cs.CV] https://arxiv.org/abs/2503.02857

[7] Robert B. Cialdini. 2006. *Influence: The Psychology of Persuasion* (revised ed.). Harper Business.

[8] Cisco Systems. 2026. What Is Social Engineering? https://www.cisco.com/site/us/en/learn/topics/security/what-is-social-engineering.html. Accessed: 2026-05-09.

[9] Coalition for Content Provenance and Authenticity (C2PA). 2025. C2PA Specification Version 2.4. https://spec.c2pa.org/specifications/specifications/2.4/specs/C2PA_Specification.html. Accessed: 2026-05-10.

[10] Content Authenticity Initiative. 2025. *5,000 members: building momentum for a more trustworthy digital world.* Content Authenticity Initiative. https://contentauthenticity.org/blog/5000-members-building-momentum-for-a-more-trustworthy-digital-world

[11] Jesse Damiani. 2019. *A Voice Deepfake Was Used To Scam A CEO Out Of $243,000.* Forbes. https://www.forbes.com/sites/jessedamiani/2019/09/03/a-voice-deepfake-was-used-to-scam-a-ceo-out-of-243000/

[12] Navya Damodar and Asha S Manek. 2025. Advances in detecting deepfake threats, methods and societal implications. In *2025 6th International Conference on IoT Based Control Networks and Intelligent Systems (ICICNIS)* (Bengaluru, India). IEEE, 2098–2104.

[13] Department for Science, Innovation and Technology and Home Office. 2024. *Cyber Security Breaches Survey 2024.* Official Statistics. UK Government. https://www.gov.uk/government/statistics/cyber-security-breaches-survey-2024/cyber-security-breaches-survey-2024 Accessed: 2026-05-09.

[14] Lara Eberl, Lisa-Marie Engländer, Caroline Löhle, Dennis Jahnecke, Aylin Baris, Donjeta Seljaci, Schabnam Shamsi, and Jochen Günther. 2025. Phishing and Identity Manipulation through Audiovisual Channels. In *Open Identity Summit 2025.* Gesellschaft für Informatik e.V., Bonn, 101–112. doi:10.18420/oid2025_07

[15] Gabriella Egozi. 2025. *Florida woman swindled into giving $15K after AI clones daughter's voice.* NBC 6 South Florida. https://www.nbcmiami.com/news/local/florida-woman-swindled-into-giving-15k-after-ai-clones-daughters-voice/3660507/

[16] David Elliott. 2025. *'This happens more frequently than people realize': Arup chief on the lessons learned from a $25m deepfake crime.* World Economic Forum. https://www.weforum.org/stories/2025/02/deepfake-ai-cybercrime-arup/

[17] Federal Bureau of Investigation. 2025. *2024 IC3 Annual Report.* Technical Report. Internet Crime Complaint Center (IC3). https://www.ic3.gov/AnnualReport/Reports/2024_IC3Report.pdf

[18] Federal Bureau of Investigation. 2026. *2025 IC3 Annual Report.* Technical Report. Internet Crime Complaint Center (IC3), U.S. Department of Justice. https://www.ic3.gov/AnnualReport/Reports/2025_IC3Report.pdf

[19] Kely Gonzaga, Sérgio Serra, Marco Gomes, and Silvestre Malta. 2026. AI-powered Social Engineering: Emerging attack vectors, vulnerabilities, and multi-layered defense strategies. *Computers* 15, 2 (Feb. 2026), 128.

[20] Ian J. Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. 2014. Generative Adversarial Networks. arXiv:1406.2661 [stat.ML] https://arxiv.org/abs/1406.2661

[21] James Hall. 2024. *CEO of WPP targeted in deepfake scam attempt.* The Guardian. https://www.theguardian.com/technology/article/2024/may/10/ceo-wpp-deepfake-scam

[22] Sales Aribe Jr. 2025. A Hybrid Deep Learning and Forensic Approach for Robust Deepfake Detection. *International Journal of Advanced Computer Science and Applications* 16, 10 (2025). doi:10.14569/ijacsa.2025.0161028

[23] Menglu Li, Yasaman Ahmadiadli, and Xiao-Ping Zhang. 2025. A Survey on Speech Deepfake Detection. arXiv:2404.13914 [cs.SD] doi:10.1145/3714458

[24] Zhouyang Liu, Yanli Chen, Yuyu He, Zhigang Wang, Hui Lu, Xinge Zhang, and Jinghang Wu. 2025. An arms race in the inbox: A systematic review of phishing generation and the rise of LLMs. In *2025 IEEE 10th International Conference on Data Science in Cyberspace (DSC)* (Baoding, China). IEEE, 94–101.

[25] Kathleen Magramo. 2024. *British engineering giant Arup revealed as $25 million deepfake scam victim.* CNN. https://www.cnn.com/2024/05/16/tech/arup-deepfake-scam-loss-hong-kong-intl-hnk

[26] Kimberly T Mai, Sergi Bray, Toby Davies, and Lewis D Griffin. 2023. Warning: Humans cannot reliably detect speech deepfakes. *PLoS One* 18, 8 (Aug. 2023), e0285333.

[27] Rina Mishra, Gaurav Varshney, and Shreya Singh. 2025. Jailbreaking Generative AI: Empowering Novices to Conduct Phishing Attacks. In *2025 55th Annual IEEE/IFIP International Conference on Dependable Systems and Networks - Supplemental Volume (DSN-S).* 251–252. doi:10.1109/DSN-S65789.2025.00022

[28] Kevin D. Mitnick and William L. Simon. 2002. *The Art of Deception: Controlling the Human Element of Security.* Wiley.

[29] Francois Mouton, Mercia M Malan, Louise Leenen, and H S Venter. 2014. Social engineering attack framework. In *2014 Information Security for South Africa* (Johannesburg, South Africa). IEEE.

[30] Walaa Mohamed Nageab, Rashed Alrasheed, and Mahmoud Khalifa. 2024. Cybersecurity in the Era of Artificial Intelligence: Risks and Solutions. In *2024 ASU International Conference in Emerging Technologies for Sustainability and Intelligent Systems (ICETSIS).* 240–245. doi:10.1109/ICETSIS61505.2024.10459584

[31] Jack Richings, Margaux Leblanc, Ian Groves, and Victoria Nockles. 2025. Performance Decay in Deepfake Detection: The Limitations of Training on Outdated Data. arXiv:2511.07009 [cs.CV] https://arxiv.org/abs/2511.07009

[32] Jennifer Vilcarino. 2026. Schools Play Game of Media Literacy Catch-Up as AI Use Rises. Education Week. https://www.edweek.org/technology/schools-play-game-of-media-literacy-catch-up-as-ai-use-rises/2026/04 Accessed: 2026-05-09.

[33] Yian Xiao. 2025. From Gans to Diffusion Models: Text-to-image generation. *Highlights in Science, Engineering and Technology* 160 (Dec. 2025), 80–87.

[34] Zhiyuan Yu, Shixuan Zhai, and Ning Zhang. 2023. AntiFake: Using Adversarial Audio to Prevent Unauthorized Speech Synthesis. 460–474. doi:10.1145/3576915.3623209

[35] Bowen Zhang, Hui Cui, Van Nguyen, and Monica Whitty. 2025. Audio deepfake detection: What has been achieved and what lies ahead. *Sensors (Basel)* 25, 7 (March 2025), 1989.

[36] Steve Zurier. 2024. *LastPass thwarts attempt to deceive employee with deepfake audio.* SC Media. https://www.scworld.com/news/lastpass-thwarts-attempt-to-deceive-employee-with-deepfake-audio