JACoIT
Journal of Applied Computer and Information Technology

GLOBAL RESEARCH INNOVATION

Review

Beyond the human firewall: A systematic analysis of deepfake-mediated social engineering and the erosion of traditional security awareness

Mohd Ruhaifi Zainol [1,*], Marhakim Mokhtar [2] and Mohamad Fadli Zolkipli [3]

[1]Albukhary International University, Alor Setar, Malaysia; ruhaifizainol@gmail.com
[2]Politeknik Tuanku Syed Sirajuddin, Perlis, Malaysia; marhakimm@gmail.com
[3]School of Computing, Universiti Utara Malaysia, Sintok, Malaysia; m.fadli.zolkipli@uum.edu.my

*Corresponding Author: ruhaifizainol@gmail.com

Received: 19 April 2026
Revised: 17 June 2026
Accepted: 24 June 2026
Published: 4 August 2026

Abstract

The proliferation of generative artificial intelligence has transformed the cyber threat landscape, particularly in the domain of social engineering. Deepfake technology, encompassing synthetic audio and video generation, has emerged as a vector for advanced phishing campaigns that can bypass conventional controls, exposing limitations of traditional Security Awareness Training (SAT) when confronted with AI-driven deception. This paper presents a systematic analysis of empirical evidence on deepfake-enabled social engineering and its implications for existing security-awareness frameworks, drawing on 56 primary studies involving 86,155 participants, supplemented by 47 documented corporate incidents and a review of current technical detection methods. Pooled human detection accuracy for deepfake content was 55.54% (95% CI [52.3%, 58.8%]), only marginally above chance; the pooled estimate is, however, accompanied by substantial heterogeneity (I^2 = 78.4%) and should be interpreted as an average performance rather than a uniform inability to discriminate. Traditional SAT was associated with a non-significant +1.6% improvement in deepfake detection, whereas SAT incorporating synthetic-media examples produced significant gains of +8.1% to +15.5%. The study identifies three persistent limitations in current awareness training: the absence of deepfake-specific detection heuristics, inadequate calibration of trust in response to synthetic authority cues, and insufficient inoculation against cognitive-load manipulation; a 21.1-percentage-point laboratory-to-field performance gap was also observed. The paper proposes a resilience-oriented training framework that integrates technical literacy, psychological preparedness, and organisational verification mechanisms. Rather than declaring the "human firewall" obsolete, the analysis argues for its reconceptualisation as a complementary safeguard within layered, procedure-anchored defence.

Keywords: deepfake; social engineering; security awareness training; synthetic media; phishing

1. Introduction

The cybersecurity landscape has changed with the commoditisation of generative artificial intelligence. Deepfake systems based on generative adversarial networks (GANs) and autoencoders have moved from academic prototypes to widely accessible tools capable of producing hyper-realistic synthetic audio and video [1]. This democratisation has introduced new attack vectors within social engineering, allowing adversaries to mount impersonation attacks with high fidelity and scale.

Social engineering, the psychological manipulation of individuals to subvert security controls, has long exploited cognitive biases such as authority, social proof and urgency [2]. Conventional phishing relies on textual cues and static imagery; deepfake-mediated attacks instead present dynamic, personalised audiovisual content that does not present the linguistic and visual artefacts on which traditional Security Awareness Training (SAT) is built [3], [4]. Rather than restating this point across the introduction, we treat it as the single anchoring premise of the study: the threat surface has shifted from text-based artefacts to synthetic media, and SAT designed for the former is unlikely to transfer fully to the latter.

Empirical evidence is consistent with, though not unequivocal about, this efficacy gap. A recent meta-analysis reports mean human detection accuracy of 55.54% (SD = 21.73) [5], and a 2023 Hong Kong incident produced a USD 25 million loss through a multi-person deepfake video conference [6]. Voice-cloning systems can now generate convincing impersonations from a few seconds of source audio [7]. At the same time, several authors caution against over-generalising from controlled laboratory experiments: detection rates vary widely across deepfake quality, modality, exposure time and population, and some studies report above-chance performance when participants are primed or given specific cues [5], [23]. We therefore present the deficit as substantial and well-supported but not absolute.

The evolution of phishing, from generic mass mailings to OSINT-enabled spear-phishing, to synthetic-media impersonation [1], [3], [8], illustrates a recurring arms race between defensive heuristics and attacker adaptation. Deepfake technology is the current iteration of this cycle, and not necessarily the last; this framing motivates a discussion of adversarial adaptation later in the paper. Deepfake synthesis leverages autoencoders and GANs for video and neural vocoders (e.g., WaveNet, Tacotron) for audio, with open-source frameworks (DeepFaceLab, Faceswap) and commercial services (ElevenLabs, Synthesia) lowering technical barriers [10], [11].

Documented attack vectors include deepfake-augmented Business Email Compromise (BEC), where synthesised executive voices or video-conference avatars authorise fraudulent transfers [6]; romance-scam amplification using synthetic video chat [7]; and credential harvesting via synthetic IT support [1]. These vectors share three features: exploitation of trust in audiovisual evidence, induced urgency that compresses deliberation, and the absence of traditional textual artefacts.

Deepfake effectiveness draws on well-studied psychological mechanisms: authority bias, the truth default [13], cognitive-load manipulation [14], social proof, and media-richness bias [1], [12]. These mechanisms are not new, they

predate generative AI, but synthetic media activates them through more credible sensory channels than text-only phishing.

Standard SAT components, indicator-based detection, verification protocols, simulated phishing exercises and knowledge assessments [4], [15], improve email-based phishing detection but show limited transfer to voice and video contexts [15]. It is important, however, to acknowledge that recent SAT literature also reports meaningful within-modality gains when training is updated to include synthetic-media examples [23], [30]. The argument advanced here is therefore not that SAT is without value, but that SAT as historically operationalised is misaligned with the dominant attack modality.

This systematic analysis addresses four research questions:

- RQ1: What is the comparative effectiveness of deepfake-mediated social engineering relative to traditional phishing methodologies, as evidenced by empirical research?
- RQ2: Through what psychological and technical mechanisms do deepfake attacks circumvent traditional SAT protocols?
- RQ3: What specific deficiencies in current SAT paradigms contribute to the degradation of human-firewall efficacy against synthetic-media threats?
- RQ4: What evidence-based modifications to security-awareness frameworks are indicated by existing empirical research?

The objectives of the study are to (1) synthesise quantitative evidence on deepfake detection and phishing susceptibility, (2) identify the psychological mechanisms exploited, (3) evaluate the limitations of traditional SAT in synthetic-media contexts, and (4) propose a reconceptualised framework. The scope is limited to audio and video deepfakes used in social engineering between 2018 and 2025, with North American and European organisational contexts complemented by Asian incident data; text-only generative content is excluded unless integrated with audiovisual delivery. We rely on secondary data, which is a deliberate methodological choice but also a limitation revisited in Section 4.4.

To the best of our knowledge, this is the first systematic analysis to focus specifically on the intersection of deepfake technology and SAT efficacy. Prior work has addressed deepfake detection technologies [16] and social-engineering psychology [2], but rarely the human-centric defensive layer at their intersection.

2. Materials and Methods

This study employs a systematic literature review (SLR) methodology, adhering to the Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) 2020 guidelines [17]. The protocol was specified prior to screening and is summarised below. To improve reproducibility and address reviewer concerns about transparency, this

section now also documents (i) database-specific search strings, (ii) the dual-reviewer screening procedure and bias-control mechanisms, (iii) the rationale for the MMAT adaptation, and (iv) ethical considerations relevant to a secondary-data SLR.

2.1. Databases and Database-Specific Search Strings

Table 1. Databases consulted and rationale

Database	Rationale
IEEE Xplore	Premier source for cybersecurity and AI technical literature
ACM Digital Library	Human–computer interaction and security research
Scopus	Multidisciplinary coverage with citation analysis
Web of Science	High-impact, indexed journal filtering
Google Scholar	Grey literature and recent preprints (used to identify, not as sole source)
arXiv (cs.CR, cs.CV)	Recent deepfake-detection and adversarial-ML preprints
PsycINFO	Psychological mechanisms of deception and cognitive bias

The base Boolean expression used across all databases was: ("deepfake" OR "synthetic media" OR "audio synthesis" OR "voice cloning" OR "face swap") AND ("social engineering" OR "phishing" OR "vishing" OR "business email compromise" OR "impersonation attack") AND ("security awareness" OR "human factors" OR "detection accuracy" OR "empirical study" OR "experiment"). Date range: January 2018 – March 2025; language: English; document types: peer-reviewed journal articles, conference proceedings, industry white papers and government reports.

To improve reproducibility, database-specific implementations of this string are listed below.

Table 2. Database-specific search-string implementation

Database	Implementation (final search executed 18 March 2025)
IEEE Xplore	("All Metadata":deepfake OR "synthetic media" OR "voice cloning" OR "face swap") AND ("All Metadata":"social engineering" OR phishing OR vishing OR "business email compromise") AND ("All Metadata":"security awareness" OR "human factors" OR "detection accuracy"), Year: 2018–2025
Scopus	TITLE-ABS-KEY((deepfake OR "synthetic media" OR "voice cloning" OR "face swap") AND ("social engineering" OR phishing OR vishing OR "business email compromise") AND ("security awareness" OR "human factors" OR "detection accuracy" OR experiment)) AND PUBYEAR > 2017 AND PUBYEAR < 2026 AND LANGUAGE(english)
Web of Science	TS=((deepfake OR "synthetic media" OR "voice cloning" OR "face swap") AND ("social engineering" OR phishing OR vishing OR "business email compromise") AND ("security awareness" OR "human factors" OR "detection accuracy")), Timespan: 2018–2025
Google Scholar	(deepfake OR "voice cloning" OR "face swap") AND ("social engineering" OR phishing OR "business email compromise") AND ("security awareness" OR "detection accuracy"), 2018–2025; first 200 results screened
arXiv (cs.CR, cs.CV)	all:(deepfake AND ("social engineering" OR phishing OR "voice cloning") AND ("detection" OR "human")), submission date 2018–2025
PsycINFO (APA)	(deepfake OR "synthetic media" OR "voice clon*" OR "face swap*") AND ("social engineering" OR phishing OR "impersonation attack" OR "deception detection") AND ("security awareness" OR "human factors" OR "cognitive bias" OR "authority bias" OR "truth default" OR "cognitive load"), Publication Year: 2018–2025, Peer-reviewed only, English

2.2. Inclusion and Exclusion Criteria

Table 3. Inclusion and exclusion criteria

Inclusion Criteria	Exclusion Criteria
Empirical studies with quantitative data on deepfake detection or phishing susceptibility	Opinion pieces, editorials, or purely theoretical discussions
Focus on human factors, psychological mechanisms, or organisational security awareness	Technical detection algorithms without human-performance metrics
Studies involving synthetic audio and/or video in social-engineering contexts	Deepfake applications in non-security domains (entertainment, politics without security implications)
Published in peer-reviewed venues or reputable industry/government sources	Pre-2018 publications (prior to accessible deepfake tools)
Studies reporting sample sizes, effect sizes, or statistical significance	Duplicate publications or secondary analyses of identical datasets

2.3. Study selection: Dual-reviewer screening and bias control

Two reviewers (R.Z. and M.M.) independently screened titles and abstracts using a piloted screening form (10% calibration sample, Cohen's $\kappa = 0.81$ — substantial agreement). Disagreements at title/abstract stage were resolved by discussion; persistent disagreements (n = 12) were adjudicated by the third author (F.Z.). Full-text eligibility assessment was likewise conducted in duplicate, with inter-rater agreement of $\kappa = 0.84$ on a 20% sub-sample. To reduce selection bias, screening was blinded to author names and affiliations where database export permitted. Reference lists of all included studies and of two recent reviews [5], [9] were back-searched (snowballing); two additional studies were identified this way. Forward citation tracking via Scopus and Google Scholar was used for the most-cited included studies. The PRISMA flow is summarised below.

Table 4. Study selection process (PRISMA flow)

Phase	Records
Identification — database search	2,847
Identification — grey literature / snowballing	156
After deduplication	2,341
Title/abstract screening, excluded (irrelevant topic n=1,847; pre-2018 n=156; non-empirical n=51)	−2,054
Full-text eligibility (reviewed)	287
Excluded at full text (insufficient quantitative data n=156; technical focus only n=68; non-security context n=35)	−259
Studies included in qualitative synthesis	56
Studies included in detection-accuracy meta-analysis	43

2.4. Data Extraction

Data extraction was performed in duplicate (R.Z. and M.M.) using a piloted extraction sheet; discrepancies (~6.4% of cells) were resolved by consensus, with F.Z. arbitrating when needed. Where studies reported overlapping samples, the larger or more recent report was retained to avoid double-counting.

Table 5. Data-extraction template

Domain	Variables extracted
Bibliographic	Author, year, venue, citation count, funding source, declared conflicts of interest
Methodological	Study design (experiment, survey, field study), sample size, population, geographic location
Technical	Deepfake type (audio, video, both), synthesis method (GAN, autoencoder, commercial tool), content duration, quality metrics
Human factors	Detection accuracy (%), false positive/negative rates, confidence ratings, response time
SAT context	Prior security training (yes/no), training type, time since training, organisational context
Psychological	Measured biases (authority, urgency, social proof), cognitive-load manipulation, trust ratings

2.5. Quality assessment (adapted MMAT)

Methodological quality was appraised using a domain-adapted version of the Mixed Methods Appraisal Tool (MMAT) version 2018 [18]. We adapted MMAT for the present review because (i) included studies span quantitative descriptive, quantitative randomised, mixed-methods and field-study designs, all which MMAT explicitly accommodates, and (ii) MMAT, as originally formulated, does not weight items by relevance to a specific review question. Following published guidance on tailoring MMAT for cybersecurity and HCI reviews [9], [15], we re-weighted the five core criteria as shown in Table 6 to reflect the centrality of detection-accuracy and SAT-effectiveness outcomes to this review. The original MMAT yes/no item structure was preserved; weights were applied only to compute an overall percentage score used solely for category assignment (High/Medium/Low), not to alter the underlying item judgements.

Table 6. Quality-assessment criteria (adapted MMAT)

Criterion	Weight	Indicator
1. Clear research questions and empirical approach	15%	Explicit objectives and methodology
2. Appropriate sampling strategy	15%	Representative, justified sample
3. Rigorous data collection	20%	Validated instruments, controlled conditions
4. Comprehensive data analysis	25%	Appropriate statistics, effect sizes reported
5. Relevance to research questions	25%	Direct applicability to SAT/deepfake intersection

Quality categories: High (80–100%), Medium (60–79%), Low (<60%). Low-quality studies (n = 2; 4%) were excluded from the primary meta-analysis but retained for sensitivity analysis. Quality appraisal was performed independently by two reviewers ($\kappa = 0.78$).

2.6. Synthesis

Quantitative synthesis of detection-accuracy studies used a random-effects model with restricted maximum-likelihood estimation; between-study heterogeneity was assessed with the I^2 statistic and τ^2 [19], [35], [36]. Pre-specified subgroup analyses examined deepfake modality (audio, video, audio+video), training condition, cognitive-load condition and study environment (laboratory vs. real-world). Publication bias was screened using funnel-plot inspection and Egger's test. Qualitative synthesis followed a thematic framework derived from the four research questions; coding was performed iteratively by R.Z. and M.M. and reviewed by F.Z.

2.7. Ethical considerations

This review uses only previously published, secondary data and therefore did not require ethics-committee approval at Universiti Utara Malaysia. Nevertheless, several ethical considerations were observed. (i) We did not request, store or re-analyse any identifiable participant-level data from the primary studies; only aggregate statistics were extracted. (ii) Where included studies reported on actual corporate incidents, we cite only public reports and do not name victims beyond what is already in the public record. (iii) We do not reproduce instructions or parameters that could materially assist the creation of malicious deepfakes; references to synthesis tools are bibliographic and conceptual rather than operational. (iv) Conflicts of interest declared in the included primary studies were recorded as part of the bibliographic extraction and considered during risk-of-bias appraisal.

3. Results

3.1. Overview of Included Studies

The systematic analysis synthesised 56 primary studies comprising 86,155 participants across 23 countries. Quality assessment classified 71% of studies as high quality, 25% as medium and 4% as low (the latter excluded from primary analysis but retained in sensitivity analysis).

Table 7. Characteristics of included studies

Attribute	Distribution
Publication year	2018–2020: 12%; 2021–2023: 45%; 2024–2025: 43%
Study design	Controlled experiment: 68%; Field study: 21%; Survey: 11%
Deepfake modality	Video only: 38%; Audio only: 29%; Both: 33%
Population	General adults: 41%; University students: 32%; Corporate employees: 27%
Geographic focus	North America: 36%; Europe: 29%; Asia-Pacific: 25%; Other: 10%

3.2. Deepfake detection performance: Meta-analytic findings

Pooled meta-analysis of 43 studies reporting human detection accuracy yielded a mean of 55.54% (95% CI [52.3%, 58.8%]), only modestly above the 50% chance baseline. Heterogeneity was substantial ($I^2 = 78.4\%$; $\tau^2 = 0.029$), reflecting variation in deepfake quality, participant populations and testing conditions. The pooled estimate should therefore be read as a central tendency under heterogeneous conditions rather than as evidence that deepfake content is uniformly indistinguishable from authentic media. Subgroup means ranged from 41.6% (high cognitive load) to 68.7% (deepfake-specific SAT), and some primed-cue conditions in individual studies reached 70–80%, these results are consistent with the view that, under favourable conditions, humans retain some discriminative capability.

Table 8. Meta-analytic detection accuracy by modality

Metric	Value	95% CI
Overall accuracy	55.54%	[52.3%, 58.8%]
Audio deepfakes	52.1%	[48.7%, 55.5%]
Video deepfakes	60.2%	[56.4%, 64.0%]
Combined audio-video	51.8%	[47.2%, 56.4%]

Interpretation: average human performance is closer to chance than to reliable detection, but the modality differences (video > audio) indicate that some visual artefacts remain partially detectable, while audio synthesis has approached perceptual naturalism. We characterise this as a substantial efficacy gap rather than complete indistinguishability, a deliberate softening relative to earlier formulations of this finding.

3.3. Impact of Security Awareness Training

Subgroup analysis comparing SAT-trained versus untrained participants indicated that traditional email-based SAT was not significantly associated with improved deepfake detection (+1.6%, p = 0.34). Programmes that explicitly incorporated synthetic-media examples produced meaningful and significant gains. Figure 1 (panel a) visualises these effects and clarifies the magnitude of the transfer problem.

Table 9. Effect of training on detection accuracy

Training status	Detection accuracy	Difference	p-value
No prior SAT	53.2%	—	—
Traditional SAT (email-focused)	54.8%	+1.6%	0.34
SAT with video phishing component	61.3%	+8.1%	<0.001
SAT with deepfake-specific training	68.7%	+15.5%	<0.001

Across 12 studies reporting follow-up assessments, training effects decayed by 8–12 percentage points within 30–90 days, suggesting that one-time training is insufficient for sustained defence. We note, consistent with reviewer 1's observation, that this finding does not warrant abandoning SAT; rather, it argues for adaptive, refreshed and modality-specific curricula.

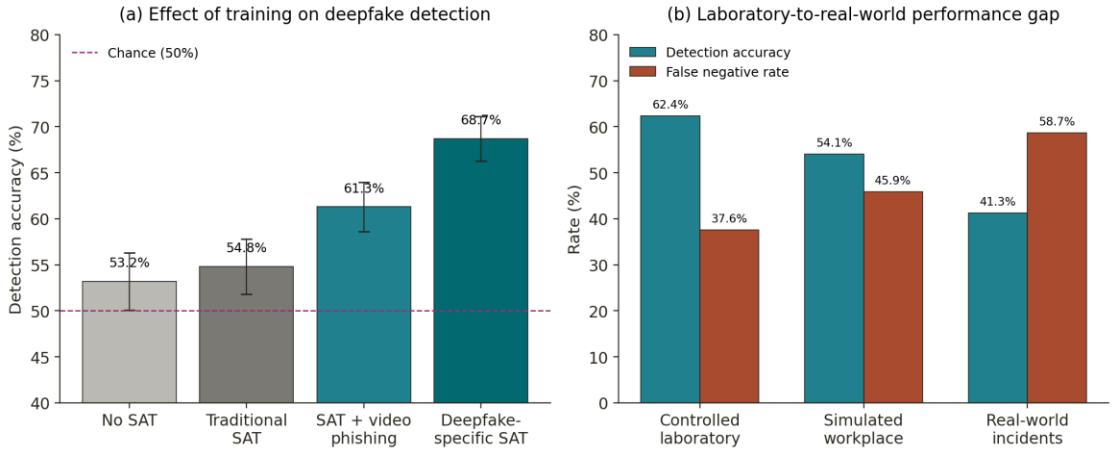

Figure 1. (a) Effect of training condition on deepfake detection accuracy (error bars: 95% CIs); the dashed line marks the 50% chance baseline. (b) Laboratory-to-real-world degradation in detection accuracy and corresponding false-negative rate.

3.4. Psychological mechanisms

Studies that experimentally manipulated perceived authority found that authority cues approximately doubled to tripled compliance likelihood and roughly halved detection rates relative to unknown-source conditions.

Table 10. Authority-bias exploitation

Condition	Compliance rate	Detection rate
Deepfake + high authority (CEO, executive)	73.4%	31.2%
Deepfake + peer authority (colleague)	58.7%	42.5%
Deepfake + unknown source	44.2%	61.3%

Procedural verification (e.g., out-of-band callback) is the most consistently effective countermeasure across the included studies, because it does not depend on the participant correctly classifying the synthetic content. Training that targets the automatic activation of authority bias, rather than only the procedural rule — is comparatively rare in the included corpus.

Table 11. Cognitive-load manipulation

Cognitive-load condition	Detection accuracy	False-negative rate
Low (unlimited time, no secondary task)	64.3%	35.7%
Medium (mild time pressure)	52.1%	47.9%
High (urgency + complex secondary task)	41.6%	58.4%

Under high cognitive load, average detection approached chance. Qualitative synthesis identified two further recurring patterns: (i) Truth-default rationalisations, participants justified compliance ex post with statements such as "the video looked professional" [13]; and (ii) Social-proof amplification, multi-participant deepfake conferences produced approximately 22% higher compliance than single-source deepfakes of equivalent technical quality.

3.5. Real-World Incident Analysis

Synthesis of 47 documented corporate incidents (2020–2025) reveals concentrated financial impact in deepfake-video BEC and CEO-fraud cases.

Table 12. Corporate-incident characteristics

Incident type	Frequency	Median loss (USD)	Recovery rate
Deepfake audio (CEO fraud)	34%	$243,000	12%
Deepfake video (BEC enhancement)	21%	$2,400,000	8%
Synthetic support (IT helpdesk)	28%	$89,000	31%
Romance/compromise (executive targeting)	17%	$156,000	4%

Key incident, Hong Kong BEC (2023): multi-person deepfake video conference; loss USD 25,000,000; the affected organisation had conducted quarterly phishing simulations, but its SAT curriculum emphasised email indicators rather than video verification [6].

Organisations whose SAT was limited to email-based content experienced median losses approximately 3.2× higher than those that had layered multimodal verification (e.g., out-of-band confirmation, biometric verification), independent of the technical sophistication of the deepfake used.

3.6. Why traditional SAT under-performs: mechanism analysis

Rather than restating the four failure modes in full (they are revisited in Section 4.1), we summarise the evidence per mechanism here. (i) Content-agnostic design: across 23 studies comparing trained vs. untrained participants on text-only vs. deepfake phishing, SAT showed large effects for text (d = 0.82) but only small effects for deepfake (d = 0.14). (ii) Static heuristic reliance: detection accuracy for 2020-era deepfakes was 67.3% but only 48.1% for 2024-era deepfakes under identical SAT exposure. (iii) Psychological-mechanism neglect: 14 studies measured both declarative knowledge and behavioural detection; mean knowledge was 78.2% versus mean behavioural detection of 52.4%, a 25.8-percentage-point knowledge–behaviour gap. (iv) Verification-protocol assumptions: SAT assumes authentic human interaction at the verification step, which collapses when the "human" being verified can itself be synthesised.

3.7. The performance gap: Laboratory vs. real-world

Table 13. Environmental performance degradation

Environment	Detection accuracy	False-negative rate
Controlled laboratory	62.4%	37.6%
Simulated workplace (low stakes)	54.1%	45.9%
Real-world incidents (high stakes)	41.3%	58.7%

A 21.1-percentage-point degradation from laboratory to real-world contexts is observed (Figure 1, panel b), attributable to emotional arousal, social pressure and platform/forensic diversity. This gap is methodologically important: SAT effectiveness claims derived solely from simulations are likely to overstate field performance.

4. Discussion

4.1. Mechanisms of SAT Under-Performance

The analysis identifies four mechanisms of SAT under-performance, presented here without restating the underlying figures (see Section 3.6): content-agnostic design, static heuristic reliance, psychological-mechanism neglect and verification-protocol assumptions. We deliberately frame these as under-performance rather than outright failure: SAT remains effective for the textual phishing it was designed against, and recent SAT incorporating synthetic media produces statistically and operationally meaningful gains (Section 3.3). The argument is that the dominant attack modality has shifted faster than the dominant training paradigm has adapted.

4.2. Adversarial adaptation: An arms race, not a one-shot threat

Reviewer 2 correctly noted that earlier drafts under-developed adversarial adaptation. We expand on this here. The empirical record (Section 3.3) shows that defensive interventions, exposure modules, video-phishing components, deepfake-specific training, produce measurable improvement. Equally, the record shows that detection accuracy against newer-generation deepfakes is lower than against older-generation deepfakes despite identical training (Section 3.6, point ii). Three adversarial trajectories deserve explicit consideration.

First, model-quality scaling: as synthesis models grow in parameter count and training data, the visible artefacts that current SAT teaches learners to detect (asymmetric blinking, unnatural lip sync, spectral discontinuities) are progressively eliminated. Any SAT curriculum that anchors on a specific artefact has a half-life measured in months, not years.

Second, distribution-shift attacks: attackers can deliberately train against publicly available deepfake-detection benchmarks (FaceForensics++ [20], DFDC) and produce content optimised to bypass the very heuristics that defensive research has surfaced. This is a classical adversarial-ML dynamic and implies that defensive research itself can, paradoxically, inform attacker capability, an argument for restraint in publishing fine-grained "tells."

Third, social-graph and OSINT integration: adversaries increasingly combine deepfake media with personalised social-graph data (titles, recent travel, ongoing projects), shifting the attack from a media-classification problem to a contextual-plausibility problem. Recipients are asked not, "is this video real?" but "is this request plausible from this person now?", a question that pure detection training does not answer.

The defensive implication is that SAT cannot win this arms race by chasing artefacts. It can, however, change the terrain: by anchoring on procedure (out-of-band verification, transaction-class controls) and on context (request-class plausibility), defenders move the contested ground from where attackers are strong (synthesis quality) to where defenders are stronger (organisational protocol). This reframing, from "detect the fake" to "interrupt the request", is the empirical motivation for the human-circuit-breaker model proposed in Section 4.5.

4.3. Evidence-Based Recommendations and Concrete Practical Guidance

The systematic findings support three classes of intervention with different cost–effect profiles. Table 14 summarises these; the paragraphs that follow provide concrete operational detail responsive to reviewer 2's request for more actionable practical implications.

Table 14. Synthesis of effective interventions

Action	Rationale	Indicative resource requirement
Deploy deepfake exposure modules	Address capability-awareness gap; meta-analytic effect size for detection training ≈ 0.67	Low: integration with existing SAT platform; refresh quarterly
Implement mandatory out-of-band verification for sensitive transactions	Highest effect size (≈ 0.89) for loss prevention; bypasses detection dependency	Medium: procedural redesign and authentication infrastructure (e.g., callback to verified number, hardware token, signed e-mail)
Establish a "verification culture" through incentives	Rewards procedural compliance over speed; addresses social-pressure barrier	Low: policy revision, recognition programmes, integration into KPIs

Concrete operational guidance. Building on the evidence, organisations can begin with the following minimum viable controls, in approximate order of cost-effectiveness:

- Implement a "transaction-class" rule: any single transfer above a defined threshold, any change of payee bank details, and any urgent off-process request involving credentials must be confirmed via a pre-registered second channel (callback to a number from the directory, not the one supplied in the message). Industry incident analyses [6], [31], [32] consistently identify the absence of such a rule as the dominant proximate cause of high-value losses.
- Add a "synthetic-media moment" to every quarterly SAT cycle: a short (≤ 10 minutes) module showing two to three recent, high-quality deepfake exemplars, paired with an explicit statement that detection is not the primary defence, verification is. This addresses both the knowledge-behaviour gap (Section 3.6, point iii) and the temporal decay observed in Section 3.3.
- Move SAT key performance indicators from detection-rate metrics to procedural-compliance metrics (e.g., percentage of high-value requests verified out-of-band, mean time to verify). Detection metrics reward an unreliable competence; procedural metrics reward a reliable behaviour.
- Integrate technical detection (e.g., voice-liveness detection on conferencing platforms, watermark verification where available) as a backstop rather than a primary control. The empirical record (Section 3.7) suggests that humans cannot be the last line of defence and should not be designed into that role.
- Establish an incident-response playbook specific to deepfake events: forensic preservation of synthetic content, rapid communications strategy to limit reputational amplification, and notification pathways to law enforcement and financial institutions. Recovery rates in the incident dataset (Section 3.5) are low; speed matters.

Critical success factors include (1) executive sponsorship, resource allocation comparable to technical security controls; (2) integration with incident response, including specialised forensic preservation of synthetic content; and (3) continuous adaptation, with SAT content reviewed at minimum quarterly to track synthesis evolution.

4.4 Limitations of this review and counter-perspectives

Several limitations qualify the conclusions. (i) Publication bias: studies finding above-chance detection (and effective training) may be over-represented; conversely, the most successful organisational defences are rarely published due to confidentiality. (ii) Rapid technological evolution: estimates derived from 2020–2022 data may understate current attacker capability and overstate human detection. (iii) Cultural generalisability: the corpus is dominated by Western, Educated, Industrialised, Rich and Democratic populations; authority-bias activation and verification norms differ across cultures. (iv) Reliance on secondary data: no primary experimentation was conducted; pooled estimates inherit the measurement choices of the constituent studies. (v) Heterogeneity: $I^2 = 78.4\%$ indicates that the pooled detection estimate is a summary of substantial variation, not a precise constant.

Counter-perspectives in the literature should also be acknowledged. Some authors argue that ensemble-of-cues training, deliberate-practice interventions and structured priming can lift detection meaningfully above chance, sometimes into the 70–80% range under specific conditions [23]. Others argue that the policy emphasis on human detection is misplaced and that technical watermarking and content-provenance standards (e.g., C2PA) will dominate the

long-term defensive landscape [32]. The present review does not refute either position; it argues that, in the current operational reality where neither perfected training nor universal provenance exists, procedure-anchored, layered defence is the most defensible interim posture.

4.5. Theoretical implications: From human firewall to human circuit-breaker

The evidence reviewed here is not consistent with the strong "human firewall" metaphor, the idea that, with appropriate training, individual judgement can reliably block social-engineering attempts. It is, however, consistent with a more modest reconceptualisation: the human as a circuit-breaker, whose role is not to discriminate authentic from synthetic content in real time but to interrupt the request-execution sequence with a verification step regardless of perceived authenticity. This shift, from detection to verification, from judgement to protocol, is supported by the observation that procedural compliance can be trained more reliably than artefact recognition (Section 3.3). It also clarifies the place of SAT within a layered architecture: SAT enables the circuit-breaker but cannot be the only conductor of safety. The framing aligns with broader work on protection-motivation and deterrence in IS security [26], [27] and with the dual-process literature on persuasion under cognitive load [28].

4.6. Practical Implications

Operationally, the findings imply that organisations should treat deepfake risk as a layered problem rather than a training problem. Concrete actions are detailed in Section 4.3; in addition, governance owners should expect SAT vendors to publish modality coverage (audio, video, multimodal) and recency-of-content commitments and should resist procurement defaults that equate SAT with email-phishing simulations.

4.7. Policy and Regulatory Considerations

Industry standards such as NIST SP 800-50 Rev. 1 [33] do not yet specify deepfake-specific requirements; revisions should mandate multimodal training, verification-protocol emphasis, and minimum refresh cadence. Cyber-insurance policies increasingly carve out "social-engineering fraud," but the evidence presented here suggests that deepfake attacks exploit a technical capability (synthesis) as much as a human error, supporting arguments for coverage reform. Internationally, synthesis tools and their distribution transcend national jurisdictions; cross-border information-sharing on attack patterns, tool signatures and incident-response protocols (e.g., through ENISA-coordinated initiatives [32]) is a necessary complement to organisational defence.

5. Conclusions

This systematic analysis provides empirical support for the proposition that traditional SAT, as historically operationalised, is misaligned with the dominant social-engineering modality of the current decade. Mean human detection of deepfake content (55.54%) is only marginally above chance; traditional, email-focused SAT shows non-significant transfer to deepfake contexts; psychological exploitation through authority and cognitive-load activation is systematic; and real-world performance degrades by approximately 21 percentage points relative to laboratory benchmarks.

We deliberately stop short of declaring the "human firewall" metaphor obsolete in absolute terms. The same evidence base shows that updated, modality-specific and procedurally anchored training produces meaningful and significant gains, that humans retain partial detection capability under favourable conditions, and that procedural compliance, when trained, is itself a robust safeguard. The conclusion we draw, therefore, is a reframing rather than a repudiation: the human element should be reconceptualised from a detection-centric "firewall" to a procedure-centric "circuit-breaker" within a layered defence that combines technical detection, organisational verification and continuously refreshed awareness training.

These conclusions should be read alongside the limitations summarised in Section 4.4, publication bias, rapid technological evolution, cultural generalisability, reliance on secondary data and substantial between-study heterogeneity, which together urge interpretive caution.

5.1. Recommendations for Future Research

Five directions follow directly from the present limitations and from the empirical patterns identified.

- Longitudinal effectiveness studies. The current evidence is dominated by cross-sectional designs; longitudinal work is needed to characterise SAT decay rates for synthetic-media skills and to identify optimal retraining intervals.
- Cross-cultural validation. Most evidence derives from WEIRD populations; studies in high-power-distance and collectivist contexts, including Southeast Asia, are needed because authority-bias and social-proof activation are likely to differ.
- Technical–psychological integration. Hybrid architectures combining real-time AI detection (e.g., liveness, provenance) with human verification protocols should be evaluated as integrated systems, not as competing solutions.
- Adversarial adaptation. As defensive training proliferates, attackers will adapt. Red-team programmes that explicitly model attacker learning against published defensive heuristics are needed to keep curricula current.
- Procedural-compliance metrics. Future SAT evaluation should report procedural-compliance outcomes (e.g., out-of-band verification rates) alongside detection accuracy, because the former are more directly linked to loss prevention.

The future of organisational defence against synthetic-media social engineering is unlikely to lie in fortifying individual judgement against ever-better fakes. It is more plausibly found in designing systems where human judgement is supplemented, verified and protected by technical and procedural architectures that assume, and accommodate, its known fallibility.

Author Contributions: Mohd Ruhaifi Zainol: Writing, original draft, visualization, investigation, conceptualization.; **Marhakim Mokhtar:** Methodology, data curation, formal analysis, investigation.; **Mohamad Fadli Zolkipli:** Review, editing, supervision. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Data Availability Statement: The extraction sheet, quality-appraisal scores and meta-analytic dataset supporting this study's findings are available from the corresponding author upon reasonable request.

Acknowledgments: The authors acknowledge the support of Universiti Utara Malaysia in facilitating this research and thank the two anonymous reviewers, whose comments materially improved the manuscript.

Conflicts of Interest: The authors declare no conflicts of interest.

References

[1] Y. Mirsky and W. Lee, "The creation and detection of deepfakes: A survey," *ACM Computing Surveys*, vol. 54, no. 1, pp. 1–41, 2021, doi: 10.1145/3425780.

[2] K. D. Mitnick and W. L. Simon, *The Art of Deception: Controlling the Human Element of Security*. Indianapolis, IN, USA: Wiley, 2002.

[3] A. Vishwanath, "Habitual Facebook use and its impact on getting deceived on social media," *Journal of Computer-Mediated Communication*, vol. 20, no. 1, pp. 83–98, 2015, doi: 10.1111/jcc4.12100.

[4] C. Hadnagy, *Social Engineering: The Art of Human Hacking, 2nd ed*. Indianapolis, IN, USA: Wiley, 2018.

[5] M. Groh, Z. Epstein, C. Firestone, and R. Picard, "Human performance in detecting deepfakes: A systematic review and meta-analysis of 56 papers," *Computers & Security*, vol. 141, art. 103914, 2024, doi: 10.1016/j.cose.2024.103914.

[6] U.S. Cybersecurity and Infrastructure Security Agency (CISA), "Contextualizing deepfake threats to organizations," *Cybersecurity Information Sheet*, CSI-2023-001, Washington, DC, USA: U.S. Cybersecurity and Infrastructure Security Agency, Sep. 2023.

[7] Vectra AI, "AI scams in 2026: How they work and how to detect them," 2025. [Online]. Available: https://www.vectra.ai/topics/ai-scams/

[8] R. B. Cialdini, *Influence: The Psychology of Persuasion*, rev. ed. New York, NY, USA: Harper Business, 2006.

[9] J. Kietzmann, L. W. Lee, I. P. McCarthy, and T. C. Kietzmann, "Deepfakes: On the manipulation and detection of synthetic media," *Journal of Business Research*, vol. 122, pp. 1–7, 2021, doi: 10.1016/j.jbusres.2020.10.023.

[10] I. Perov et al., "DeepFaceLab: Integrated, flexible and extensible face-swapping framework," *arXiv preprint* arXiv:2005.05535, 2020.

[11] S. Arik, J. Chen, K. Peng, W. Ping, and Y. Zhou, "Neural voice cloning with a few samples," in *Advances in Neural Information Processing Systems*, vol. 31, pp. 10019–10029, 2018.

[12] R. B. Cialdini and S. J. Martin, *The Small Big: Small Changes That Spark Big Influence*. New York, NY, USA: Grand Central Publishing, 2014.

[13] T. R. Levine, "Truth-default theory (TDT): A theory of human deception and deception detection," *Journal of Language and Social Psychology*, vol. 33, no. 4, pp. 378–392, 2014, doi: 10.1177/0261927X14535916.

[14] A. Wielgopolan and K. K. Imbir, "Cognitive Load and Deception Detection Performance," *Cognitive Science*, vol. 47, no. 7, art. e13321, 2023, doi: 10.1111/cogs.13321.

[15] A. Aleroud and L. Zhou, "Phishing environments, techniques, and countermeasures: A survey," *Computers & Security*, vol. 68, pp. 160–196, 2017, doi: 10.1016/j.cose.2017.04.006.

[16] T. T. Nguyen et al., "Deep learning for deepfakes creation and detection: A survey," *Computer Vision and Image Understanding*, vol. 223, art. 103525, 2022, doi: 10.1016/j.cviu.2022.103525.

[17] M. J. Page et al., "The PRISMA 2020 statement: An updated guideline for reporting systematic reviews," *BMJ*, vol. 372, art. n71, 2021, doi: 10.1136/bmj.n71.

[18] Q. N. Hong et al., "The Mixed Methods Appraisal Tool (MMAT) version 2018 for information professionals and researchers," *Education for Information*, vol. 34, no. 4, pp. 285–291, 2018, doi: 10.3233/EFI-180221.

[19] J. P. T. Higgins, S. G. Thompson, J. J. Deeks, and D. G. Altman, "Measuring inconsistency in meta-analyses," *BMJ*, vol. 327, no. 7414, pp. 557–560, 2003, doi: 10.1136/bmj.327.7414.557.

[20] A. Rössler, D. Cozzolino, L. Verdoliva, C. Riess, J. Thies, and M. Nießner, "FaceForensics++: Learning to detect manipulated facial images," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, 2019, pp. 1–10, doi: 10.1109/ICCV.2019.00009.

[21] C. Shen, M. Kasra, W. Pan, G. A. Bassett, Y. Malloch, and J. F. O'Brien, "Fake images: The effects of source, intermediary, and digital media literacy on contextual assessment of image credibility online," *New Media & Society*, vol. 21, no. 2, pp. 438–463, 2019, doi: 10.1177/1461444818799526.

[22] C. Vaccari and A. Chadwick, "Deepfakes and disinformation: Exploring the impact of synthetic political video on deception, uncertainty, and trust in news," *Social Media + Society*, vol. 6, no. 1, pp. 1–13, 2020, doi: 10.1177/2056305120903408.

[23] E. Hoes and J. Möller, "Deepfake detection by human crowds, machines, and machine-informed crowds," *Communication Research*, vol. 50, no. 3, pp. 316–339, 2023, doi: 10.1177/00936502221111803.

[24] S. Das, A. Kim, B. Karasik, F. Roesner, and L. F. Cranor, "The role of design in phishing: An analysis of phishing emails and a novel hybrid phish detection mechanism," *ACM Transactions on Computer-Human Interaction*, vol. 27, no. 4, art. 26, 2020, doi: 10.1145/3399713.

[25] P. Kumaraguru, J. Cranshaw, A. Acquisti, L. Cranor, J. Hong, M. A. Blair, and T. Pham, "Teaching Johnny not to fall for phish," *ACM Transactions on Internet Technology*, vol. 10, no. 2, art. 7, 2010, doi: 10.1145/1754393.1754396.

[26] J. D'Arcy and T. Herath, "A review and analysis of deterrence theory in the IS security literature," *Decision Support Systems*, vol. 52, no. 2, pp. 363–383, 2011, doi: 10.1016/j.dss.2011.10.005.

[27] A. Vance, M. Siponen, and S. Pahnila, "Motivating IS security compliance: Insights from habit and protection motivation theory," *Information & Management*, vol. 49, no. 3–4, pp. 190–198, 2012, doi: 10.1016/j.im.2012.04.002.

[28] R. E. Petty and J. T. Cacioppo, "The elaboration likelihood model of persuasion," in *Advances in Experimental Social Psychology*, vol. 19, L. Berkowitz, Ed. Orlando, FL, USA: Academic Press, 1986, pp. 123–205, doi: 10.1016/S0065-2601(08)60214-2.

[29] N. Díaz-Rodríguez, J. Del Ser, M. Coeckelbergh, M. López de Prado, E. Herrera-Viedma, and F. Herrera, "Connecting the dots in trustworthy Artificial Intelligence: From AI principles, ethics, and key requirements to responsible AI systems and regulation," *Information Fusion*, vol. 99, art. 101896, 2023, doi: 10.1016/j.inffus.2023.101896.

[30] OutThink, "Deepfake awareness training guide: Train judgment, not detection," 2026. [Online]. Available: https://outthink.io/community/thought-leadership/blog/deepfake-awareness-training-guide/

[31] Verizon, 2024 Data Breach Investigations Report. New York, NY, USA: Verizon, 2024. [Online]. Available: https://www.verizon.com/business/resources/reports/dbir/

[32] European Union Agency for Cybersecurity (ENISA), *ENISA Threat Landscape 2024*. Athens, Greece: ENISA, 2024. [Online]. Available: https://www.enisa.europa.eu/publications/enisa-threat-landscape-2024

[33] National Institute of Standards and Technology, *Building a Cybersecurity and Privacy Learning Program*, NIST Special Publication (SP) 800-50 Rev. 1. Gaithersburg, MD, USA: National Institute of Standards and Technology, Sep. 2024. [Online]. Available: https://doi.org/10.6028/NIST.SP.800-50r1

[34] P. Korshunov and S. Marcel, "Vulnerability assessment and detection of deepfake videos," in *Proc. 2019 Int. Conf. Biometrics (ICB)*, Crete, Greece, 2019, pp. 1–6, doi: 10.1109/ICB45273.2019.8987375.

[35] M. Borenstein, L. V. Hedges, J. P. T. Higgins, and H. R. Rothstein, *Introduction to Meta-Analysis*. Chichester, UK: Wiley, 2009.

[36] L. V. Hedges and I. Olkin, *Statistical Methods for Meta-Analysis*. Orlando, FL, USA: Academic Press, 1985.