

AI-Enabled Social Engineering and Deepfake Fraud in Higher Education: An Integrative Review of Threats, Human Detection Limits and Layered Defences, with a Research Agenda for the Global South

Shamsuddeen Mohammed Sada¹, Nnenna Harmony Nwobodo-Nzeribe², Adesegun Nurudeen Osijirin^{3*}

¹Department of Software Engineering, Federal University of Allied Health Sciences, Enugu, Nigeria

²Department of Computer Engineering, Enugu State University of Science and Technology, Enugu, Nigeria

³Division of General Studies and Digital Literacy, Federal University of Allied Health Sciences, Enugu, Nigeria

DOI: <https://doi.org/10.36347/sjet.2026.v14i10.005>

| Received: 17.08.2026 | Accepted: 23.10.2026 | Published: 26.10.2026

*Corresponding author: Adesegun Nurudeen Osijirin

Division of General Studies and Digital Literacy, Federal University of Allied Health Sciences, Enugu, Nigeria

Abstract

Review Article

Purpose: Generative artificial intelligence (AI) has changed the economics of social engineering. Large language models write fluent, personalised phishing at near-zero cost, voice-cloning tools imitate a lecturer or parent from seconds of audio, and real-time deepfake video can impersonate an official on a call. Universities combine an open culture, high staff and student turnover, large financial flows and young, heavily connected users, which makes them attractive targets. Yet the literature on AI-enabled deception and the literature on cybersecurity in higher education have developed largely in parallel. This review brings them together. **Method:** An integrative review following Whittemore and Knafl's (2005) approach. Structured searches of Scopus, Web of Science, ScienceDirect, SpringerLink, IEEE Xplore and the ACM Digital Library (2017–2026), supplemented by citation chasing, identified empirical studies, systematic reviews and authoritative reports on AI-enabled social engineering, human deepfake detection, phishing susceptibility and literacy-based interventions. The evidence was synthesised thematically and mapped onto a conceptual model. **Findings:** Four findings stand out. First, generative AI amplifies six threat vectors relevant to universities, from large language model (LLM) spear-phishing to synthetic-persona scholarship and romance fraud. Second, pooled human deepfake detection accuracy is close to chance (about 55% across 56 papers), and people consistently overestimate their own ability. Third, awareness training and media-literacy interventions produce real but modest and decaying gains, and they work better against misinformation than against targeted fraud. Fourth, research on students is concentrated in high-income countries, with very little from sub-Saharan Africa despite the region's specific exposure to financial fraud. **Contribution:** We propose a Layered Human–Technical Defence Model that links threat vectors, human detection limits, individual, procedural, technical and institutional defences, and contextual moderators. We derive six testable propositions and a research agenda for Global South universities. **Keywords:** deepfakes; generative AI; social engineering; phishing; higher education; cybersecurity awareness; media literacy; AI literacy; fraud detection; Global South.

Copyright © 2026 The Author(s): This is an open-access article distributed under the terms of the Creative Commons Attribution 4.0 International License (CC BY-NC 4.0) which permits unrestricted use, distribution, and reproduction in any medium for non-commercial use provided the original author and source are credited.

1. INTRODUCTION

Social engineering, which is manipulation that leads people to give up information, money or access, has long been the most reliable way into an organisation's systems (Krombholz *et al.*, 2015; Salahdine & Kaabouch, 2019). Its effectiveness has always depended on the attacker's ability to craft a credible pretext: an email that sounds like the bursar, a call that sounds like the head of department, a profile that looks like a real person. Generative artificial intelligence (AI) removes much of the cost and skill this used to require. Large language models (LLMs) produce grammatical, context-

aware messages in many languages; voice-cloning systems reproduce a speaker from a short sample; and face-swap and lip-sync tools generate convincing video, increasingly in real time (Mirsky & Lee, 2021; Tolosana *et al.*, 2020; Schmitt & Flechais, 2024).

The consequences are no longer hypothetical. Hazell (2023) showed that an LLM could generate personalised spear-phishing emails for more than 600 British Members of Parliament at a cost of a fraction of a cent each. Heiding *et al.*, (2024) found that fully automated LLM-written phishing emails achieved click-

through rates close to those of emails crafted by human experts using an established persuasion model. In early 2024, a finance employee at a multinational engineering firm in Hong Kong transferred roughly US\$25 million after a video call in which the "chief financial officer" and other colleagues were deepfakes (Chen & Magramo, 2024). Law-enforcement agencies have warned that deepfakes are moving from novelty to routine criminal tooling (Europol, 2022).

Universities are exposed for structural reasons. They are open by design, with large and changing populations, decentralised IT, and a culture that values sharing over restriction (Ulven & Wangen, 2021). They hold valuable research, personal and financial data, and they move money in predictable cycles of tuition, scholarships, bursaries and stipends. Their main users are young adults who live much of their social and financial lives on mobile devices. In parts of the Global South, students are also both targets and recruits for organised online fraud. Nigerian studies have documented the social organisation of internet fraud among undergraduates (Tade & Aliyu, 2011; Ojedokun & Eraye, 2012), and recent survey work in Enugu State has begun to examine how cybersecurity awareness and digital and media literacy relate to students' preparedness to detect AI-powered scams and deepfakes (Osijirin *et al.*, 2026a).

Despite this, the relevant evidence sits in separate bodies of work. Computer science surveys catalogue deepfake generation and detection algorithms (Mirsky & Lee, 2021; Rana *et al.*, 2022; Tolosana *et al.*, 2020). Psychology and communication research measures how well people detect synthetic media (Diel *et al.*, 2024; Somoray *et al.*, 2025). Information-systems research explains phishing susceptibility and security behaviour (Desolda *et al.*, 2021; Vishwanath *et al.*, 2011). Education research addresses AI literacy and academic integrity in the age of ChatGPT (Kasneji *et al.*, 2023; Ng *et al.*, 2021). Reviews of cybersecurity in higher education predate the generative AI shift (Ulven & Wangen, 2021), and our own earlier review addressed the broader threat landscape of digital learning rather than AI-enabled deception specifically (Osijirin *et al.*, 2026b). No review has yet asked the integrated question that matters to university leaders and educators: given what we know about AI-enabled deception and the limits of human detection, which combination of literacy, procedure and technology can realistically protect students and staff?

This review addresses that gap. It is guided by four questions:

- **RQ1.** Which AI-enabled social engineering threat vectors are most relevant to higher-education institutions?
- **RQ2.** What does the evidence say about people's ability to detect AI-generated deception, and which factors shape it?

- **RQ3.** Which individual, procedural, technical and institutional defences have evidence of effectiveness, and what are their limits?
- **RQ4.** What gaps remain, particularly for universities in the Global South?

The paper makes three contributions. First, it offers a higher-education-specific taxonomy of AI-enabled social engineering. Second, it synthesises human detection evidence across text, image, audio and video, and draws out implications for training design. Third, it proposes a Layered Human–Technical Defence Model with testable propositions and a prioritised research agenda.

2. CONCEPTUAL BACKGROUND

2.1 Social engineering and the psychology of persuasion

Social engineering attacks follow a familiar sequence of reconnaissance, pretext construction, engagement and exploitation, carried through channels such as email, phone, messaging and in-person contact (Krombholz *et al.*, 2015; Salahdine & Kaabouch, 2019). Their success rests on persuasion cues such as authority, urgency, scarcity, reciprocity and social proof, which trigger fast, heuristic judgement rather than careful scrutiny. The Integrated Information Processing Model of phishing (Vishwanath *et al.*, 2011) showed that attention to urgency cues and habitual email use predicted victimisation more than knowledge did. A systematic review of human factors in phishing (Desolda *et al.*, 2021) similarly concluded that situational and cognitive factors, not just awareness, explain who falls for phishing.

Dual-process accounts help explain why. People who rely on intuition rather than analytical reasoning are more susceptible to false content, a pattern that reflects "lazy" thinking more than motivated bias (Pennycook & Rand, 2019). Protection Motivation Theory (Rogers, 1975) and self-efficacy theory (Bandura, 1977) add that protective behaviour depends on perceived threat, perceived efficacy of the response and confidence in one's own ability to carry it out. These constructs are well established in studies of security policy compliance (Ifinedo, 2012) and security behaviour (Ng *et al.*, 2009).

2.2 Generative AI as a force multiplier

Generative AI does not create new categories of social engineering so much as it removes the constraints that limited existing ones. Schmitt and Flechais (2024) describe three effects: realism (fluent text and lifelike media), personalisation at scale (targeting informed by scraped data) and automation (end-to-end campaigns with little human effort). Gupta *et al.*, (2023) document how safety guardrails on conversational models can be bypassed to produce malicious content. Deepfake generation, built on autoencoders, generative adversarial networks and, more recently, diffusion models, has

moved from research laboratories to consumer apps (Mirsky & Lee, 2021; Westerlund, 2019). The broader social effect may be as important as individual deceptions. Deepfakes raise general uncertainty about what is real, which erodes trust even when people are not directly fooled (Vaccari & Chadwick, 2020; Hancock & Bailenson, 2021).

2.3 Literacies as protective resources

Three literacies recur in the defence literature. Cybersecurity awareness covers knowledge of threats and safe practices (Zwilling *et al.*, 2022). Digital and media literacy covers the ability to evaluate sources, recognise manipulation techniques and verify claims (Guess *et al.*, 2020; Roozenbeek & van der Linden, 2019). AI literacy is the ability to understand, use and critically evaluate AI systems, including knowing how generative systems work and where they fail (Long & Magerko, 2020; Ng *et al.*, 2021). These literacies overlap, and the evidence on how they combine to protect against AI-enabled fraud is still emerging.

3. METHOD

3.1 Review design

We used an integrative review design because the evidence base is heterogeneous. It combines experimental detection studies, surveys, systematic reviews, technical surveys and policy reports, and the aim was to build a conceptual model rather than to estimate a single pooled effect (Snyder, 2019; Torraco, 2005; Whittemore & Knafl, 2005). The review followed the five stages Whittemore and Knafl recommend: problem identification, literature search, data evaluation, data analysis and presentation.

3.2 Search strategy

Searches covered Scopus, Web of Science Core Collection, ScienceDirect, SpringerLink, IEEE Xplore and the ACM Digital Library for January 2017 to September 2026. 2017 marks the public emergence of "deepfakes". The core Scopus string was:

TITLE-ABS-KEY ((deepfake* OR "synthetic media" OR "voice clon*" OR "generative AI" OR "large language model*" OR "AI-generated") AND (phish* OR "social engineering" OR scam* OR fraud* OR deception OR impersonat* OR detect*) AND (student* OR universit* OR "higher education" OR human* OR user* OR participant*)) AND PUBYEAR > 2016

Equivalent strings were adapted for each database. Backward and forward citation chasing from key reviews (Desolda *et al.*, 2021; Diel *et al.*, 2024; Schmitt & Flechais, 2024; Somoray *et al.*, 2025; Ulven & Wangen, 2021) identified further studies. Foundational theory (for example, protection motivation and self-efficacy) and pre-2017 phishing studies were included where they anchor the conceptual model.

3.3 Eligibility and appraisal

Sources were eligible if they (a) reported empirical data on AI-enabled or synthetic-media deception, human detection performance, phishing or scam susceptibility, or the effect of an educational or procedural intervention; or (b) were systematic reviews, technical surveys or authoritative institutional reports on these topics. Opinion pieces without evidence were excluded. Empirical studies were appraised for design, sample, measurement of detection (accuracy, sensitivity, d') and ecological validity, following the concern raised by Somoray *et al.*, (2025) that heterogeneous metrics make detection studies hard to compare. Table 1 summarises key empirical evidence. The full synthesis draws on the sources listed in the reference list.

3.4 Synthesis

Evidence was coded deductively against the four research questions and inductively for emergent themes. Codes were then organised into the layers of the conceptual model (Figure 1), and propositions were derived where convergent evidence supported a directional claim.

4. RESULTS

4.1 RQ1: AI-enabled threat vectors in higher education

The literature supports six threat vectors with clear relevance to universities (Table 2).

T1. LLM-generated phishing and spear-phishing.

LLMs remove the spelling and grammar errors that awareness campaigns have long taught users to look for, and they adapt tone and content to the target (Schmitt & Flechais, 2024). Experimental evidence shows that automated LLM emails can approach expert-crafted emails in effectiveness and can be produced at negligible cost (Hazell, 2023; Heiding *et al.*, 2024). In universities, plausible pretexts are abundant: registration deadlines, results releases, scholarship awards, accommodation, examination timetables and fee payment.

T2. Voice-clone impersonation.

Synthetic speech can now imitate a specific speaker. Mai *et al.*, (2023) found that listeners identified speech deepfakes with only about 73% accuracy, and training brought only small improvements. In universities, plausible targets include calls that appear to come from a lecturer asking a student to pay for "result correction", from a parent requesting urgent money, or from a senior officer instructing a bursary clerk.

T3. Deepfake video calls and identity verification.

Real-time face and voice synthesis allows impersonation in live calls. The Hong Kong case shows that a multi-person deepfake conference can override normal organisational caution (Chen & Magramo, 2024). As universities expand online interviews, remote proctoring and video-based identity checks, synthetic

video creates risks for admissions, recruitment and assessment integrity.

T4. Synthetic personas.

AI-generated faces are indistinguishable from real faces for most viewers and are rated as *more* trustworthy (Nightingale & Farid, 2022). This enables convincing fake profiles for romance fraud, sextortion and fake job, internship and scholarship offers. Cross (2022) shows that synthetic images defeat reverse-image search, which has been a standard piece of advice in romance-fraud prevention.

T5. AI-built fake portals and payment pages.

Generative tools speed up the creation of convincing clone websites for student portals, learning

management systems and payment gateways. Combined with T1, they enable credential-harvesting campaigns timed to the academic calendar. This vector extends long-standing phishing anatomies (Alkhalil *et al.*, 2021) rather than replacing them.

T6. Deepfake image-based abuse and reputational attacks.

Non-consensual synthetic intimate images and fabricated audio or video of staff and students can be used for harassment, extortion or reputational damage (Westerlund, 2019; Hancock & Bailenson, 2021). Although less often framed as "cybersecurity", these harms fall on the same population and call for coordinated institutional responses.

Table 1: Key empirical evidence on human detection of AI-generated deception

Study	Modality	Design and sample	Main finding relevant to this review
Köbis <i>et al.</i> , (2021), <i>iScience</i>	Video	Experiment, N = 210	Detection was unreliable; neither raising awareness nor financial incentives improved it; participants overestimated their ability and leaned towards judging videos authentic.
Nightingale & Farid (2022), <i>PNAS</i>	Image (faces)	Experiments on StyleGAN2 faces	Accuracy was near chance and improved only modestly with training; synthetic faces were rated more trustworthy than real ones.
Groh <i>et al.</i> , (2022), <i>PNAS</i>	Video	Large online experiment comparing crowds with a leading model	Crowds performed comparably to the model; human-machine combinations could help but were vulnerable to wrong machine advice.
Mai <i>et al.</i> , (2023), <i>PLOS ONE</i>	Audio (speech)	Experiment, English and Mandarin	Listeners detected speech deepfakes at about 73% accuracy; training gave only small gains.
Bray <i>et al.</i> , (2023), <i>Journal of Cybersecurity</i>	Image (faces)	Online experiment	Accuracy was modestly above chance; familiarisation interventions gave little improvement; confidence was poorly calibrated.
Hwang <i>et al.</i> , (2021), <i>Cyberpsychology, Behavior, and Social Networking</i>	Video	Experiment	Deepfake disinformation was more persuasive than text alone; media-literacy education reduced its effect.
Appel & Prietzel (2022), <i>Journal of Computer-Mediated Communication</i>	Video (political)	Experiments	Analytical thinking and political interest were associated with better detection.
Delchev <i>et al.</i> , (2024), <i>Springer Computing Conference</i>	Video	Survey task, N = 179 UK computer science students	Even technically trained students could not reliably identify a deepfake video among three on a mobile phone.
Heiding <i>et al.</i> , (2024), <i>IEEE Access</i>	Text (email)	Field experiment with phishing emails	LLM-generated phishing achieved click-through rates approaching expert-crafted emails; LLMs also showed promise in detecting phishing intent.
Diel <i>et al.</i> , (2024), <i>Computers in Human Behavior Reports</i>	All	Systematic review and meta-analysis, 56 papers	Pooled accuracy was 55.54%; confidence intervals for most modalities crossed chance; detection-improvement strategies raised accuracy to about 65%.
Somoray <i>et al.</i> , (2025), <i>Human Behavior and Emerging Technologies</i>	All	Systematic review	Accuracy varies widely and is shaped by person- and stimulus-level factors; interventions gave mixed results; inconsistent metrics limit comparison.

Table 2: Taxonomy of AI-enabled social engineering threats in higher education

Vector	Typical university scenarios	Primary target	Persuasion levers exploited
T1 LLM phishing / spear-phishing	Fake portal-update, results, scholarship or fee notices; supervisor impersonation	Students, staff, researchers	Authority, urgency, relevance
T2 Voice-clone impersonation	"Lecturer" asks for payment to fix a grade; "parent" in distress; "registrar" instructs bursary	Students, finance staff, parents	Authority, emotional urgency
T3 Deepfake video / identity spoofing	Fake senior-officer video calls; spoofed online interviews and proctoring	Finance and admissions staff	Authority, social proof (multiple "colleagues")
T4 Synthetic personas	Romance and sextortion; fake recruiters, internships and foreign scholarships	Students (especially final-year and postgraduate)	Liking, scarcity, reciprocity
T5 AI-built fake portals	Cloned LMS, email or payment pages timed to the academic calendar	Students and staff	Familiarity, urgency
T6 Deepfake image-based abuse	Synthetic intimate images; fabricated recordings of staff	Students (disproportionately women), staff	Fear, shame, reputational threat

4.2 RQ2: The limits of human detection

4.2.1 Accuracy is close to chance

The most robust finding in this literature is that unaided people are poor at detecting synthetic media. Diel *et al.*'s (2024) meta-analysis of 56 papers estimated total detection accuracy at 55.54% (95% CI 48.87–62.10). Accuracy was 62.08% for audio, 53.16% for images, 52.00% for text and 57.31% for video, and confidence intervals crossed 50% for most modalities. Somoray *et al.*, (2025) reached a similar conclusion in a separate systematic review: accuracy varies widely across studies, and the field lacks common measurement standards. Individual experiments point the same way across faces (Nightingale & Farid, 2022; Bray *et al.*, 2023), video (Köbis *et al.*, 2021) and speech (Mai *et al.*, 2023). The evidence on students specifically is thin but consistent. Even computer science students could not reliably pick out a deepfake on a mobile phone (Delchev *et al.*, 2024).

4.2.2 Overconfidence and a bias towards believing

Two further patterns make the accuracy problem worse. First, people overestimate their detection ability. Köbis *et al.*, (2021) found that participants were confident despite near-chance performance, and Bray *et al.*, (2023) found that confidence was poorly calibrated to accuracy. Second, judgements lean towards "real". People are more likely to accept fakes as authentic than to reject real content, the "seeing-is-believing" heuristic (Köbis *et al.*, 2021). Together, these patterns mean that people who feel prepared may be the least protected, which has direct implications for how preparedness is measured in survey research.

4.2.3 Person- and stimulus-level moderators

Detection is shaped by analytical thinking (Appel & Prietzel, 2022; Pennycook & Rand, 2019), by cognitive ability and network characteristics (Ahmed, 2021) and by stimulus quality and modality (Diel *et al.*,

2024; Somoray *et al.*, 2025). For phishing more broadly, habitual and hurried email processing, attention to urgency cues and impulsivity increase susceptibility (Hadlington, 2017; Vishwanath *et al.*, 2011). Viewing conditions matter too. Small screens, compression and low-bandwidth playback hide the artefacts that detection advice tells people to look for. This is especially relevant where students mainly use mobile phones. Many of these moderators have not been tested in African or South Asian student populations, and we treat this as a priority gap (Section 6).

4.2.4 Generative AI erodes the old cues

Classic awareness training taught people to look for spelling errors, generic greetings and odd formatting. LLMs remove these cues (Schmitt & Flechais, 2024). Advice to look for unnatural blinking, mismatched lighting or lip-sync errors in video is also becoming outdated as generation quality improves (Tolosana *et al.*, 2020; Rana *et al.*, 2022). Training built around artefact-spotting therefore has a short shelf life. More durable advice focuses on the *context* of a request (is this channel, timing and demand normal?) rather than on the *content*.

4.3 RQ3: What defends against AI-enabled deception?

4.3.1 Individual layer: literacies and self-efficacy

Cybersecurity awareness training reduces phishing susceptibility, but effects decay over time and depend on design. Embedded, repeated and contextual training outperforms one-off sessions (Jampen *et al.*, 2020). Awareness alone does not reliably produce protective behaviour (Zwilling *et al.*, 2022), which is consistent with the gap between knowledge and action in the security-behaviour literature (Ifinedo, 2012; Ng *et al.*, 2009).

Media-literacy and inoculation interventions have the strongest experimental record against

misinformation. A short media-literacy intervention improved discernment between true and false news in the United States and India, although effects faded (Guess *et al.*, 2020). "Prebunking" games and videos improved recognition of manipulation techniques (Roozenbeek & van der Linden, 2019; Roozenbeek *et al.*, 2022). For deepfakes specifically, media-literacy education reduced the persuasive effect of deepfake disinformation (Hwang *et al.*, 2021). These interventions, however, target belief in false news. Whether they transfer to targeted, interactive fraud, such as a cloned voice asking for money, is largely untested.

Detection training for synthetic media gives modest gains. Diel *et al.*, (2024) found that detection-improvement strategies raised pooled accuracy to about 65%, but individual studies report small or null effects (Köbis *et al.*, 2021; Mai *et al.*, 2023), and Somoray *et al.*, (2025) characterise the evidence as mixed.

AI literacy is a promising but under-tested protective factor. Frameworks emphasise understanding how AI systems work, using them and critically evaluating their outputs (Long & Magerko, 2020; Ng *et al.*, 2021). The plausible mechanism is that students who understand how easily synthetic media is produced will hold a more realistic prior about what might be fake, and so rely less on their eyes and ears and more on verification. Empirical tests of this mechanism are scarce.

Security self-efficacy links knowledge to action. Within protection motivation and social cognitive theory, people act when they believe a response works and that they can carry it out (Bandura, 1977; Rogers, 1975). For deepfakes, efficacy should be directed at verification behaviours that work regardless of media quality (calling back on a known number, checking through an official channel) rather than at perceptual detection, where confidence is already miscalibrated.

4.3.2 Procedural layer: verification over detection

Given near-chance perception, the most reliable defence is procedural. Requests involving money, credentials or sensitive data should be verified through an independent, pre-established channel, whatever the request looks or sounds like. Fraud-prevention scholarship increasingly recommends this shift from "spot the fake" to "verify the request" (Cross, 2022; Europol, 2022). In universities, this can mean published rules ("staff will never ask you to pay into a personal account"), call-back procedures for bursary and finance offices, code words for high-risk transactions and a single trusted portal for all payment instructions.

4.3.3 Technical layer: authentication, provenance and detection

Technical controls reduce how much depends on human judgement. Multi-factor authentication limits the damage of stolen credentials. Email authentication

(SPF, DKIM and DMARC) makes domain spoofing harder. Automated deepfake detectors have advanced (Rana *et al.*, 2022; Tolosana *et al.*, 2020) but generalise poorly to new generation methods and should support, not replace, procedural checks. Groh *et al.*, (2022) caution that human-machine teaming can fail when the machine is wrong. Content-provenance standards and labelling obligations, such as the transparency duties for deepfakes in the European Union's AI Act (Regulation (EU) 2024/1689), may help over time, but they will not cover content produced by criminals.

4.3.4 Institutional layer: policy, reporting and support

Higher-education cybersecurity reviews stress governance, risk management and awareness programmes alongside technology (Ulven & Wangen, 2021). For AI-enabled fraud, institutions also need clear reporting routes, fast takedown and support for victims, especially victims of image-based abuse and sextortion, and joint work with law enforcement under national frameworks such as Nigeria's Cybercrimes (Prohibition, Prevention, etc.) Act. Institutional communication also shapes the "normal" against which students judge requests. Inconsistent channels, such as staff sometimes using personal WhatsApp accounts to request payments, make fraudulent requests harder to distinguish.

4.4 RQ4: Gaps in the evidence

The synthesis reveals five gaps. (1) Geographic concentration: detection experiments are dominated by samples from North America, Western Europe and East Asia, and evidence from sub-Saharan African universities is largely limited to descriptive surveys. (2) Fraud versus misinformation: most intervention research targets belief in false news rather than interactive financial fraud. (3) Measurement: inconsistent detection metrics (Somoray *et al.*, 2025) and heavy reliance on self-reported preparedness, which is vulnerable to the overconfidence documented above. (4) Language and accent: very few studies test synthetic speech or text in African languages or accented English, where both generation quality and listener expectations may differ. (5) Institutional outcomes: very little research evaluates university-level procedural controls, such as call-back rules or payment-channel policies, against real incident data.

5. A Layered Human-Technical Defence Model

Figure 1 integrates the findings into a Layered Human-Technical Defence Model (LHTDM). The model reads from left to right. AI-enabled threat vectors (T1-T6) exploit human detection limits (H1-H5). These limits are compensated for by defences at four layers (D1-D7), which produce outcomes ranging from individual fraud-detection preparedness to institutional resilience. Contextual moderators, emphasised for the Global South, condition both the effectiveness of defences and the outcomes.

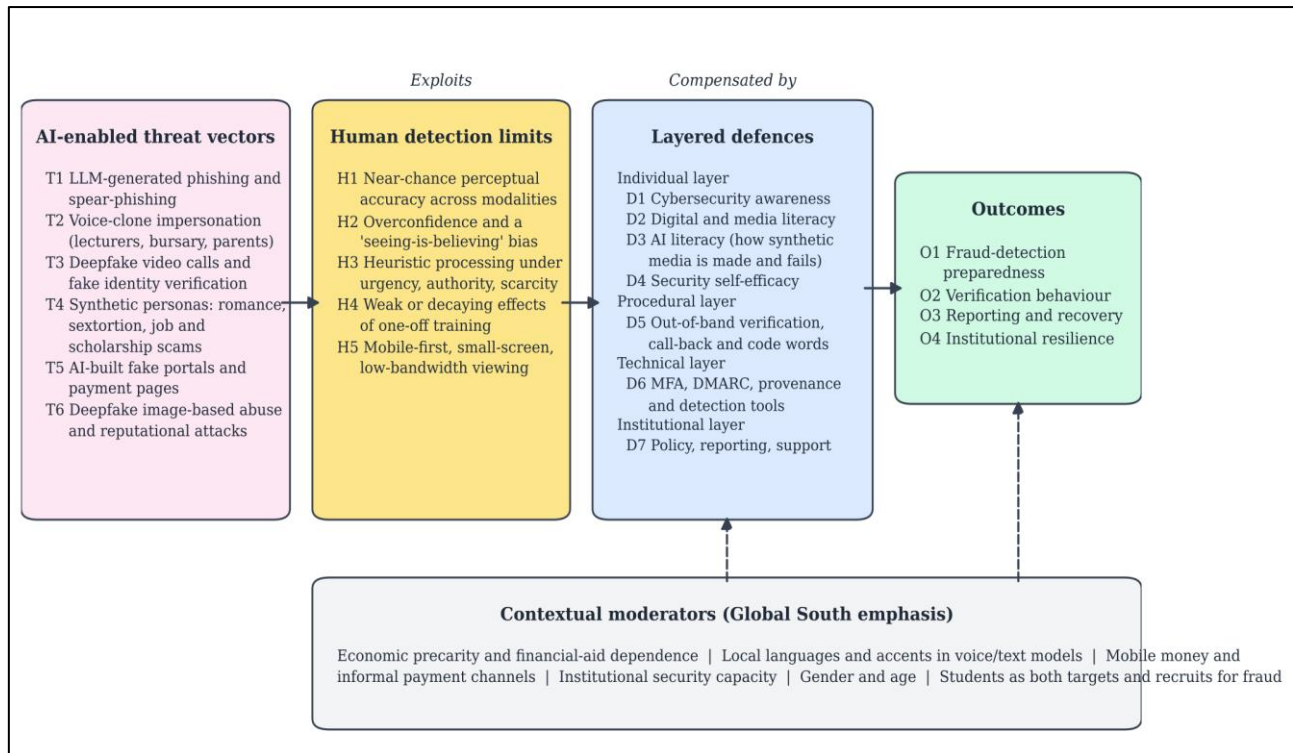


Figure 1: The Layered Human-Technical Defence Model (LHTDM) for AI-enabled social engineering in higher education.
Source: author's synthesis

The model's central claim is that, as generative quality improves, the burden of defence must shift from perception (can I tell this is fake?) to verification (is this request legitimate, whatever it looks like?). The evidence supports six propositions for empirical testing:

- **P1.** Cybersecurity awareness, digital literacy and media literacy will each predict fraud-detection preparedness, but their combined effect will be greater than any single literacy alone.
- **P2.** AI literacy will predict verification behaviour more strongly than it predicts perceptual deepfake-detection accuracy.
- **P3.** Self-reported confidence in detecting deepfakes will be weakly related, or unrelated, to measured detection accuracy. Studies relying only on self-report will therefore overestimate preparedness.
- **P4.** Interventions that train verification routines will produce larger and more durable reductions in fraud victimisation than interventions that train artefact-spotting.
- **P5.** Institutional procedural controls (a single payment channel, published "we will never ask" rules, call-back procedures) will moderate the relationship between individual literacy and victimisation, reducing the role of individual skill.
- **P6.** Economic precarity and dependence on financial aid will increase susceptibility to scholarship, job and payment-related AI fraud (T1, T4, T5), independently of literacy.

6. Research agenda for Global South universities

6.1 Measure behaviour, not only perception.

Survey studies in Nigeria and elsewhere have usefully mapped awareness and literacy (Osijirin *et al.*, 2026a), but the overconfidence findings mean that self-reported preparedness must be complemented with objective tasks: deepfake detection tests, simulated phishing with ethical approval, and scenario-based verification choices. Reporting should adopt consistent metrics (accuracy with confidence intervals, sensitivity, d') to allow future meta-analysis (Somoray *et al.*, 2025).

6.2 Test locally realistic stimuli.

Studies should use synthetic voices and texts in Nigerian Pidgin, Igbo, Yoruba, Hausa and accented English, and pretexts that match local fraud scripts (result-upgrade payments, fake federal scholarships, recruitment scams). Mobile-first presentation conditions should be standard.

6.3 Evaluate interventions with designs that allow causal claims.

Randomised or quasi-experimental trials should compare artefact-spotting training, inoculation-style prebunking adapted to fraud, and verification-routine training, with follow-ups to measure decay (Jampen *et al.*, 2020; Roozenbeek *et al.*, 2022).

6.4 Study the institution, not only the student.

Researchers should partner with bursary, ICT and student-affairs offices to evaluate procedural

controls against incident records, and audit official communication channels for inconsistency.

6.5 Address the dual role of students.

Where students may be drawn into fraud networks (Tade & Aliyu, 2011; Ojedokun & Eraye, 2012), AI literacy curricula should cover the ethics and criminal liability of using generative tools for deception, not only how to avoid victimisation.

6.6 Attend to gender and image-based abuse.

Deepfake image-based abuse falls disproportionately on women. Research should examine prevalence, reporting barriers and institutional support in African universities, where evidence is almost absent.

7. Practical implications

For university leaders, the evidence points to a few low-cost, high-value actions: publish a short list of things the institution will never ask students or staff to do; channel all payments through a single authenticated portal; require call-back verification for any change in payment details or urgent transfer request; enforce multi-factor authentication and DMARC; and set up a simple reporting and support route for fraud and image-based abuse. For educators, cybersecurity and media-literacy content should be redesigned around verification habits and realistic local scenarios, embedded across the curriculum rather than delivered once at orientation. For students, the most protective single habit is to treat any unexpected request for money, codes or credentials as unverified until confirmed through a channel they already trust, however convincing the voice or video.

8. Limitations

As an integrative review, this study prioritises conceptual synthesis over statistical pooling. Although searches covered major databases, relevant grey literature, non-English studies and institutional incident reports may have been missed. The field is moving quickly, and generation quality, and therefore detection accuracy, may change faster than the literature can record. The propositions are derived from convergent evidence but remain to be tested, particularly in Global South contexts where direct evidence is limited.

9. CONCLUSION

Generative AI has lowered the cost and raised the quality of social engineering, and universities present an attractive combination of open systems, financial flows and young, mobile-first users. The evidence is clear that people cannot reliably see or hear their way to safety: detection accuracy hovers near chance, and confidence outruns competence. Literacy still matters, but its main role is to motivate verification rather than to sharpen perception. The Layered Human–Technical Defence Model proposed here offers a structure for research and practice that combines individual literacies, verification procedures, technical controls and institutional policy. Testing it in the universities of

Africa and the wider Global South, where exposure is high and evidence is thin, is the most pressing next step.

DECLARATIONS

Funding: No external funding was received for this work.

Conflict of interest: The author declares no competing interests.

Ethics approval: Not applicable; this review analyses published literature.

Data availability: No new data were generated. Search strings are reported in Section 3.2.

Author contributions: S.M.S.: Conceptualization, Methodology, Formal analysis, Investigation, Data curation, Software, Visualization, Writing – original draft, Writing – review & editing.

N.H.N: Supervision, Data curation, Validation, Writing – review & editing.

A.N.O.: Methodology, Writing – review & editing.

REFERENCES

- Ahmed, S. (2021). Who inadvertently shares deepfakes? Analyzing the role of political interest, cognitive ability, and social network size. *Telematics and Informatics*, 57, 101508. <https://doi.org/10.1016/j.tele.2020.101508>
- Alkhalil, Z., Hewage, C., Nawaf, L., & Khan, I. (2021). Phishing attacks: A recent comprehensive study and a new anatomy. *Frontiers in Computer Science*, 3, 563060. <https://doi.org/10.3389/fcomp.2021.563060>
- Appel, M., & Prietzel, F. (2022). The detection of political deepfakes. *Journal of Computer-Mediated Communication*, 27(4), zmac008. <https://doi.org/10.1093/jcmc/zmac008>
- Bandura, A. (1977). Self-efficacy: Toward a unifying theory of behavioral change. *Psychological Review*, 84(2), 191–215. <https://doi.org/10.1037/0033-295X.84.2.191>
- Bray, S. D., Johnson, S. D., & Kleinberg, B. (2023). Testing human ability to detect 'deepfake' images of human faces. *Journal of Cybersecurity*, 9(1), tyad011. <https://doi.org/10.1093/cybsec/tyad011>
- Chen, H., & Magramo, K. (2024, February 4). Finance worker pays out \$25 million after video call with deepfake 'chief financial officer'. *CNN*. <https://edition.cnn.com/2024/02/04/asia/deepfake-cfo-scam-hong-kong-intl-hnk/index.html>
- Cross, C. (2022). Using artificial intelligence (AI) and deepfakes to deceive victims: The need to rethink current romance fraud prevention messaging. *Crime Prevention and Community Safety*, 24(1), 30–41. <https://doi.org/10.1057/s41300-021-00134-w>

- Delchev, K., Safieddine, F., & Hammad, R. (2024). Identification of A.I. generated deep fake video by higher education students. In *Intelligent Computing: Proceedings of the 2024 Computing Conference* (Lecture Notes in Networks and Systems). Springer.
- Desolda, G., Ferro, L. S., Marrella, A., Catarci, T., & Costabile, M. F. (2021). Human factors in phishing attacks: A systematic literature review. *ACM Computing Surveys*, 54(8), 1–35. <https://doi.org/10.1145/3469886>
- Diel, A., Lalgı, T., Schröter, I. C., MacDorman, K. F., Teufel, M., & Bäuerle, A. (2024). Human performance in detecting deepfakes: A systematic review and meta-analysis of 56 papers. *Computers in Human Behavior Reports*, 16, 100538. <https://doi.org/10.1016/j.chbr.2024.100538>
- Europol. (2022). *Facing reality? Law enforcement and the challenge of deepfakes* (Europol Innovation Lab Observatory Report). Publications Office of the European Union.
- European Union. (2024). Regulation (EU) 2024/1689 of the European Parliament and of the Council laying down harmonised rules on artificial intelligence (Artificial Intelligence Act). *Official Journal of the European Union*, L, 2024/1689.
- Groh, M., Epstein, Z., Firestone, C., & Picard, R. (2022). Deepfake detection by human crowds, machines, and machine-informed crowds. *Proceedings of the National Academy of Sciences*, 119(1), e2110013119. <https://doi.org/10.1073/pnas.2110013119>
- Guess, A. M., Lerner, M., Lyons, B., Montgomery, J. M., Nyhan, B., Reifler, J., & Sircar, N. (2020). A digital media literacy intervention increases discernment between mainstream and false news in the United States and India. *Proceedings of the National Academy of Sciences*, 117(27), 15536–15545. <https://doi.org/10.1073/pnas.1920498117>
- Gupta, M., Akiri, C., Aryal, K., Parker, E., & Praharaj, L. (2023). From ChatGPT to ThreatGPT: Impact of generative AI in cybersecurity and privacy. *IEEE Access*, 11, 80218–80245. <https://doi.org/10.1109/ACCESS.2023.3300381>
- Hadlington, L. (2017). Human factors in cybersecurity; examining the link between Internet addiction, impulsivity, attitudes towards cybersecurity, and risky cybersecurity behaviours. *Heliyon*, 3(7), e00346. <https://doi.org/10.1016/j.heliyon.2017.e00346>
- Hancock, J. T., & Bailenson, J. N. (2021). The social impact of deepfakes. *Cyberpsychology, Behavior, and Social Networking*, 24(3), 149–152. <https://doi.org/10.1089/cyber.2021.29208.jth>
- Hazell, J. (2023). *Spear phishing with large language models* (arXiv:2305.06972). arXiv. <https://doi.org/10.48550/arXiv.2305.06972>
- Heiding, F., Schneier, B., Vishwanath, A., Bernstein, J., & Park, P. S. (2024). Devising and detecting phishing emails using large language models. *IEEE Access*, 12, 42131–42146. <https://doi.org/10.1109/ACCESS.2024.3375882>
- Hwang, Y., Ryu, J. Y., & Jeong, S.-H. (2021). Effects of disinformation using deepfake: The protective effect of media literacy education. *Cyberpsychology, Behavior, and Social Networking*, 24(3), 188–193. <https://doi.org/10.1089/cyber.2020.0174>
- Ifinedo, P. (2012). Understanding information systems security policy compliance: An integration of the theory of planned behavior and the protection motivation theory. *Computers & Security*, 31(1), 83–95. <https://doi.org/10.1016/j.cose.2011.10.007>
- Jampen, D., Gür, G., Sutter, T., & Tellenbach, B. (2020). Don't click: Towards an effective anti-phishing training. A comparative literature review. *Human-centric Computing and Information Sciences*, 10, 33. <https://doi.org/10.1186/s13673-020-00237-7>
- Kasneci, E., Sessler, K., Küchemann, S., Bannert, M., Dementieva, D., Fischer, F., et al., (2023). ChatGPT for good? On opportunities and challenges of large language models for education. *Learning and Individual Differences*, 103, 102274. <https://doi.org/10.1016/j.lindif.2023.102274>
- Köbis, N. C., Doležalová, B., & Soraperra, I. (2021). Fooled twice: People cannot detect deepfakes but think they can. *iScience*, 24(11), 103364. <https://doi.org/10.1016/j.isci.2021.103364>
- Krombholz, K., Hobel, H., Huber, M., & Weippl, E. (2015). Advanced social engineering attacks. *Journal of Information Security and Applications*, 22, 113–122. <https://doi.org/10.1016/j.jisa.2014.09.005>
- Long, D., & Magerko, B. (2020). What is AI literacy? Competencies and design considerations. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems* (pp. 1–16). ACM. <https://doi.org/10.1145/3313831.3376727>
- Mai, K. T., Bray, S., Davies, T., & Griffin, L. D. (2023). Warning: Humans cannot reliably detect speech deepfakes. *PLOS ONE*, 18(8), e0285333. <https://doi.org/10.1371/journal.pone.0285333>
- Mirsky, Y., & Lee, W. (2021). The creation and detection of deepfakes: A survey. *ACM Computing Surveys*, 54(1), 1–41. <https://doi.org/10.1145/3425780>
- Ng, B.-Y., Kankanhalli, A., & Xu, Y. C. (2009). Studying users' computer security behavior: A health belief perspective. *Decision Support Systems*, 46(4), 815–825. <https://doi.org/10.1016/j.dss.2008.11.010>
- Ng, D. T. K., Leung, J. K. L., Chu, S. K. W., & Qiao, M. S. (2021). Conceptualizing AI literacy: An exploratory review. *Computers and Education: Artificial Intelligence*, 2, 100041. <https://doi.org/10.1016/j.caeai.2021.100041>
- Nightingale, S. J., & Farid, H. (2022). AI-synthesized faces are indistinguishable from real

- faces and more trustworthy. *Proceedings of the National Academy of Sciences*, 119(8), e2120481119. <https://doi.org/10.1073/pnas.2120481119>
- Ojedokun, U. A., & Eraye, M. C. (2012). Socioeconomic lifestyles of the Yahoo-boys: A study of perceptions of university students in Nigeria. *International Journal of Cyber Criminology*, 6(2), 1001–1013.
 - Osijirin, A. N., Sada, S. M., Edmond, V. U., Anigbo, L. C., & Okechukwu, O. (2026a). AI-powered scams and deepfakes in tertiary institutions in Enugu State, Nigeria: The roles of cybersecurity awareness, digital literacy and media literacy in students' fraud detection preparedness. *Saudi Journal of Engineering and Technology*. <https://doi.org/10.36348/sjet.2026.v11i04.021>
 - Osijirin, A. N., Anigbo, L. C., Okechukwu, O., Chima, A. O., & Sada, S. M. (2026b). Cybersecurity and digital learning in higher education: Emerging threats, protective technologies, and future directions. *Asian Journal of Research in Computer Science*. <https://doi.org/10.9734/ajrcos/2026/v19i7884>
 - Pennycook, G., & Rand, D. G. (2019). Lazy, not biased: Susceptibility to partisan fake news is better explained by lack of reasoning than by motivated reasoning. *Cognition*, 188, 39–50. <https://doi.org/10.1016/j.cognition.2018.06.011>
 - Rana, M. S., Nobi, M. N., Murali, B., & Sung, A. H. (2022). Deepfake detection: A systematic literature review. *IEEE Access*, 10, 25494–25513. <https://doi.org/10.1109/ACCESS.2022.3154404>
 - Rogers, R. W. (1975). A protection motivation theory of fear appeals and attitude change. *The Journal of Psychology*, 91(1), 93–114. <https://doi.org/10.1080/00223980.1975.9915803>
 - Roozenbeek, J., & van der Linden, S. (2019). Fake news game confers psychological resistance against online misinformation. *Palgrave Communications*, 5, 65. <https://doi.org/10.1057/s41599-019-0279-9>
 - Roozenbeek, J., van der Linden, S., Goldberg, B., Rathje, S., & Lewandowsky, S. (2022). Psychological inoculation improves resilience against misinformation on social media. *Science Advances*, 8(34), eabo6254. <https://doi.org/10.1126/sciadv.abo6254>
 - Salahdine, F., & Kaabouch, N. (2019). Social engineering attacks: A survey. *Future Internet*, 11(4), 89. <https://doi.org/10.3390/fi11040089>
 - Schmitt, M., & Flechais, I. (2024). Digital deception: Generative artificial intelligence in social engineering and phishing. *Artificial Intelligence Review*, 57, 324. <https://doi.org/10.1007/s10462-024-10973-2>
 - Snyder, H. (2019). Literature review as a research methodology: An overview and guidelines. *Journal of Business Research*, 104, 333–339. <https://doi.org/10.1016/j.jbusres.2019.07.039>
 - Somoray, K., Miller, D. J., & Holmes, M. (2025). Human performance in deepfake detection: A systematic review. *Human Behavior and Emerging Technologies*, 2025, 1833228. <https://doi.org/10.1155/hbe2/1833228>
 - Tade, O., & Aliyu, I. (2011). Social organization of internet fraud among university undergraduates in Nigeria. *International Journal of Cyber Criminology*, 5(2), 860–875.
 - Tolosana, R., Vera-Rodriguez, R., Fierrez, J., Morales, A., & Ortega-Garcia, J. (2020). Deepfakes and beyond: A survey of face manipulation and fake detection. *Information Fusion*, 64, 131–148. <https://doi.org/10.1016/j.inffus.2020.06.014>
 - Torraco, R. J. (2005). Writing integrative literature reviews: Guidelines and examples. *Human Resource Development Review*, 4(3), 356–367. <https://doi.org/10.1177/1534484305278283>
 - Ulven, J. B., & Wangen, G. (2021). A systematic review of cybersecurity risks in higher education. *Future Internet*, 13(2), 39. <https://doi.org/10.3390/fi13020039>
 - Vaccari, C., & Chadwick, A. (2020). Deepfakes and disinformation: Exploring the impact of synthetic political video on deception, uncertainty, and trust in news. *Social Media + Society*, 6(1). <https://doi.org/10.1177/2056305120903408>
 - Vishwanath, A., Herath, T., Chen, R., Wang, J., & Rao, H. R. (2011). Why do people get phished? Testing individual differences in phishing vulnerability within an integrated, information processing model. *Decision Support Systems*, 51(3), 576–586. <https://doi.org/10.1016/j.dss.2011.03.002>
 - Westerlund, M. (2019). The emergence of deepfake technology: A review. *Technology Innovation Management Review*, 9(11), 39–52. <https://doi.org/10.22215/timreview/1282>
 - Whittemore, R., & Knafl, K. (2005). The integrative review: Updated methodology. *Journal of Advanced Nursing*, 52(5), 546–553. <https://doi.org/10.1111/j.1365-2648.2005.03621.x>
 - Zwilling, M., Klien, G., Lesjak, D., Wiecheteck, L., Cetin, F., & Basim, H. N. (2022). Cyber security awareness, knowledge and behavior: A comparative study. *Journal of Computer Information Systems*, 62(1), 82–97. <https://doi.org/10.1080/08874417.2020.1712269>