TEME, Vol. L, Nº 2, April – June 2026, pp. 601–623

Review article
Received: February 23, 2026
Revised: March 10, 2026
Accepted: March 17, 2026

https://doi.org/10.22190/TEME260223036D
UDC 343.851:004.7
343.9:004.738.5
004.738.5:351.817
343.533::004

SOCIAL ENGINEERING AND THE LIMITS OF CYBERCRIME PREVENTION: A REVIEW OF SOCIAL, EDUCATIONAL, AND TECHNOLOGICAL CHALLENGES

Dalibor Doležal[1*], Tanesh Kumar[2]

[1]University of Zagreb, Faculty of Education and Rehabilitation Sciences, Department of Criminology, Zagreb, Croatia
[2]Aalto University, Department of Information and Communications Engineering, Aalto, Finland

ORCID iDs: Dalibor Doležal https://orcid.org/0000-0003-1558-5905
Ivana Marković https://orcid.org/0000-0002-5907-8414

Abstract

Social engineering remains a leading pathway to cyber harm, not primarily because users "lack awareness," but because attackers exploit common cognitive shortcuts, social norms, and organisational routines – now intensified by rapid, platform-mediated communication. This study aims to clarify why social engineering persists as a prevention challenge and to assess the limitations of current prevention strategies. The paper employs a narrative (critical) literature review. Interdisciplinary research on social engineering and cybercrime prevention is synthesised using a thematic-analytic structure across three strands: (1) social vulnerability and victimisation, (2) educational and organisational prevention (training, awareness, culture, routines, and usable warning/interface design), and (3) technological asymmetry and escalation, including AI-enabled deception and deepfake-enabled impersonation. The reviewed literature converges on four findings. First, susceptibility is contextual and patterned, shaped by roles, routines, socio-emotional dynamics, and platform cues rather than stable individual deficits. Second, training and awareness can help but often show limited transfer and sustainability when generic, weakly reinforced, or poorly integrated into workflows. Third, warning-based approaches are constrained by habituation and cognitive load. Fourth, accelerating AI-enabled deception and impersonation reduces the reliability of surface cues, increasing attacker–defender asymmetry. Social engineering should be addressed as a systemic socio-technical prevention problem rather than an "awareness deficit." Effective prevention is layered and institutional: role-sensitive capability-building, verification norms, workflow-based controls (e.g., second-channel checks), usable-security and platform interventions, and governance arrangements that allocate responsibility and enable continuous evaluation. Future research should strengthen behavioural measurement, include longer

* Corresponding author: Dalibor Doležal, Faculty of Education and Rehabilitation Sciences, Department of Criminology, Borongajska cesta 83f, 10000 Zagreb, Croatia, dalibor.dolezal@erf.unizg.hr

follow-up, replicate across contexts, and more closely connect susceptibility studies with organisational and policy analysis.

Key words: Social engineering, cybercrime prevention, cybercrime.

СОЦИЈАЛНИ ИНЖЕЊЕРИНГ И ГРАНИЦЕ ПРЕВЕНЦИЈЕ САЈБЕР КРИМИНАЛА: ПРЕГЛЕД ДРУШТВЕНИХ, ОБРАЗОВНИХ И ТЕХНОЛОШКИХ ИЗАЗОВА

Апстракт

Социјални инжењеринг остаје водећи пут ка штети у сајбер простору, не првенствено зато што корисницима 'недостаје свест', већ зато што нападачи искоришћавају когнитивне пречице, друштвене норме и организационе рутине, појачане комуникацијом посредованом дигиталним платформама. Циљ рада јесте да разјасни зашто социјални инжењеринг представља изазов за превенцију и да процени ограничења стратегија. Рад примењује преглед литературе. Интердисциплинарна истраживања о социјалном инжењерингу и превенцији сајбер криминала синтетизована су кроз три целине: (1) социјалну рањивост и виктимизацију, (2) образовну и организациону превенцију, укључујући обуку, подизање свести, културу, рутине и дизајн упозорења и интерфејса, и (3) технолошку асиметрију и ескалацију, укључујући обману омогућену вештачком интелигенцијом и лажно представљање засновано на дипфејк технологији. Литература указује на четири налаза. Прво, подложност је контекстуална; обликују је улоге, рутине, социо-емоционална динамика и сигнали платформе, а не индивидуални недостаци. Друго, обука и подизање свести могу помоћи, али често имају ограничен пренос и одрживост када су генерички осмишљени или лоше интегрисани у радне токове. Треће, приступи засновани на упозорењима ограничени су хабитуацијом и когнитивним оптерећењем. Четврто, обмана и лажно представљање омогућени вештачком интелигенцијом умањују поузданост сигнала и повећавају асиметрију између нападача и одбране. Социјални инжењеринг треба посматрати као социо-технички проблем превенције, а не као 'дефицит свести'. Делотворна превенција мора бити слојевита: изградња капацитета осетљива на улоге, норме верификације, контроле засноване на радним токовима, интервенције у безбедности платформи и управљачки аранжмани који расподељују одговорност и омогућавају евалуацију. Будућа истраживања треба да унапреде бихејвиорално мерење, укључе дужа праћења, потврде налазе у контекстима и тешње повежу студије подложности са организационом и јавнополитичком анализом.

Кључне речи: социјални инжењеринг, превенција сајбер криминала, сајбер криминал.

INTRODUCTION

The persistence of social engineering cannot be adequately explained by a simple lack of knowledge. Much of what makes these attacks effective is rooted in normal social and cognitive processes and involves the use of so called “weapons of influence” – reciprocity, authority, urgency, scarcity, and the use of heuristics when decisions must be made quickly or with incomplete information (Cialdini, 2007). Social engineer-

ing threats can be classified using three levels, i.e., environment, approaches and mediums. Environment level mainly focuses whether the attacker is physically present or not (Zaoui et al., 2024). Approaches level covers the tactics and psychological triggers attackers use to exploit human behaviour and motivation. Medium level involves the channels and tools through which attacks are carried out.

In digitally connected environments, these processes are amplified by the evolution of various relevant enabling technologies such as AI, smart platform architectures, and communication practices that encourage speed, precision, visibility, and continuous responsiveness. In the modern digital landscape, the nature of social engineering attacks has become much more ubiquitous and can manifested in various forms ranging from physical, social, technical and socio-technical threats (Birthriya et al.,, 2025). This is why social engineering is often described not as a marginal category of cyber offending but as a recurrent feature of contemporary cybercrime: it exploits ordinary interaction rather than exceptional technical weakness (Wang et al., 2020; Zaoui et al., 2024).

Cybercrime prevention has increasingly become a question of how societies manage risk in the current digitally mediated social life and for the future hyperconnected world. Although technical security controls remain essential, many harmful incidents still depend on ordinary human actions: complying with requests, trusting messages, sharing information, approving access, or following routines that offenders can exploit. Social engineering is central to this issue because it relies on manipulation and influence as the primary means of offending, with technology providing distribution, credibility cues, and operational scale (Mitnick and Simon, 2002; Wang et al., 2020). In terms of prevention, social engineering creates a persistent mismatch between the rapid adaptability of offenders and the slower pace at which individuals, institutions, and regulatory systems learn and adjust.

Prevention efforts have typically focused on education and awareness. Governments, organisations, and educational institutions invest in campaigns and training designed to improve detection, decision-making, and reporting. However, recent reviews highlight persistent limitations in these methods. Training is often generic, compliance-focused, or insufficiently reinforced over time, and evaluation frequently prioritises short-term knowledge gains rather than lasting behavioural change in realistic settings (Aldawood and Skinner, 2019; Prümmer et al., 2024; Prümmer et al., 2025). These issues are especially apparent in complex institutional environments such as higher education, where diverse user groups, open communication norms, and extensive digital attack surfaces increase exposure and complicate the standardisation of preventive routines (Alsulami et al., 2021; Abdulla et al., 2023). Usable-security research identifies another challenge: frequent prompts and warnings can cause habituation and dis-

missal, undermining prevention strategies that rely on sustained vigilance (Anderson et al., 2016; Kirwan et al., 2020; Zaaba et al., 2021).

At the same time, literature has become more explicit about the social and organisational dynamics that shape prevention outcomes. One influential critique is that cybersecurity discourse often constructs a "deficient user," implicitly locating failure in individuals rather than in workflows, incentives, system design, or institutional responsibility (Klimburg-Witjes and Wentland, 2021). This is important for crime prevention because responsibility allocation affects which interventions are prioritised and how legitimacy is built: punitive or blame-oriented approaches can discourage reporting, reduce learning, and create a gap between formal policy and everyday practice. From this perspective, education is necessary but unlikely to be sufficient unless it is embedded in organisational routines, supported by leadership, and coupled with technical and procedural safeguards that reduce reliance on perfect human performance (Syafitri et al., 2022; Li et al., 2023).

Recent research also highlights that vulnerability to social engineering is best understood as contextual and patterned, rather than fixed. In social networking settings, susceptibility and victimisation are linked to combinations of habitual behaviour, socio-emotional dynamics, perceptual processing, and message and sender characteristics. Importantly, empirical findings are often heterogeneous across platforms and populations (Albladi and Weir, 2018; Algarni, 2019; Alshammari et al., 2025). This complicates "one-size-fits-all" prevention messaging and supports differentiated approaches that focus on high-exposure roles, decision contexts, and the specific interaction patterns that offenders exploit.

The urgency of these issues is heightened by the rapid technological change. Social engineering has always been adaptive, but current research suggests that automation and generative AI can reduce the cost of producing persuasive, tailored, high-volume deceptive communications, while deepfake-enabled impersonation can undermine the practical reliability of cues people use to judge authenticity (Schmitt and Flechais, 2024; Matecas et al., 2025; Pedersen et al., 2025). This does not render human-centred prevention obsolete, but it does shift the balance towards verification infrastructures and workflow-based controls – authenticated channels, escalation norms, provenance signals, and technical guardrails that support cautious decision-making under time pressure (Li et al., 2023; Pedersen et al., 2025). In other words, the prevention question becomes less about teaching people to "spot" deception and more about designing socio-technical environments in which deception is harder to operationalise and easier to report.

Methodology

For the purposes of this study's main goals, we used a narrative (critical) literature review approach, synthesising interdisciplinary research on

social engineering and cybercrime prevention to develop an integrative argument about the limits of prevention. The review is structured thematically and analytically around three strands that are often treated separately: (1) social vulnerability and victimisation, (2) educational and organisational prevention (training, awareness, culture, routines, and usable warning/interface design), and (3) technological asymmetry and escalation, including AI-enabled deception and deepfake-enabled impersonation. The approach emphasises critical interpretation of heterogeneous evidence (empirical and conceptual), identification of recurring constraints and evidence gaps, and synthesis into a socio-technical account of why layered prevention is necessary but persistently challenged.

A narrative review is a form of knowledge synthesis that can include a wide range of study types and provide an overall account with interpretation and critique, rather than limiting inclusion to narrowly comparable study designs, as is typical in systematic reviews. This approach is particularly appropriate here because social engineering prevention spans multiple disciplines and levels of analysis – individual cognition and routines, organisational culture and workflow, platform and interface design, and rapidly evolving technical infrastructures. As a result, the evidence base is conceptually and methodologically diverse and not easily reducible to a single effect-size question.

The review is also explicitly critical in orientation. In Grant and Booth's (2009) typology a critical review demonstrates that the author has extensively engaged with the literature and critically evaluated its quality, moving beyond description towards analysis and conceptual contribution, often culminating in a model, interpretation, or reframing of the problem. In practice, the paper adopts this critical narrative stance to challenge "awareness-deficit" framings and argues that social engineering should be considered as a systemic socio-technical prevention problem requiring layered, institutionally embedded controls, rather than relying on idealised user vigilance.

Main goal of this paper is to present a narrative review of interdisciplinary literature on social engineering and contemporary prevention approaches. It examines the social, educational, and technological challenges in crime prevention by integrating three strands of evidence that are often discussed separately:

1. Social vulnerability and victimisation: how susceptibility is shaped by social roles, routines, and platform-mediated interaction (Alshammari et al., 2025; Albladi and Weir, 2018).

2. Educational and organisational prevention: what the evidence suggests about training, awareness, organisational culture, and warning/interface design, including recurring limitations in behavioural transfer and sustainability (Aldawood and Skinner, 2019; Prümmer et al., 2025; Zaaba et al., 2021).

3. Technological asymmetry and escalation: how evolving infrastructures – especially AI-enabled deception – reshape the practical feasibility of prevention and increase the importance of verification and governance measures (Schmitt and Flechais, 2024; Pedersen et al., 2025).

From this aim, the goals of this paper are to synthesise interdisciplinary scholarship to explain why social engineering persists as a prevention challenge rooted in ordinary social and cognitive processes amplified by digital infrastructures, to critically assess the limits of dominant prevention approaches, and to identify evidence gaps and research needs.

SOCIAL ENGINEERING IN CYBERCRIME: CONCEPTS, MECHANISMS, AND EVOLVING CONTEXTS

The development of artificial intelligence has increased both the volume and forms of manipulation by lowering production costs, increasing variability, and improving credibility in communication channels (Matecas et al., 2025; Pedersen et al., 2025). AI-assisted social engineering has demonstrated that widely available large language models (LLMs) and diffusion models can rapidly generate convincing templates and support interactive conversational strategies, enabling faster design of diverse message variants used for deception (Matecas et al., 2025). Analysis of deepfake-assisted social engineering targeting companies has shown that organisations remain vulnerable partly because common controls are not designed for high-probability impersonation, and because procedural checks and attack simulations are unevenly institutionalised (Pedersen et al., 2025). These results are important for prevention policy because they undermine strategies that rely on "low-quality" cues and instead highlight the need to strengthen process-level authentication and verification routines, including second-channel verification and role-based approval controls (Pedersen et al., 2025; Nowakowski, 2025). Overall, the AI turn reinforces the socio-technical interpretation of social engineering: technological change does not replace social manipulation, but reorganises its scope and economics, widening the gap between the development of attacker capabilities and prevention strategies that remain focused on individual awareness as the primary defence (Matecas et al., 2025; Aldawood and Skinner, 2019).

Changes in the digital and physical environment – such as rapid digital migration during times of crisis – illustrate how the risk of social engineering is socially embedded and amplified by changes in routines, exposure, and uneven cyber literacy (Venkatesha et al., 2021). A review of social engineering attacks during the COVID-19 pandemic argues that increased online presence without appropriate education creates favourable conditions for exploitation, supporting prevention models that treat vulnerability as a moving target shaped by social change and technological dependency (Venkatesha et al., 2021). Summarising the findings of the rele-

vant literature, social engineering is best conceptualised as a structured, staged, and mechanism-driven form of cyber-enabled deception in which risk arises from interactions between cognitive constraints, message engineering, organisational workflows, and platform capabilities (Wang et al., 2021; Nowakowski, 2025). Prevention strategies that remain focused on the end-user risk misdiagnosing the problem, while approaches that integrate education with screening infrastructures and sociotechnical design are more consistent with the empirical and conceptual evidence base (Klimburg-Witjes and Wentland, 2021; Aldawood and Skinner, 2019).

Technological Asymmetry and Escalation in Contemporary Social Engineering

In the past decade, digital transformation has fundamentally altered the landscape of social engineering. Global social networks, mobile connectivity, and affordable Internet access have created significant opportunities to combine technical and human attack techniques, resulting in higher success rates. The emergence of the "network society", built on instant and inexpensive communication, enables attackers to reach and influence targets on a larger scale and at a faster pace. As technologies have evolved at an exponential rate, attackers have quickly incorporated them into deception strategies, while defenders have often struggled to adapt. This dynamic has produced a form of technological asymmetry – advanced tools in the hands of attackers versus unprepared or underequipped targets – and has led to an escalation or arms race in tactics.

Contemporary social engineering increasingly employs AI-generated content to deceive human targets with a high degree of realism. Generative AI tools, especially large-scale language models and deepfake generators, have introduced sophisticated attack vectors that blur the line between authentic and fraudulent communication (Schmitt and Flechais, 2024). For example, AI-generated phishing emails are more likely to contain correct grammar, context-aware phrasing, and a convincing tone, making them much harder to identify as fraudulent compared to the error-filled spam of the past (Schmitt and Flechais, 2024). Similarly, deepfake technologies enable the creation of hyper-realistic voice messages and video impersonations, allowing attackers to convincingly present themselves as trusted colleagues or public figures (Schmitt and Flechais, 2024). For instance, Collier (2025) notes that voice-deepfake synthesis can be used in family-impersonation vishing (e.g., "grandparent" scams), increasing plausibility and success likelihood This has significantly increased the success rate of high-risk fraud schemes such as business email compromise (BEC) and executive impersonation fraud (Matecaș et al., 2025). These examples illustrate how the asymmetric use of advanced AI provides attackers with new opportunities to exploit human trust, surpassing the ability of victims to distinguish reality from fiction. In addition to targeting live individuals,

social engineering increasingly uses bots and agents, such as AI-powered customer service, to further this trend (Matecaș et al., 2025). Generative models can also create synthetic social media profiles with realistic photos and personal information, which are then used to befriend targets and build trust over time. The technological asymmetry is clear: average users assume they are communicating with a real person, unaware that an AI agent, guided by vast amounts of available data and using algorithms to orchestrate the conversation, is behind the interaction. Some studies show that generative AI effectively removes many of the telltale signs of fraud that users once relied on (Alahmed et al., 2024). Furthermore, attacks are no longer limited to email phishing; phone calls, text messages, and social media messages generated by artificial intelligence can also be used to socially engineer victims (Alahmed et al., 2024).

These innovations have significantly lowered the entry barrier to conducting sophisticated social engineering. Previously, only well-equipped actors – such as state intelligence agencies or organised crime rings – could execute complex deception operations, often investing considerable time in devising excuses, rehearsing fake scenarios, or coordinating teams (Stoica, 2021). Today, however, openly available AI tools enable even modestly skilled attackers to generate numerous customised attack templates within seconds, effectively automating the creative workload (Matecaș et al., 2025). From an offensive cyber operations perspective, AI compresses reconnaissance and tailoring work, enabling rapid production of highly personalised social engineering content and scaling campaigns far beyond what manual social engineering could sustain (Collier, 2025). This democratisation of capability erodes the asymmetry that once favoured only the most skilled or resource-rich individuals or groups. A recent study showed that, with minimal guidance, readily available large language models can produce credible spear-phishing emails and vishing (voice phishing) call scripts, often indistinguishable from those written by experts (Matecaș et al., 2025).

In summary, AI and automation have escalated human-targeted deception by increasing the realism, variety, and interactivity of attacks, thereby raising the likelihood that targets will be deceived. Furthermore, AI-driven human deception has multiplied both the quantity of attacks (through automation and scale) and their quality (through sophistication and personalisation), redefining what constitutes realistic deception in the digital age. In addition to targeting human perception, attackers are also engaging in technical deception – the manipulation of digital and physical systems to deceive victims or evade security measures (Stoica, 2021). One aspect is the creation of fraudulent infrastructure, such as counterfeit websites, emails, or online assets that falsely present themselves as legitimate. By mimicking trusted portals or communication channels, adversaries can trick users into unknowingly interacting with malicious systems. A flood

of fraudulent emails and domains can be automatically generated and rotated, confusing traditional blacklists and reputation-based defences (Schmitt and Flechais, 2024).

Another aspect of technical deception is deceiving machines that protect or inform human decision-making. In the social engineering context, this could involve crafting phishing emails or chat messages specifically optimised to fool spam filters and AI-based detectors. Rather than focusing solely on deceiving humans, the attacker also targets defensive AI. For example, by using generative models to test and improve phishing content, attackers can identify phrases or formats that consistently bypass email classifiers, thereby neutralising automated defences (Schmitt and Flechais, 2024). This phenomenon exemplifies escalation: while organisations use AI to detect fraud, attackers respond by using AI to generate evasive attacks – a duel of algorithms underlying human-targeted deception. As a result, conventional security tools, which often rely on static rules or past attack signatures, become less effective (Alahmed et al., 2024).

Attackers are increasingly using sensor spoofing and technical manipulation of physical systems as part of blended social engineering attacks. In some scenarios, a "social engineer" may target an automated system trusted by a human, rather than directly exploiting human psychology (Kumar et al., 2025). For example, by using deepfake audio trained to mimic the victim's voice, attackers can deceive automated speaker recognition systems into believing the victim has authenticated. Once the system is deceived, individuals can be socially engineered through an intermediary – they trust the machine's validation and, consequently, the caller. Similarly, in the fields of autonomous vehicles and the Internet of Things, recent studies have shown that carefully crafted signals can spoof sensors (Xu et al., 2023). In one study, attackers projected fake traffic and environmental signs, causing autonomous car sensors to perceive non-existent obstacles or to miss real ones, potentially leading to dangerous behaviour (Xu et al., 2023). While not a classic social engineering attack on a person, such technical deceptions create false realities to which engineers and users may respond inappropriately. They illustrate the broader concept of infrastructure deception – where any component of the digital ecosystem (from AI systems to network protocols) can be manipulated to support an attack. These developments demonstrate how modern social engineering operates on multiple fronts: by deceiving the human mind with convincing content and by manipulating the technical environment to reinforce the illusion. This multi-layered approach marks a clear escalation from earlier eras, requiring equally sophisticated awareness to counter.

A common theme underlies these trends: digital automation and AI have transformed the feasibility, scope, and sophistication of social engineering. Attackers now routinely use big data and machine learning to increase the success of their attacks. Artificial intelligence can analyse vast

amounts of personal data, social media posts, and other digital footprints to profile potential victims in detail (Blauth et al., 2022). Automated reconnaissance, once a labour-intensive process, can now be performed rapidly by algorithms that identify those most vulnerable in a given system (f, new employees or individuals with publicly available contact information and specific interests). Blauth et al. (2022) note that AI-driven target selection reduces the time and effort required for attackers to find vulnerable individuals, making AI-driven social engineering an effective and inexpensive threat. With AI handling the bulk of the work, even mass-personalised attacks become feasible. An attacker can craft 1,000 individually tailored phishing messages as easily as 10 generic ones, undermining the economic constraints that once limited the reach of fraud. Furthermore, the issue is not only the scale and volume of attacks, but also their variation: AI-powered models can generate an almost infinite series of unique message phrases, avoiding repetition that could trigger detection. Such polymorphic social engineering ensures that each victim receives a slightly different bait, bypassing universal warning signs (Pakina et al., 2025).

In the cybercrime literature, this phenomenon has been identified as a shift towards "attack as a service," where threat actors use AI platforms to carry out much of the work involved in committing cybercrime (Broadhurst et al., 2019; Hayward and Maas, 2020). Furthermore, Blauth et al. (2022) argue that advanced AI tools have blurred the distinction between human perpetrator and tool, as AI can autonomously generate and execute social engineering attacks with minimal human input. The question arising from this observation is whether AI itself can be considered a "motivated perpetrator" in criminological models or merely an extension of user intent.

It is important to note that technological asymmetry can have reciprocal consequences in this escalation. On one hand, attackers from this cybercrime group continue to enjoy disproportionate advantages such as access to zero-day exploits, vast AI computing resources, or specialised knowledge, which they integrate with social engineering to devastating effect. These actors coordinate campaigns that combine hacking and psychological warfare, as seen in state-led disinformation and phishing operations (Stoica, 2021; Collier, 2025). On the other hand, the proliferation of AI tools has empowered attackers with fewer resources, narrowing the gap between high- and low-level threat actors ihyper-ren terms of social engineering capabilities (Matecaș et al., 2025). The ability to synthesise human deception on a large scale is no longer exclusive to intelligence agencies; it is now available through APIs and open-source models to anyone with malicious intent.

From the perspective of those involved in preventing such attacks, the contemporary landscape of social engineering is defined by both the sheer volume of attacks and their high sophistication. Every email, call, or

message can be supported by AI-selected personal information, written in grammatically correct text, potentially deepfaked, and delivered through a fraudulent channel. Psychological tactics – such as inducing trust, fear, or urgency – remain rooted in traditional principles of manipulation, but their execution is now enhanced by 21st-century automation and AI.

In summary, technological asymmetry and escalation have fundamentally transformed social engineering. Human-centric social engineering has reached new heights with AI, deepfakes, and chatbots that exploit human trust on a massive scale and with remarkable realism. Technical deception – from fake websites and infrastructure to hostile exploitation of AI and sensors – adds another dimension, undermining the tools and channels we rely on to distinguish truth from falsehood. Digital infrastructures, especially artificial intelligence and automated systems, have proven to be doubly effective: they empower defenders, but also greatly enhance attackers' ability to manipulate and adapt.

The current trend in social engineering is therefore characterised by an accelerated arms race in which advances in generative artificial intelligence and automation continue to outpace deception, at least until equally advanced countermeasures emerge. The emerging evolution of the AI-enabled social engineering threats can be classified into '3E phases', i.e. i) Enlarging, the growth of attacks through digital media, ii) Enriching, adding new threat vectors and approaches, and iii) Emerging new threats and methods (Yu et all., 2024). While a detailed discussion of mitigation is beyond the scope of this chapter, recognising the nature of the threat is a crucial first step. Technological asymmetry means that attackers will seek any advantage that innovation provides; escalation means those advantages will be ruthlessly exploited and disseminated. Recognising these realities is essential for both academia and industry as they strive to safeguard the human element in an era of intelligent, automated deception.

Prevention and Intervention: Educational, Organisational, and Socio-technical

Effective prevention of social engineering requires a layered sociotechnical approach that combines people-focused measures with supporting technical controls (Anderson et al., 2016). In practice, this involves raising users' awareness of potential dangers in decision-making, making it part of the defence mechanism within organisational processes and systems. Cletus et al. (2022) compare this to building a "human firewall", that is, a workforce trained to recognise and combat manipulation. However, research (Li, Song and Pang, 2023; Cletus et al., 2022) consistently shows that general, one-off awareness training has limited impact. If training is too focused on compliance with legal norms or the content is too abstract in relation to the real work context, employees tend to ignore it, and the "human firewall" can fail in practice (Cletus et al., 2022). In educational

institutions such as universities, which have open communication norms and diverse users, simplistic ad hoc warning messages, annual seminars, or thematic workshops are particularly insufficient. A more comprehensive strategy is needed, focusing on continuous education, a supportive culture, and usable security design (Li, Song and Pang, 2023).

By reviewing the literature on social engineering and its associated problems, guidelines can be identified that serve as a basis for creating prevention models and strategies to reduce the likelihood of successful social engineering attacks.

One important guideline concern adapting education and programmes to the actual roles individuals have in their organisations, using real or highly realistic scenarios. Security awareness programmes are most credible when they reflect the actual working conditions and decision points employees face. Instead of universal content, education and training should be based on the specifics of individual roles – for example, staff in the financial sector should learn to recognise engineering attempts related to accounts or financial operations, while IT staff should be trained to verify the identity of callers. Aldawood and Skinner (2019) note that many organisations fail to adapt their training to different staff groups, which limits effectiveness, while adapting content to the job context improves relevance and engagement (Alsulami et al., 2021). From the technical aspect, this approach must include customized roles or jobs specific training platforms and realistic simulated phishing exercises.

The second principle underpinning training is to make it interactive and experiential. Simulated phishing exercises and dashboards, just-in-time prompts, scenario-based workshops, and guided reflections on past incidents are much more effective in helping employees build practical detection skills. Cletus et al. (2022) found that programmes relying solely on PowerPoint presentations or quizzes often fail to achieve lasting impact, while those incorporating active participation, such as role-playing, are more effective in promoting safer behaviours in digital environments. Similarly, Syafitri et al. (2022) emphasise that experiential learning components can bridge the gap between knowledge and action in preventing social engineering. For example, live-fire exercises (phishing simulations with immediate feedback) allow users to learn from mistakes in a safe environment. Additionally, interactive e-learning platforms, sandboxing, and collaboration and engagement tools can play vital roles in such trainings.

The third principle emphasises reinforcement and regular follow-up of training to maintain and refine learned skills. One-off training sessions have only short-term benefits if not reinforced. Numerous studies (Cletus et al., 2022; Syafitri et al., 2022; Aldawood and Skinner; Yu et al., 2024) encourage a shift from annual training to ongoing awareness and education programmes. This can include periodic refresher modules, monthly microlearning sessions, manager-led follow-ups or reinforcements, and surprise

simulation exercises to keep users engaged and better prepared for challenges. In their research, Aldawood and Skinner (2019) highlighted that attackers are constantly updating their tactics, so training for staff responsible for prevention must also be updated and repeated to keep pace. Cletus et al. (2022) propose a behavioural change framework called ATAC (Awareness → Transition → Adaptation → Consolidation) to highlight the decline after initial training. In their model, initial awareness must be transformed into adaptive responses through practical work and active training and ultimately consolidated into long-term habits through continuous reinforcement. The key message of this model is that habits are formed over time – a single training session or seminar, no matter how good or how expert the facilitators, is unlikely to permanently change behaviour without accompanying practical training. Reinforcement prompts can enable context-aware automated security reminders at important and needful moments. Moreover, micro-learning apps and analytics dashboards may also be useful for this approach.

The fourth principle, which is otherwise ever-present when addressing issues related to theory and practical work, is bridging the gap between knowledge and action. Simply knowing about social engineering threats does not ensure that a person will respond correctly and promptly when under pressure or otherwise hindered in their work. Some studies have confirmed this gap, noting that many users who can identify phishing still click on suspicious links if they do not receive practice and support. For example, Alsulami et al. (2021) found that even participants with sufficient knowledge of phishing concepts often failed to apply that knowledge in real-world scenarios. The implication of these findings is that awareness must be translated into action through well-designed interventions.

However, technical defences alone will fail if the organisation's culture and policies do not reinforce safe practices, which is the fifth principle on which prevention should be based. Borkovich and Skovira's (2019) research introduce the concept of "cybersecurity inertia", where organisations recognise threats such as social engineering but delay action due to internal resistance or complacency.

Common issues also include fragmented responsibility ("it is IT's job, not mine"), budget constraints, and a tendency to blame individual employees for breaches rather than addressing systemic weaknesses. Overcoming this inertia requires visible leadership commitment and a culture that treats security as everyone's responsibility. As Borkovich and Skovira (2019) note, elevating cybersecurity to a "top-down, organisation-wide priority" ensures that training and policies for system security are taken seriously at all levels.

The sixth principle underlying the development of prevention strategies is to foster an environment that focuses on learning from mistakes rather than assigning blame for failures. A key aspect of the organisational

context is how errors and incidents are managed. An approach is to create 'Non-Punitive Culture' in which individuals who have fallen victim to social engineering are punished without addressing the circumstances that led to the incident can deter others from reporting similar events (Klimburg-Witjes and Wentland, 2021). A more effective strategy is to share responsibility – acknowledging that people are imperfect and designing systems that support them. One way to achieve this is to implement a policy of anonymous reporting of incidents or "near misses" (where employees can report a suspicious attempt without fear of punishment), coupled with direct training. In short, people should feel safe admitting mistakes or uncertainties, enabling the organisation to continuously improve its defences. Organizations may also introduce a "Security Champions" program where the individuals who report phishing incidents can earn rewards or get recognized for their efforts, and they help encourage everyone to stay alert and active (Jayatilaka, 2021). The most common and quick channels for the incidents reporting can be dedicated email addresses, web-forms, browser extensions, and mobile apps.

The foundations of the sixth principle also underpin the seventh principle, which is to embed a secure environment into daily operational routines and accountability structures. To sustain efforts to prevent damage from social engineering, organisations can integrate security checkpoints into standard workflows, such as additional protocols for larger transactions or password generation for email accounts. For example, integrating verifications workflows, incident reporting mechanisms, and interactive analytics dashboards among others are highly important. Borkovich and Skovira (2019) also suggest measures such as ethical hacking simulations to detect vulnerabilities in systems at regular intervals, ensuring the overarching goal of embedding security awareness into the organisational structure – from training for new employees, through performance appraisals, to internal communications (sharing tips, celebrating "security heroes", etc.). When security is reinforced by organisational structures and culture, employees are more likely to remember and apply what they have learned in training.

However, care must be taken not to overdo security instructions and activities. It should be remembered that even well-trained, conscientious users can make mistakes, especially under time pressure or cognitive load. Usability research in security shows that if warnings and controls are too frequent, confusing, or obstructive, users will eventually ignore them – a phenomenon known as security warning fatigue or habituation (Anderson et al., 2016).

Zaaba et al. (2021) investigated the design of warnings and instructions about the dangers of falling victim to social engineering in this context. Their research showed that the design and clarity of instructions related to social engineering risks are important factors for more effective defence. Security warnings should be concise, actionable, and context spe-

cific. The same authors emphasise that, instead of lengthy, generic messages, instructions that briefly state the risk and guide the user on what action to take (such as whom to report to and what type of threat is present) are more effective. Additionally, the filtering system should be adjusted, as warnings that appear too frequently or for messages flagged solely because they come from a particular email address can have the opposite effect: users may begin to ignore these warnings, increasing the prospect of becoming victims of social engineering. Li et al. (2023) describe a pattern-based framework that includes such countermeasures for each phase of a social engineering attack. In essence, technical and procedural safeguards provide support.

In conclusion, based on previous research and proposals, the most effective defence against social engineering is a combination of all measures – a multi-component prevention strategy (Zaaba et al., 2021; Li et al. 2023; Syafitri et al., 2022). Contemporary prevention research advocates multi-component programmes that combine education, process improvements, technology, and continuous evaluation. No single measure is infallible, but together they create defence in depth.

By applying such a multi-layered approach – education, organisational strengthening, and enabling technology – the risk of social engineering can be significantly reduced. It is crucial that each layer complements the others: informed and vigilant users, supported by a strong security culture and effective technical safeguards, create an environment in which attackers do not find easy targets. This integrated strategy goes beyond viewing people as the weakest link and instead empowers them as a key line of defence, supported by their institution's policies and infrastructure.

DISCUSSION

Taken as a whole, the literature supports a relatively stable conclusion that social engineering remains difficult to prevent because it exploits ordinary social and cognitive processes, while modern communication infrastructures amplify its reach, speed, and adaptability. As a result, prevention strategies that rely primarily on individual vigilance are substantially unstable. This does not mean that user behaviour is beside the point; rather, prevention effectiveness depends on how individual capability-building is integrated into organisational routines and technical safeguards (Aldawood and Skinner, 2019; Li et al., 2023; Syafitri et al., 2022).

A recurring contribution of recent conceptual work is greater clarity about what constitutes social engineering and why definitions matter for prevention evaluation. When social engineering is defined too broadly, prevention claims become difficult to interpret because interventions are measured against heterogeneous phenomena (Wang et al., 2020). In opposition, when prevention discourse over-narrows the phenomenon to “care-

less clicking", it tends to ignore social and organisational mechanisms that shape exposure (Klimburg-Witjes and Wentland, 2021). The review therefore supports using definitional anchors that distinguish social engineering by its primary mechanism – manipulation and compliance-seeking – while remaining attentive to the socio-technical environments that make those mechanisms effective (Wang et al., 2020; Zaoui et al., 2024).

The evidence base also indicates that vulnerability is best understood as contextual and patterned, rather than as a single trait or deficit. In social networking settings, susceptibility and victimisation are linked to combinations of factors spanning habits, socio-emotional dynamics, cognition, message features, and sender cues, with effects varying by sample and platform context (Alshammari et al., 2025; Albladi and Weir, 2018). Message-level studies similarly suggest that persuasive effectiveness depends on how requests are framed and which cues are emphasised, which is relevant for designing both educational and platform interventions (Algarni, 2019). Trait-based explanations appear less useful as standalone prevention tools; even when personality correlates are identified, they do not justify operational "profiling" approaches and risk reinforcing unproductive "weak link" narratives (Muhanad et al., 2025; Klimburg-Witjes and Wentland, 2021).

Regarding education and training, the literature is consistent on two points. First, training often improves knowledge and reported confidence. Second, the translation of knowledge into durable practice is uneven and contingent on context, reinforcement, and integration with organisational processes (Aldawood and Skinner, 2019; Prümmer et al., 2024; Prümmer et al., 2025). This aligns with organisational perspectives that highlight compliance constraints and inertia: users may understand risks yet continue unsafe routines when secure actions are cumbersome or socially costly (Beautement et al., 2008; Borkovich and Skovira, 2019). Usable-security research provides an additional constraint: repeated warnings can lead to habituation and reduced attention, undermining prevention strategies that assume sustained vigilance in high-friction environments (Anderson et al., 2016; Kirwan et al., 2020; Zaaba et al., 2021). These bodies of work converge on a prevention implication that is especially relevant for institutional and legal audiences: education is necessary, but it should be treated as just one component of an engineered multi layered prevention system.

Rapid technological change intensifies these challenges. Recent research on AI-enabled deception and deepfake-driven social engineering indicates a shift in the cost structure of persuasion: content can now be produced at scale, adapted rapidly, and made more plausible, reducing the diagnostic value of surface-level cues (Matecas et al., 2025; Pedersen et al., 2025; Schmitt and Flechais, 2024). In practical terms, prevention approaches that rely on users detecting poor language, generic templates, or implausible identities are likely to become less effective. This strengthens

the case for verification infrastructures (authenticated channels, escalation norms, provenance signals) and for organisational workflows that enable verification under time pressure (Li et al., 2023; Pedersen et al., 2025). As the normal working culture of an organization is dynamic and based on trust, informal norms, unwritten rules, power dynamics, different communication styles and behavioural pattens, it is highly likely that the AI models cannot fully understand the actual context and situation (Rahwan, 2019) . It may be because that data training of the AI-based models is made based on the past events or incidents and does not fully consider or sense the current organisational culture or norms.

Technological platforms such as AI-enabled threat detection and behaviour analytics tools can be highly useful in detecting the anomalies and potential malicious entities, but they can also produce higher false positive or negative rates (FPR/FNR) (Tiwari, 2025). This situation may lead to the 'alert fatigue' where the security team receives massive volume of the alerts from AI/ML tools that significantly hampers detection capabilities and corresponding responses (Tariq et al., 2025). Moreover, with increased use of AI/ML in the prevention approaches, there is limited transparency and explainability in AI-based decision-making that can cause various challenges, e.g., organisations cannot validate the system decisions and compliance requirements may not be properly fulfilled.

Two evidence gaps are relevant for a cautious, objective interpretation. First, many studies rely on self-report measures, short follow-up periods, and non-comparable outcomes, which limits inference about sustained prevention effects (Prümmer et al., 2024; Prümmer et al., 2025). Second, much of the literature is context-specific (sector, country, platform), and generalisability is not always tested; the field would benefit from more cross-context replication and multi-method designs that combine behavioural measures with organisational or process analysis (Syafitri et al., 2022; Alshammari et al., 2025). For prevention policy, this suggests that institutions should implement interventions with built-in evaluation (behavioural indicators, reporting rates, response times, simulation outcomes), rather than assuming training is effective by default (Aldawood and Skinner, 2019; Syafitri et al., 2022).

CONCLUSION

This review examined social engineering in relation to social, educational, and technological challenges in crime prevention. The literature consistently describes social engineering as a socio-technical form of offending: manipulation and compliance-seeking are central mechanisms, while digital infrastructures enable scale, repetition, and rapid adaptation (Wang et al., 2020; Mitnick and Simon, 2002). Evidence from user-, message-, and context-level studies indicates that vulnerability is shaped by

patterned combinations of habitual behaviour, socio-emotional dynamics, cognitive processing, and platform cues, with effects varying across settings (Alshammari et al., 2025; Albladi and Weir, 2018; Algarni, 2019). Training and awareness programmes contribute to prevention, but their effects are contingent and often limited when they are generic, weakly reinforced, or poorly integrated into organisational routines and usable security design (Aldawood and Skinner, 2019; Prümmer et al., 2025; Zaaba et al., 2021).

The review further suggests that accelerating technological change – particularly AI-enabled deception and deepfake-based impersonation – reduces the reliability of superficial detection cues and increases the need for verification infrastructures and workflow-based controls (Matecas et al., 2025; Pedersen et al., 2025; Schmitt and Flechais, 2024). Consequently, prevention strategies that rely on perfect individual vigilance are unlikely to remain adequate. A more robust prevention approach is layered and institutional: role-sensitive capability-building, supportive reporting and verification norms, interface and platform interventions that reduce cognitive burden, and governance arrangements that allocate responsibility appropriately and enable continuous evaluation (Li et al., 2023; Syafitri et al., 2022; Klimburg-Witjes and Wentland, 2021).

Future research would benefit from stronger behavioural measurement, longer follow-up periods, cross-context replication, and closer integration between empirical susceptibility research and organisational or policy analysis. For practitioners and policymakers, the practical implication is clear: effective cybercrime prevention for social engineering requires coordinated action across education, organisational design, technology, and governance, rather than isolated interventions targeted at individual users.

REFERENCES

Abdulla, R. M., Faraj, H. A., Abdullah, C. O., Amin, A. H., and Rashid, T. A. (2023). *Analysis of social engineering awareness among students and lecturers*. IEEE Access, 11, 101098–101111. https://doi.org/10.1109/ACCESS.2023.3311708

Akhawe, D., and Felt, A. P. (2013). *Alice in Warningland: A large-scale field study of browser security warning effectiveness*. In Proceedings of the 22nd USENIX Security Symposium (USENIX Security 13) (pp. 257–272). USENIX Association.

Alahmed, Y., Abadla, R., and Al Ansari, M. J. (2024). *Exploring the potential implications of AI-generated content in social engineering attacks*. In Proceedings of the 2024 International Conference on Multimedia Computing, Networking and Applications (MCNA) (pp. 64–73). IEEE.

Albladi, S. M., and Weir, G. R. S. (2018). User characteristics that influence judgment of social engineering attacks in social networks. Human-centric Computing and Information Sciences, 8, Article 5. https://doi.org/10.1186/s13673-018-0128-7

Aldawood, H., and Skinner, G. (2019). Reviewing cyber security social engineering training and awareness programs—Pitfalls and ongoing issues. *Future Internet, 11*(3), 73. https://doi.org/10.3390/fi11030073

Algarni, A. (2019). What message characteristics make social engineering successful on Facebook: The role of central route, peripheral route, and perceived risk. *Information, 10*(6), Article 211. https://doi.org/10.3390/info10060211

Almaliki, M. (2019). *Misinformation-aware social media: A software engineering perspective*. *IEEE Access, 7*, 182451–182459. https://doi.org/10.1109/ACCESS.2019.2960270

Alshammari, S. S., Soh, B., and Li, A. (2025). Understanding social engineering victimisation on social networking sites: A comprehensive review of factors influencing user susceptibility to cyber-attacks. *Information, 16*(2), 153. https://doi.org/10.3390/info16020153

Alshammari, S. S., Soh, B., and Li, A. (2025). *Understanding social engineering victimisation on social networking sites: A comprehensive review of factors influencing user susceptibility to cyber-attacks*. *Information, 16*(2), 153. https://doi.org/10.3390/info16020153

Alsulami, M. H., Alharbi, F., Almutairi, H., Almutairi, B. A. A., Alotaibi, M. B., Alanzi, M. E., Alotaibi, K. D., and Al-Harthi, S. E. (2021). Measuring awareness of social engineering in the educational sector in the Kingdom of Saudi Arabia. Information, 12(5), 208. https://doi.org/10.3390/info12050208

Anderson, B. B., Jenkins, J. L., Vance, A., Kirwan, C. B., and Eargle, D. (2016). Your memory is working against you: How eye tracking and memory explain habituation to security warnings. *Decision Support Systems, 92*, 3–13. https://doi.org/10.1016/j.dss.2016.09.010

Bada, M., Sasse, A. M., and Nurse, J. R. C. (2019). Cyber security awareness campaigns: Why do they fail to change behaviour? https://arxiv.org/abs/1901.02672

Beautement, A., Sasse, M. A., and Wonham, M. (2008). The compliance budget: Managing security behaviour in organisations. In *Proceedings of the 2008 New Security Paradigms Workshop (NSPW '08)* (pp. 47–58). Association for Computing Machinery. https://doi.org/10.1145/1595676.1595684

Birthriya, S. K., Ahlawat, P., and Jain, A. K. (2025). *A comprehensive survey of social engineering attacks: Taxonomy of attacks, prevention, and mitigation strategies*. *Journal of Applied Security Research*. https://doi.org/10.1080/19361610.2024.2372986

Blauth, T. F., Gstrein, O. J., and Zwitter, A. (2022). *Artificial intelligence crime: An overview of malicious use and abuse of AI*. IEEE Access, 10, 77110–77122. https://doi.org/10.1109/ACCESS.2022.3191790

Borkovich, D. J., and Skovira, R. J. (2019). Cybersecurity inertia and social engineering: Who's worse, employees or hackers? Issues in Information Systems, 20(3), 139–150. https://doi.org/10.48009/3_iis_2019_139-150

Broadhurst, R., Broadhurst, M., Alazab, M., Chon, S., and Jiang, C. (2019). *Artificial intelligence and crime*. SSRN.

Cialdini, R. B. (2007). *Influence: The psychology of persuasion* (Rev. ed.). Harper Business.

Cletus, A., Weyory, B., and Opoku, A. (2022). Improving social engineering awareness, training and education (SEATE) using a behavioral change model. *International Journal of Advanced Computer Science and Applications, 13*(5), 606–613.

Cohen, L. E., and Felson, M. (1979). Social change and crime rate trends: A routine activity approach. *American Sociological Review, 44*(4), 588–608. https://doi.org/10.2307/2094589

Collier, H. (2025). *AI in social engineering: The next generation of offensive cyber operations*. In *Proceedings of the 24th European Conference on Cyber Warfare and Security (ECCWS 2025)* (pp. 80–83). Academic Conferences International

LimitedCross, C. (2018). *Cybercrime: Victimisation, offenders and policing*. Routledge. https://doi.org/10.4324/9781315644574

Ferrari, R. (2015). Writing narrative style literature reviews. *Medical Writing, 24*(4), 230–235. https://doi.org/10.1179/2047480615Z.000000000329

Grant, M. J., and Booth, A. (2009). A typology of reviews: An analysis of 14 review types and associated methodologies. *Health Information and Libraries Journal, 26*(2), 91–108. https://doi.org/10.1111/j.1471-1842.2009.00848.x

Green, B. N., Johnson, C. D., and Adams, A. (2006). Writing narrative literature reviews for peer-reviewed journals: Secrets of the trade. *Journal of Chiropractic Medicine, 5*(3), 101–117. https://doi.org/10.1016/S0899-3467(07)60142-6

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

Hayward, K. J., and Maas, M. (2021). *Artificial intelligence and crime: A primer for criminologists*. *Crime, Media, Culture, 17*(2), 209–233.

Hindelang, M. J., Gottfredson, M. R., and Garofalo, J. (1978). *Victims of personal crime: An empirical foundation for a theory of personal victimization*. Ballinger.

Jayatilaka, A., Beu, N., Baetu, I., Zahedi, M., Babar, M. A., Hartley, L., and Lewinsmith, W. (2021). *Evaluation of security training and awareness programs: Review of current practices and guidelines.* https://doi.org/10.48550/arXiv.2112.06356

Khan, N. F., Ikram, N., Murtaza, H., and Javed, M. (2023). Evaluating protection motivation based cybersecurity awareness training on Kirkpatrick's model. *Computers and Security, 125*, 103049. https://doi.org/10.1016/j.cose.2022.103049

Kirwan, C. B., Bjornn, D. K., Anderson, B. B., Vance, A., Eargle, D., and Jenkins, J. L. (2020). Repetition of computer security warnings results in differential repetition suppression effects as revealed with functional MRI. *Frontiers in Psychology, 11*, 528079. https://doi.org/10.3389/fpsyg.2020.528079

Klimburg-Witjes, N., and Wentland, A. (2021). Hacking humans? Social engineering and the construction of the "deficient user" in cybersecurity discourses. *Science, Technology, and Human Values, 46*(6), 1316–1339. https://doi.org/10.1177/0162243921992844

Kumar, N., and Muhammad Salman. (2025). RTBTS: A Real-Time Behavioural Training System to Mitigate Psychological Vulnerabilities in Social Engineering Attacks. *International Journal of Electrical, Computer, and Biomedical Engineering*, *3*(1), 188–211. https://doi.org/10.62146/ijecbe.v3i1.103

Leukfeldt, E. R. (2014). Cybercrime and social ties. *Trends in Organized Crime, 17*(4), 231–249. https://doi.org/10.1007/s12117-014-9223-2

Leukfeldt, E. R., and Yar, M. (2016). Applying routine activity theory to cybercrime: A theoretical and empirical analysis. *Deviant Behavior, 37*(3), 263–280. https://doi.org/10.1080/01639625.2015.1012409

Li, T., Song, C., and Pang, Q. (2023). Defending against social engineering attacks: A security pattern-based analysis framework. IET Information Security, 17(4), 703–726. https://doi.org/10.1049/ise2.12125

Matecas, A.-R., Kieseberg, P., and Tjoa, S. (2025). *Social engineering with AI*. Future Internet, 17(11), 515. https://doi.org/10.3390/fi17110515

Mitnick, K. D., and Simon, W. L. (2002). *The art of deception: Controlling the human element of security*. Wiley.

Montañez, R., Golob, E., and Xu, S. (2020). *Human cognition through the lens of social engineering cyberattacks*. Frontiers in Psychology, 11, 1755. https://doi.org/10.3389/fpsyg.2020.01755

Muhanad, A., Abuelezz, I., Haris, R., Khan, K. M., and Ali, R. (2025). Personality traits as predictors of vulnerability to persuasion in social engineering amongst risk-

aware targets. Computing, 107, Article 168. https://doi.org/10.1007/s00607-025-01521-z

Nowakowski, W. (2025). Social engineering analysis framework: A comprehensive playbook for human hacking. IEEE Access, 13, 18827–18849. https://doi.org/10.1109/ACCESS.2025.3532999

Pakina, R., Kejriwal, D., Pujari, R., and Rane, A. (2025). *Adversarial AI in social engineering attacks*. *International Journal of Science and Technology, 4*(1). https://doi.org/10.56127/ijst.v4i1.1964

Parsons, K., Calic, D., Pattinson, M., Butavicius, M., McCormac, A., and Zwaans, T. (2017). The Human Aspects of Information Security Questionnaire (HAIS-Q): Two further validation studies. *Computers and Security, 66*, 40–51. https://doi.org/10.1016/j.cose.2017.01.004

Pedersen, K. T., Pepke, L., Stærmose, T., Papaioannou, M., Choudhary, G., and Dragoni, N. (2025). Deepfake-driven social engineering: Threats, detection techniques, and defensive strategies in corporate environments. *Journal of Cybersecurity and Privacy, 5*(2), 18. https://doi.org/10.3390/jcp5020018

Prümmer, J., van Steen, T., and van den Berg, B. (2024). A systematic review of current cybersecurity training methods. *Computers and Security, 136*, 103585. https://doi.org/10.1016/j.cose.2023.103585

Prümmer, J., van Steen, T., and van den Berg, B. (2025). Assessing the effect of cybersecurity training on end-users: A meta-analysis. *Computers and Security, 150*, 104206. https://doi.org/10.1016/j.cose.2024.104206

Rahwan, I., Cebrian, M., Obradovich, N., Bongard, J., Bonnefon, J.-F., Breazeal, C., … Wellman, M. (2019). *Machine behaviour*. *Nature, 568*, 477–486. https://doi.org/10.1038/s41586-019-1138-y

Schmitt, M., and 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

Siddiqi, M. A., Pak, W., and Siddiqi, M. A. (2022). A study on the psychology of social engineering-based cyberattacks and existing countermeasures. Applied Sciences, 12(12), 6042. https://doi.org/10.3390/app12126042

Stoica, A. (2021). *Social engineering as the new deception game*. *Romanian Journal of Information Technology and Automatic Control, 31*(3), 57–68. https://doi.org/10.33436/v31i3y202105

Syafitri, W., Shukur, Z., Mokhtar, U. A., Sulaiman, R., and Ibrahim, M. A. (2022). Social engineering attacks prevention: A systematic literature review. IEEE Access, 10, 39325–39343. https://doi.org/10.1109/ACCESS.2022.3162594 OUCI

Tariq, S., Singh, A., Chhetri, S. R., Nepal, S., and Paris, C. (2025). *Bridging expertise gaps: The role of LLMs in human-AI collaboration for cybersecurity.*

Tiwari, S. (2025). *Social engineering attacks: Trends, psychological triggers, and defense mechanisms* (preprint). Preprints.org.

Tversky, A., and Kahneman, D. (1974). Judgment under uncertainty: Heuristics and biases. *Science, 185*(4157), 1124–1131. https://doi.org/10.1126/science.185.4157.1124

Venkatesha, S., Reddy, K. R., and Chandavarkar, B. R. (2021). Social engineering attacks during the COVID-19 pandemic. SN Computer Science, 2, 78. https://doi.org/10.1007/s42979-020-00443-1

Vishwanath, A. (2015). *Examining the distinct antecedents of e-mail habits and its influence on the outcomes of a phishing attack*. Journal of Computer-Mediated Communication, 20(5), 570–584. https://doi.org/10.1111/jcc4.12126

Vishwanath, A., Harrison, B., and Ng, Y. J. (2018). Suspicion, cognition, and automaticity model of phishing susceptibility. *Communication Research, 45*(8), 1146–1166. https://doi.org/10.1177/0093650215627483

Walklate, S. (2007). *Imagining the victim of crime*. Open University Press.
Wang, Z., Sun, L., and Zhu, H. (2020). *Defining social engineering in cybersecurity*. IEEE Access, 8, 85094–85115. https://doi.org/10.1109/ACCESS.2020.2992807
Wang, Z., Zhu, H., and Sun, L. (2021). Social engineering in cybersecurity: Effect mechanisms, human vulnerabilities and attack methods. IEEE Access, 9, 11895–11910. https://doi.org/10.1109/ACCESS.2021.3051633
Wang, Z., Zhu, H., Liu, P., and Sun, L. (2021). *Social engineering in cybersecurity: a domain ontology and knowledge graph application examples*. *Cybersecurity, 4*(1), 31. https://doi.org/10.1186/s42400-021-00094-6
Workman, M. (2008). Wisecrackers: A theory-grounded investigation of phishing and pretext social engineering threats to information security. *Journal of the American Society for Information Science and Technology, 59*(4), 662–674. https://doi.org/10.1002/asi.20779
Xu, Y., Han, X., Deng, G., Li, J., Liu, Y., and Zhang, T. (2023). *SoK: Rethinking sensor spoofing attacks against robotic vehicles from a systematic view*. In 2023 IEEE 8th European Symposium on Security and Privacy (EuroSandP) Workshops (pp. 1082–1100). IEEE.
Yar, M. (2005). The novelty of "cybercrime": An assessment in light of routine activity theory. *European Journal of Criminology, 2*(4), 407–427. https://doi.org/10.1177/147737080556056
Yu, J., Yu, Y., Wang, X., Lin, Y., Yang, M., Qiao, Y., and Wang, F. Y. (2024). The shadow of fraud: The emerging danger of ai-powered social engineering and its possible cure. arXiv preprint arXiv:2407.15912.
Zaaba, Z. F., Yi, C. L. X., Amran, A., and Omar, M. A. (2021). Harnessing the challenges and solutions to improve security warnings: A review. *Sensors, 21*(21), 7313. https://doi.org/10.3390/s21217313
Zaoui, M., Yousra, B., Sadqi, Y., Maleh, Y., and Ouazzane, K. (2024). *A comprehensive taxonomy of social engineering attacks and defense mechanisms: Toward effective mitigation strategies*. IEEE Access, 12, 72224–72241. https://doi.org/10.1109/ACCESS.2024.3403197

СОЦИЈАЛНИ ИНЖЕЊЕРИНГ И ОГРАНИЧЕЊА ПРЕВЕНЦИЈЕ САЈБЕРКРИМИНАЛА: ПРЕГЛЕД ДРУШТВЕНИХ, ОБРАЗОВНИХ И ТЕХНОЛОШКИХ ИЗАЗОВА

Далибор Долежал[1], Танеш Кумар[2]

[1]Универзитет у Загребу, Факултет за едукацијско-рехабилитацијске науке, Одељење за криминологију, Загреб, Хрватска
[2]Универзитет Алто, Одељење за информационо и комуникационо инжењерство, Алто, Финска

Резиме

У раду се полази од питања због чега су облици преваре засновани на манипулацији људима и даље међу најуспешнијим начинима извршења сајберкриминала, иако организације већ годинама улажу у безбедносне алате, интерне процедуре и различите обуке запослених. Уместо тумачења које успех напада пре свега приписује „непажњи корисника", рад показује да је реч о сложенијем проблему у

коме се укрштају психолошки механизми убеђивања, обрасци свакодневне комуникације и начин на који су радни задаци организовани. Тежиште није на појединачним грешкама, већ на околностима у којима се одлуке доносе брзо, под притиском и уз ослањање на уобичајене сигнале поверења (позната адреса, очекивана порука, ауторитет пошиљаоца, осећај хитности, рутина одговарања). У том оквиру социјални инжењеринг се приказује као облик криминалног деловања који користи друштвене односе и комуникационе навике исто онолико колико користи техничку инфраструктуру.

Средишњи део рада систематизује литературу из више области и прати три линије расправе. У првом делу разматрају се обрасци виктимизације и различити облици рањивости, при чему се наглашава да успех напада не зависи само од знања или „типа личности“, већ од конкретне ситуације: радне улоге, организационих очекивања, канала комуникације и емоционалног стања примаоца поруке. Посебно се истиче да исти појединац може у једном контексту препознати превару, а у другом реаговати управо онако како нападач жели, што потврђује да је процена ризика снажно условљена окружењем и тренутним задатком. У другом делу рада анализирају се едукативне и организационе мере. Показује се да тренинзи и кампање могу имати ефекта, али да ти ефекти често слабе када су садржаји општи, када се спроводе повремено и када нису повезани са стварним процедурама у установи или компанији. Такође, рад указује на ограничења стандардних упозорења у интерфејсима: корисници временом престају да их примећују, аутоматски кликћу кроз поруке система или их тумаче у складу са радним приоритетима, а не безбедносним правилима. У трећем делу обрађује се савремени технолошки развој, нарочито употреба генеративне вештачке интелигенције, која нападачима омогућава бржу припрему уверљивих порука, већи обим кампања и прецизније прилагођавање садржаја мети. Додатну тежину имају алати за имитацију гласа и слике, јер поткопавају ослањање на „препознатљив глас“, видео-позив или стил писања као знакове аутентичности.

На основу приказане грађе, рад предлаже промену фокуса у превенцији: уместо модела који одговорност пребацује готово искључиво на појединца, потребна је изградња отпорности на нивоу организационог система. То подразумева да се провере идентитета и осетљивих захтева не остављају само процени запослених, већ да буду уграђене у радне токове (нпр. обавезна потврда другим каналом, јасни протоколи за финансијске и кадровске захтеве, успоравање процеса када се тражи хитна реакција). Истовремено, важни су и дизајн комуникационих платформи, боља видљивост поузданих сигнала, као и мере које отежавају лажно представљање и злоупотребу идентитета. Завршно, рад наглашава да будућа истраживања треба да се мање ослањају на изолована тестирања „препознавања пхисхинга“, а више на праћење понашања у реалним условима, кроз дуже временске периоде и у различитим секторима, како би се прецизније проценило које комбинације организационих, едукативних и техничких мера заиста смањују успешност социјалног инжењеринга у пракси.