Borneo International Journal eISSN 2636-9826; Vol.8 (4); 2025; 47-55
Published by *Majmuah Enterprise*
www.majmuah.com

Beyond Romance and Investment: A Forward-Looking Analysis of AI-Enabled Social Engineering Attacks on WhatsApp, Telegram, and TikTok in the Post-2025 Era

MARHAKIM MOHAMAD MOKHTAR and MOHAMAD FADLI BIN ZOLKIPLI
School of Computing, College of Art and Science, Universiti Utara Malaysia (UUM), 06010 Sintok, Kedah, MALAYSIA
Email: marhakimm@gmail.com m.fadli.zolkipli@uum.edu.my| Tel: +60194903221 |+60177247779 |

Received: December 14, 2025
Accepted: December 18, 2025
Online Published: December 18, 2025

Abstract
The rapid advancement of generative artificial intelligence (AI) has reshaped the nature of social engineering attacks, extending far beyond conventional romance and investment scams. Rather than relying on static deception techniques, contemporary attacks increasingly leverage hyper-personalised content, real-time adaptation, and synthetic media to exploit trusted communication environments. Encrypted messaging platforms and short-form video applications, in particular, have become fertile ground for these evolving threats. This study presents a forward-looking comparative analysis of AI-enabled social engineering attacks across WhatsApp, Telegram, and TikTok, with a specific focus on how platform architecture influences attacker capability in the post-2025 landscape. Using a mixed-methods approach, 1,247 threat intelligence reports published between 2023 and 2025 were examined alongside three representative case studies: HackOnChat (WhatsApp session hijacking), DarkGram (Telegram's cybercriminal ecosystem), and AI-generated scam content on TikTok. The findings reveal marked differences in platform risk profiles. Telegram demonstrates the most resilient criminal infrastructure, with hundreds of channels sustaining large-scale malicious activity and high engagement rates. WhatsApp, while hosting fewer attack vectors, presents the greatest financial impact due to business-targeted attacks driven by voice cloning and contextual phishing. TikTok emerges as the fastest-growing vector, where algorithmic amplification enables AI-generated scam content to reach large audiences before moderation occurs. By synthesising these findings, this study proposes a platform-specific vulnerability framework and evidence-based mitigation strategies that address both technical and behavioural dimensions of risk. The research extends Social Cognitive Theory by incorporating the influence of synthetic media and offers practical recommendations for platform governance, organisational security training, and AI-assisted detection mechanisms necessary to navigate the evolving threat environment of 2026–2027.
Keywords: AI-enabled social engineering, WhatsApp security, Telegram cybercrime, TikTok threats

1. Introduction

1.1 Research Background

Cybersecurity landscape is currently experiencing a drastic paradigm shift result of artificial intelligence (AI) with advanced generative capabilities. Attacks relying on traditional social engineering which depend on romance scam and fake investment (Muscanell et al., 2014) now considered obsolete compared to post-2025 attack vector which combines synthetic media, real time deepfakes, and personas generated by Large Language Model (LLM) (Gupta et al., 2023; Albladi & Weir, 2023). These attacks go beyond conventional phishing which creates hyper-personalized and context-aware fraudulent content, which is statistically difficult to distinguish from authentic communication (Cocelli et al., 2024). Social media platforms—specifically WhatsApp, Telegram, and TikTok— have become a major conduit for these attacks due to their unique architectural features. WhatsApp's end-to-end encryption system and user-trusted interface paradoxically facilitates session hijacking campaigns such as HackOnChat which affected over 3,000 accounts in Asia in 2025 (CTM360, 2025). Telegram's privacy-oriented architecture and broadcast channels have inadvertently spawned a resilient cybercriminal ecosystem, with 339 main cybercrime activity channels reach 23.8 million customers (Pourabbas Vafa et al., 2024). TikTok's algorithmic content distribution system amplifies dangerous AI-generated videos, with one infostealer campaign reaching 500,000 views before being detected (Trend Micro, 2025). The convergence between these platform-specific vulnerabilities and the capabilities of generative AI demands a proactive analytical framework that goes beyond retrospective scam categorization. (Abdullah et al., 2021; Frauenstein & Flowerday, 2020).

1.2 Problem Statement

Although there is increase in threat, the existing literature points to three critical gaps. First, existing social engineering research has largely focused on email-based phishing and traditional confidence scams, with 73% of studies published

before 2022 ignoring AI-enabled multimedia attacks 2022 ignore AI-enabled multimedia attacks (Yamamoto & Yamada, 2023). Second, platform-specific analysis is still isolated; no systematic comparative studies have assessed how WhatsApp's QR authentication, Telegram's API permissions, and TikTok's recommendation algorithm facilitate new attack vectors differently. Third, Predictions of future threats are non-existent in academic discourse whereas most studies analyze historical incidents rather than predicting the evolution of attacks post-2025 (Gartner, 2024). This study addresses this gap by investigating AI-driven social engineering attacks that have surpassed romance and investment scams in sophistication and penetration.

1.3 Research Questions

This study guided by the following questions:

1. RQ1: What are the AI-enabled social engineering attack vectors beyond the romance and investment scams that are now appearing on WhatsApp, Telegram, and TikTok?
2. RQ2: How platform-specific architectural features (e.g. WhatsApp's linked device system, Telegram channel broadcasting, TikTok's algorithmic amplification) facilitate these attacks?
3. RQ3: What is the evolutionary trajectory of these AI-enabled attacks in the post-2025 era (2026–2027)?
4. RQ4: To what extent can a multi-layered mitigation strategy, comprising platform governance, user training, and technology detection, effectively address these emerging threats?

1.4 Research Objectives

Main Objective

Conduct proactive comparative analysis of AI-enabled social engineering attack vectors on WhatsApp, Telegram, and TikTok, identifying platform-specific vulnerabilities and predicting post-2025 threat trajectories that go beyond traditional romance and investment scams.

Specific Objectives

1. To systematically categorize new AI-driven attack vectors — including real-time voice cloning, deepfake video calls, LLM-generated phishing content, and synthetic influencer personification—across all three target platforms, based on threat intelligence from 2023–2025 (CBS News, 2024; Trend Micro, 2025).
2. Assess the vulnerability of the platform architecture by analyzing how WhatsApp session authentication mechanisms, Telegram API permissions, and TikTok content recommendation algorithms facilitate or hinder AI-powered social engineering campaigns (CTM360, 2025; Pourabbas Vafa et al., 2024).
3. Evaluate user vulnerability factors in the digital-native demographic (aged 18–30), measuring the effectiveness of current security awareness training against AI-generated synthetic media using Social Cognitive Theory principles (Abdullah et al., 2021; Ezeaka & Ewetuobi, 2024).
4. Predicting the evolution of attacks post-2025 by combining Gartner forecasts (2024), AI capability roadmap, and cybercriminal ecosystem migration patterns observed in Telegram channel resilience (Albladi & Weir, 2023; Gupta et al., 2023).
5. Develop a platform-specific mitigation framework which combines AI-based detection (e.g. the 96% accuracy DarkGram model), strengthened authentication protocols, and an adaptive user education program proven to reduce vulnerability scores from 24.02 to 13.97 through continuous training (Kyi & Stobert, 2022; Al-Mudahi et al., 2022).
6. Provide evidence-based policy recommendations for platform developers and cybersecurity practitioners addressing the tension between privacy preservation and security responsibilities in encrypted messaging environments (Keepnet Labs, 2025; Huntress, 2025).

1.5 Significance of the Study

This study makes three main contributions. Academically, it extends social engineering theory by integrating the threat of synthetic AI media into the Social Cognitive Framework., addressing critical gaps in cybersecurity literature (Albladi & Weir, 2023; Frauenstein & Flowerday, 2020). Practically, it offers the first universal comparative vulnerability assessment across WhatsApp, Telegram, and TikTok, providing actionable intelligence for platform security teams and corporate defenses (Pourabbas Vafa et al., 2024; Trend Micro, 2025). Socially, it predicts threats post-2025—predicting global losses of $40 billion from deepfake scams by 2027—enabling proactive policy development before these attacks reach mainstream acceptance (CBS News, 2024; Gartner, 2024).

2. Literature Review

2.1 Theoretical Framework: Social Cognitive Theory and Synthetic Media Threat

Social Cognitive Theory (SCT) serves as the conceptual basis of this study, explains the reciprocal interactions between personal, environmental, and behavioral factors in determining vulnerability to social engineering attacks (Abdullah et al., 2021). In the post-2025 era, SCT needs to be expanded to take into account the capabilities of generative AI that is

capable of manipulating all three factors simultaneously. AI not only changes the environment (e.g. fake messages that appear authentic), but also influences personal factors (e.g. reducing risk perceptions) and shapes behavior (e.g. normalizing clicks without checking) (Muscanell et al., 2014; Frauenstein & Flowerday, 2020). Albladi and Weir (2023) proposed a "Synthetic Media Threat Framework" that classifies AI attacks according to three dimensions: persona manipulation (creating a convincing false identity), temporal manipulation (attacks that change according to real-time responses), and contextual manipulation (adapting to the victim's current events). This framework is critical to understanding how real-time deepfakes don't just mimic faces, but build complete "fake environments," forcing victims to interact with nonexistent entities (Cocelli et al., 2024). Gupta et al. (2023) adding that LLMs like GPT-4 have completely changed the scale - attacks that used to require weeks of manual compilation are now completed in seconds, with personalization taking into account the victim's conversation history, demographics, and cultural cues.

2.2 Attack Evolution: From Static to AI-Dynamic

Early literature described phishing as a "static" attack—template email, fixed fake page (Abdullah et al., 2021). Study by Frauenstein and Flowerday (2020) uses personality information processing models to predict vulnerabilities, but does not take into account the ability of attacks to "learn" and adapt. Muscanell et al. (2014) identified Cialdini's five principles of influence (trust, citizenship, commitment, comfort, scarcity) in internet scams, but personalization is still limited to public data (name, location). Generative AI goes beyond this by producing dynamically responsive content. Abdullah et al. (2021) proving that habituation to authentic notifications reduces the effectiveness of warnings—AI is now optimizing this cognitive weakness through real-time adaptation. Yamamoto and Yamada (2023) in their literature review they stated that 73% of studies before 2022 ignored AI-enabled multimedia attacks, creating a critical knowledge gap. Three generations of attacks have been identified: First Generation (static), Second Generation (data-driven personalization), and now Third Generation (AI-dynamic) which combines deepfake, LLM-persona, and contextual adaptation (Albladi & Weir, 2023). This study contributes by introducing the predicted Fourth Generation—AI-autonomous that operates without human direction.

2.3 Platform Specific Vulnerabilities

2.3.1 WhatsApp: The Trust Paradox

Ezeaka and Ewetuobi (2024) found that 78% of students were aware of the risks, but 63% still clicked on links due to blind trust in the encryption icon. HackOnChat 2025 compromised 3,000+ business accounts via QR code exploitation (CTM360, 2025). Key vulnerabilities: trust exploitation and contextual AI fraud (Keepnet Labs, 2025).

2.3.2 Telegram: A Criminal Ecosystem The DarkGram dataset (Pourabbas Vafa et al., 2024) analyzed 339 channels with 23.8 million subscribers—78% positive reactions to malicious links. API allows for mass phone enumeration, channel migration takes just 14 minutes making it resilient (Huntress, 2025). 28.1% phishing links, 38% malicious exe.

2.3.3 TikTok: Algorithmic Amplification

Trend Micro (2025) documented a deepfake video reaching 500,000 views before takedown. For You Page (FYP) prioritizes engagement, not validity. Hackers use AI to optimize captions, manipulate engagement via bot farms, and hidden QRishing (Gupta et al., 2023; Cocelli et al., 2024).

2.4 AI-Sophisticated Attack Vectors: Behind Romance & Investment

Voice Cloning & Vishing 2.0: CBS News (2024) reports a 442% increase in AI vishing in 2024-2025. VALL-E technology only requires 3 seconds of recording for 99% accurate imitation (Cocelli et al., 2024). TOAD (Telephone Oriented Attack Delivery): hybrid email + AI call for fund confirmation (Gartner, 2024). Kimsuky APT uses this technique to target activists and government officials (Genians, 2024). LLM-Generated Content: Gupta et al. (2023) show GPT-4 can generate 1,000 phishing messages in 10 minutes, taking into account the victim's LinkedIn and social media. This is completely beyond manual capabilities. Real-time deepfake: DeepFaceLive and Magicam (Cocelli et al., 2024) allows hackers to change faces and voices live in video calls—a threat that leaves no forensic metadata. Session Hijacking 2.0: HackOnChat shows that AI is not just exploiting QR codes, but automating the entire attack chain—from reconnaissance, exploitation, persistence, to lateral movement within the victim's contact network. (CTM360, 2025).

2.5 User Factors & Existing Training Failures

Abdullah et al. (2021) and Kyi & Stobert (2022) proving training >3 times a year lowered the vulnerability score from 24.02 to 13.97 (Chetioui et al., 2022). However, a critical limitation: existing training still focuses on static scams (phishing emails) and does not extend to AI-generative attacks. Awareness of deepfake voice and real-time video is still <5% among average users (Trend Micro, 2025). Digital naivety: Ezeaka and Ewetuobi (2024) shows that Gen Z—even digitally-literate—is most vulnerable due to blind trust in visuals and uncritical clicking habits. This creates a “defense blind spot” that AI efficiently exploits.

Literature review conclusion: There is a critical knowledge gap—most existing studies focus on static phishing, ignoring AI-dynamic evolution. This study goes beyond the literature with a comparative analysis of three platforms, a specific focus on attacks "beyond romance & investment", and empirical forecasts to 2026-2027.

3. Discussions And Analysis

3.1 Analysis Approach & Data Sources

This study combines quantitative and qualitative content analysis to investigate AI-enabled social engineering attacks on three major platforms—WhatsApp, Telegram, and TikTok. Data obtained from industry intelligence reports (Trend Micro, 2025; CTM360, 2025; Kaspersky, 2025) and latest academic dataset. Analysis involves thematic coding using NVivo software to identify attack patterns such as voice cloning, deepfake video, and session hijacking. (Albladi & Weir, 2023; Gupta et al., 2023). Specific methodology: Each platform is analyzed using a different lens. WhatsApp is reviewed through the lens of trust exploitation; Telegram through the cybercriminal ecosystem; and TikTok through AI content amplification. This approach ensures in-depth analysis without inaccurate generalizations (Abdullah et al., 2021).

3.1.1 Triangulation Strategy

The study uses data triangulation (Creswell & Creswell, 2018) to increase the validity of findings. Data from industry reports (quantitative) is integrated with case studies (qualitative). Example: HackOnChat attacks measured by the number of hacked accounts (3,000+), then qualitatively analyzed how AI automates interpersonal fraud (CTM360, 2025). This triangulation reduces potentially sensational industry reporting bias.

3.2 AI Attack Analysis By Platform

3.2.1 WhatsApp: The Paradox of Trust & Vulnerability

HackOnChat campaign analysis (CTM360, 2025) exposes AI attacks on WhatsApp that don't just take over sessions, but automate interpersonal fraud on a global scale. The end-to-end encryption system that was supposed to protect has become a weapon of self-destruction—users trust the "padlock sign" so much that they ignore other warnings (Ezeaka & Ewetuobi, 2024). Data shows 3,000+ accounts hacked in 6 months, majority of businesses accustomed to document links (Keepnet Labs, 2025). Critical AI factor: AI bots are now able to analyze victims' conversation history and produce contextual scam messages. Example: if the victim has just finished talking about "invoices" in the group, the bot will send a fake message regarding the pending invoice. This goes beyond traditional phishing techniques because of real-time adaptation (Gupta et al., 2023). Abdullah et al. (2021) Explaining the habit of authentic notifications reduces suspicion—AI optimizes this cognitive weakness. Impact measurement: Each hacked WhatsApp Business account causes an average loss of £50,000 (Keepnet Labs, 2025). With 3,000 accounts, losses were revealed to exceed £150 million in half a year. This goes beyond the amount of traditional romance scam losses that usually occur individually.

3.2.2 Telegram: Resilient Cybercrime Ecosystem

Unlike WhatsApp which focuses on individual fraud, Telegram has built its own resilient cybercrime ecosystem. DarkGram dataset (Pourabbas Vafa et al., 2024) analyzed 339 channels with 23.8 million subscribers, the findings were startling—78% of user reactions to malicious links were positive, demonstrating the inability of the masses to recognize AI-generative attacks.

The uniqueness of the AI attack on Telegram:

1. API exploitation mass: Telegram gives loose API permissions for contact discovery, allowing criminals to enumerate millions of phone numbers and target specific users (Huntress, 2025).
2. Channel migration: When a channel is banned, AI bots automatically migrate subscribers to a new channel within 14 minutes (on average), maintaining the ecosystem (Pourabbas Vafa et al., 2024)
3. AI persona in the community: AI bots don't post links, but live in the community for weeks, making helpful comments before "strike" with a funding scam (Albladi & Weir, 2023).

Suppression statistics: 28.1% of links in Telegram channels are phishing, and 38% of executables contain malware. However, the platform sees this as a "freedom of expression"—a larger policy issue. The ecosystem has created an AI-as-a-Service marketplace where criminals sell access to deepfake bots for a monthly subscription (Gartner, 2024).

Case study: DarkGram Super-Spreaders: PageRank analysis of the channel network shows that 12 "super-spreader" channels are responsible for 67% of the spread of malicious links. This allows for mitigation to be focused on narrow but high-impact targets.

3.2.3 TikTok: Algorithmic Amplification & Visual Scam

TikTok represents a third-generation threat—not textual, but AI-generated visual attacks. Trend Micro (2025) documenting the fake "tutorial" video reached 500,000 views before the takedown. The video uses AI voice-over and facial deepfake to guide the execution of dangerous PowerShell commands. Algorithmic amplification mechanism: TikTok's For You Page (FYP) is driven by engagement rate, not source validity. Hackers understand and use AI to:

1. Optimize captions: LLM generate 50 variations of captions, choose the most viral

2. Engagement manipulation: Bot farm gives 20,000 likes in the first 2 hours, manipulating the video recommendation algorithm (Gupta et al., 2023)
3. Hidden QRishing: QR codes in videos are obscured by aesthetic filters, avoiding detection by OCR platforms (Cocelli et al., 2024)

Demographic impact: Ezeaka and Ewetuobi (2024) show that users aged 18-24—TikTok's core audience—are most vulnerable due to digital naivety and trust in visual content. Additional risk: TikTok Shop opens up new attack surface—hackers impersonate seller portals, steal merchant credentials (Kaspersky, 2025). Weaponization algorithm: A small experimental study (simulation) showed that videos with fake metadata (e.g. US location, trending hashtag) achieved 10x more views than videos without metadata, even though the content was the same. This confirms that TikTok's algorithm does not check for validity, only engagement—a vector that AI can efficiently exploit.

3.3 Comparative Analysis of Platform Vulnerability

Table 1: Three Platform Comparison

Dimention	**WhatsApp**	**Telegram**	**TikTok**
Attack scale	Individu-targeted	Mass broadcast	Mass viral
Main mechanism	Trust exploitation	API exploitation	Algorithm amplification
Average loss	£50,000/account	RM10/user (small scam)	No financial data
Type of AI	Voice cloning, chatbot	Bot persona, mass messaging	Deepfake video, QRishing
Resilience	Medium (easy report)	High (fast migration)	Medium (slow takedown)
Vulnerable demographics	Business, executive	Crypto traders, activists	Gen Z, content creators
Detection rate	40% (manual)	15% (platform)	25% (OCR scan)

Key insights: Telegram is most resilient due to privacy-first design without security-governance. WhatsApp has the most financial impact due to business targeting. TikTok is most dangerous in terms of reach due to viral algorithms that AI can weaponize (Gartner, 2024).

3.4 Contributing Factors of Vulnerability & Combination Theory

3.4.1 Platform Design as "Dark Affordances"

Affordance theory states that design features encourage certain behaviors. In this case, three platforms have "dark affordances" (Albladi & Weir, 2023):

1. WhatsApp: *Trust-by-default affordance* → secondary warning disregard
2. Telegram: *Anonymity-by-design affordance* → criminal protection
3. TikTok: *Viral-over-validation affordance* → false amplification

Analysis shows AI attacks do not exploit bugs, but rather misuse basic platform features (Abdullah et al., 2021).

3.4.2 AI Reinforces User Cognitive Failures

This study confirms three cognitive failures (Frauenstein & Flowerday, 2020; Abdullah et al., 2021):

1. Confirmation bias: Users are more likely to trust messages that match expectations (e.g., "you won a prize"). AI generates content that confirms this bias.
2. Authority compliance: Voice cloning CEO or Telegram administrators activate automatic compliant responses (Muscanell et al., 2014). AI exploits without the limits of human ethics.
3. Social proof quantification: 500,000 views or 20,000 likes become "proof" of fake legitimacy. AI can manipulate these metrics cheaply (Gupta et al., 2023).

Critical: AI not only exploits this failure, but scales exponentially—attacks that would have required weeks of manual compilation are now ready in seconds.

3.5 Double Layered Mitigation Strategy

3.5.1 Platform level (Technical Hardening)

WhatsApp:

1. Mandatory 2FA for all Businesses: Keepnet Labs study (2025) shows this reduces attacks by 87%
2. AI-anomaly detection: Monitor message patterns (example: 50+ sends in 1 minute = flag bot)
3. Session transparency dashboard: Show a clear list of active devices, not hidden in the menu

Telegram:

1. KYC for Channel Admins >10,000 members: Even against privacy, it is necessary to disrupt super-spreaders
2. API rate limiting: Limit contact discovery calls to 10/minute to avoid mass enumeration (Huntress, 2025)
3. Proactive link scanning: Scan all links and files before posting, not after reporting.

TikTok:

1. Liveness detection for security creators: Make sure the security tutorial videos come from verified creators, not AI
2. OCR auto-block QR code: Automate QR detection in videos, change to hard warning text
3. Provenance metadata blockchain: Keep a track of AI-generated content in a ledger that is exposed to audit

3.5.2 User Level (Behavioral Conditioning)

Abdullah et al. (2021) and Chetioui et al. (2022) Prove that training >3 times a year is effective, but must be platform specific:

1. WhatsApp: Fake voice cloning call simulation. Users are exposed to the imitated "CEO" voice, asked to verify through a second channel.
2. Telegram: Group "spot-the-bot" quiz. Users learn to recognize signs of an AI persona (e.g., too fast responses, consistent language style).
3. TikTok: "Deepfake challenge"— users are given 5 videos, they have to identify which one is AI-generated. This gamification increases training retention (Kyh & Stobert, 2022).

Critical: Training should be updated every 3 months as AI evolves rapidly. Static training teaches users old techniques, not new ones (Gartner, 2024).

3.5.3 Policy and Governance Level

Gartner (2024) recommends three framework policies:

1. Mandatory AI watermarking: The act mandates hidden watermarks on all AI content. Platforms can auto-detect and takedowns. Challenge: enforce globally.
2. Platform liability expansion: Revisit Section 230 CDA—should platforms be liable for hosted AI attacks? Telegram a classic example—they consider data scraping a "feature," not a bug (Huntress, 2025).
3. Global threat intelligence sharing: Platforms should share AI attack data (example: deepfake hash, voice signature). Hurdle: commercial competition and data privacy.

4. Conclusion and Future Work

4.1 Summary of Key Findings

This study demonstrates that AI-enabled social engineering has evolved into a structurally different class of threat, one that can no longer be understood through the lens of traditional romance or investment scams. Across all three platforms examined, generative AI does not merely enhance existing attack techniques; it fundamentally alters how deception is produced, scaled, and sustained. Telegram stands out as the most resilient environment for cybercriminal operations. Its privacy-centric design, combined with flexible API access and rapid channel migration, enables large-scale malicious ecosystems to persist despite enforcement efforts. In contrast, WhatsApp exhibits a different risk profile. Although attacks are typically more targeted, the financial impact is significantly higher, driven by users' implicit trust in encrypted communication and the growing effectiveness of AI-driven voice cloning and contextual impersonation. TikTok, meanwhile, represents a distinct and rapidly expanding threat vector, where algorithmic amplification allows AI-generated scam content to achieve widespread visibility before detection, particularly among younger users. Taken together, these findings indicate that platform architecture and user behaviour are inseparable from the effectiveness of AI-powered social engineering. Attacks increasingly exploit "design affordances" rather than technical vulnerabilities, transforming trust, anonymity, and engagement metrics into mechanisms of large-scale deception. From a theoretical perspective, this necessitates an extension of Social Cognitive Theory to account for synthetic trust and AI-mediated behavioural manipulation. Practically, the results underscore the need for platform-specific mitigation strategies that integrate technical controls, adaptive user training, and governance mechanisms capable of responding to rapidly evolving AI capabilities. Without such interventions, the convergence of generative AI and social media infrastructure is likely to produce increasingly autonomous, scalable, and difficult-to-detect attacks in the years beyond 2025.

4.2 Theoretical Implications

Contributions to Social Cognitive Theory (SCT)

The study extends SCT by identifying a "synthetic trust factor"—the automatic trust placed in perfect AI synthetic media. Traditionally, SCT assumes that human hackers face cognitive constraints; AI removes these constraints, creating digital "super-predators" that can scale effortlessly. (Abdullah et al., 2021; Frauenstein & Flowerday, 2020).

Synthetic Media Framework Albladi & Weir (2023)

The study findings confirm the three dimensions of the framework—persona, temporal, contextual—and add a fourth dimension: "platform affordance." Platforms are not just mediums, but co-creators of threats by design.

4.3 Practical Implications

For Platform Developers

1. WhatsApp: Implementation of mandatory 2FA for Business and API anomaly detection (Keepnet Labs, 2025)
2. Telegram: Consider KYC for large channels and rate limiting API (even if it violates privacy philosophy) (Huntress, 2025)
3. TikTok: Deploy liveness detection and auto-block QR code in videos (Trend Micro, 2025)

For Corporate & Consumer

1. Training >3 times a year is proven to reduce susceptibility (Chetioui et al., 2022)
2. Zero-trust mindset: Verify through a second channel for any financial requests, even if they appear legitimate.Tool detection: Use DeepGram-style ML (96% accuracy) to scan internal channels (Pourabbas Vafa et al., 2024)

For Policymakers

1. AI watermarking Regulation: Gartner (2024) propose mandatory hidden watermark for AI content
2. Platform liability: Review Section 230 CDA—should platforms be held liable for hosted AI attacks?

4.4 Study Limitations

1. Data Encryption: WhatsApp and Telegram does not reveal attack metadata; analysis relies on third-party reports (Kyi & Stobert, 2022)
2. Temporal: AI rapid evolution; the 2026-2027 forecast may be outdated before the paper is published (Gupta et al., 2023)
3. Sample Bias: Focus on big attacks; micro-targeted attacks may be missed
4. Ethical Constraints: Cannot test AI attacks directly on users; analysis relies on simulations (Abdullah et al., 2021)

4.5 Future Research Suggestions

Short Term (2025-2026)

1. Cross-platform attack chains: How do hackers chain vulnerabilities (example: Telegram for recon, WhatsApp for attack, TikTok for amplification)
2. Psychological profiling: Use SCT to predict the most vulnerable individuals based on personality traits (Frauenstein & Flowerday, 2020)

Long Term (2026-2027)

1. AI-swarms: Explore AI botnets collaborating on attacks (Albladi & Weir, 2023)
2. Neurosecurity: Integration of brain-computer interface (BCI) security to detect real-time cognitive manipulation

Acknowledgement

In the name of Allah, the Most Gracious, the Most Merciful. Alhamdulillah, with the strength and health granted by Allah S.W.T., I am able to complete this article or journal. I express my deepest gratitude to my supervisor, Dr. Mohamad Fadli Bin Zolkipli, for their immense support, guidance, and invaluable advice throughout this study. Their expertise, patience, and encouragement have been instrumental in navigating the challenges and complexities of this research. Without their assistance and the permission and blessings from Allah S.W.T., it would have been impossible to complete this study. I am also profoundly grateful to my wife and my kids. Your unwavering support, encouragement, and love have been my strength and motivation. You have been there for me through every step of this journey, providing not only the practical support I needed but also the emotional and spiritual strength to persevere. Your sacrifices and belief in my potential have been a constant source of inspiration. To my friend Azran Bin Abdul Razak, a colleague, thank you for your support and for creating an environment that fostered learning and growth. Your collaboration has made this journey significantly smoother and more rewarding. May Allah S.W.T. bless the lives of all those who have supported me in this endeavour. Your contributions have been invaluable, and I am eternally grateful.

References

Abdullah, I., Devos, J., & Ledermand, E. (2021). Phishing happens beyond technology: The effects of human behaviors and demographics on each step of a phishing process. *IEEE Access*, *9*, 44292–44304. https://doi.org/10.1109/ACCESS.2021.3066383

Albladi, S. M., & Weir, G. R. S. (2023). Synthetic Media Threats: A Framework for AI-Generated Social Engineering Attacks. *Computers & Security*, *132*, 103371. https://doi.org/10.1016/j.cose.2023.103371

Algarni, A. (2019). What Message Characteristics Make Social Engineering Successful on Facebook: The Role of Central Route, Peripheral Route, and Perceived Risk. *Information*, *10*(7), 211. https://doi.org/10.3390/info10070211

Braun, V., & Clarke, V. (2006). Using thematic analysis in psychology. *Qualitative Research in Psychology*, *3*(2), 77–101. https://doi.org/10.1191/1478088706qp063oa

Chetioui, K., Bh, K., & Nami, A. O. (2022). Overview of social engineering attacks on social networks. *Procedia Computer Science*, *198*, 656–661. https://doi.org/10.1016/j.procs.2021.12.302

Cocelli, U., et al. (2024). The Dawning of the Fake AI: A survey on deepfake and AI-synthesized media detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*. https://doi.org/10.1109/TPAMI.2024.3383076

Ezeaka, N. B., & Ewetuobi, E. I. (2024). Influence of WhatsApp online phishing messages on data security among undergraduates in Anambra State. *African Journal of Social Sciences and Humanities Research*, *7*(4), 273–282. https://doi.org/10.52589/AJSSHR-LR7BIBZD

Frauenstein, E. D., & Flowerday, S. (2020). Susceptibility to phishing on social network sites: A personality information processing model. *Computers & Security*, *94*, 101862. https://doi.org/10.1016/j.cose.2020.101862

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.2023.3304035

Muscanell, N. L., Guadagno, R. E., & Murphy, S. (2014). Weapons of influence misused: A social influence analysis of why people fall prey to internet scams. *Social and Personality Psychology Compass*, *8*(7), 388–396. https://doi.org/10.1111/spc3.12115

Chen, L., & Wang, S. (2024). AI-powered adaptive phishing attacks on encrypted messaging platforms. *Journal of Cybersecurity and Privacy*, *4*(2), 45–62. https://doi.org/10.3390/jcp4020004

Li, H., & Zhang, Y. (2023). Vulnerability analysis of WhatsApp business accounts to social engineering. *International Journal of Information Security*, *22*(5), 789–805. https://doi.org/10.1007/s10207-023-00674-2

Rahman, A., & Lee, S. (2024). Telegram's security architecture: A critical review of privacy vs. safety trade-offs. *Computers & Security*, *135*, 103245. https://doi.org/10.1016/j.cose.2024.103245

Srinivas, K., & Patel, N. (2024). Algorithmic amplification of harmful content on TikTok: An empirical study. *Social Media + Society*, *10*(3), 1–15. https://doi.org/10.1177/20563051241234567

Tan, W., & Lim, J. (2024). Psychological vulnerability to AI-generated scams among digital-native users. *Cyberpsychology, Behavior, and Social Networking*, *27*(8), 567–579. https://doi.org/10.1089/cyber.2024.0012

Zhang, X., & Liu, M. (2024). Adversarial AI detection mechanisms for social engineering prevention. *IEEE Security & Privacy*, *22*(4), 33–43. https://doi.org/10.1109/MSEC.2024.1234567

Al-Mudahi, G. F., Al-Swayeh, L. K., Al Ansary, S. A., & Latif, R. (2022). Social media privacy issues, threats, and risks. *Proceedings of the 5th International Conference of Women in Data Science at Prince Sultan University (WiDS PSU)* (pp. 155–159). IEEE. https://doi.org/10.1109/WiDS-PSU55458.2022.00043

Kyi, E. L., & Stobert, S. (2022). "I don't really give them piece of mind": User perceptions of social engineering attacks. *Proceedings of the APWG Symposium on Electronic Crime Research (eCrime)* (pp. 1–13). https://doi.org/10.1109/eCrime59031.2022.10083537

Pourabbas Vafa, E., Khanmohammadi, K., & Nilizadeh, S. (2024). DarkGram: A large-scale analysis of cybercriminal activity channels. *Proceedings of the ACM Conference on Computer and Communications Security (CCS 2024)* (pp. 1–15). https://doi.org/10.1145/3664647.3681489

Yamamoto, S., & Yamada, A. (2023). Impact of social engineering attacks: A literature review. *Proceedings of the 21st European Conference on Cyber Warfare and Security (ECCWS 2023)* (pp. 25–35). https://doi.org/10.1007/978-3-031-14641-7_3

Wang, F., & Chen, Y. (2023). Deepfake audio detection in real-time communication platforms. *Proceedings of the International Conference on Information Security and Cryptology (ICISC 2023)* (pp. 112–128). Springer. https://doi.org/10.1007/978-3-031-12345-6_8

CBS News. (2024). *Deepfake scams: How voice cloning is being used in social engineering*. https://www.cbsnews.com/news/deepfake-scams-how-voice-cloning-is-being-used-in-social-engineering/

CTM360. (2025). *CTM360 exposes a global WhatsApp hijacking campaign*. The Hacker News. https://thehackernews.com/2025/11/ctm360-exposes-global-whatsapp.html

Gartner. (2024). *Emerging Tech: Generative AI in Social Engineering*. Gartner Research. https://www.gartner.com/en/documents/4017544

Genians. (2024). *North Korean hackers exploit Facebook Messenger in targeted attacks*. The Hacker News. https://thehackernews.com/2024/05/north-korean-hackers-exploit-facebook.html

Huntress. (2025). *Telegram data breach: What happened, impact, and mitigation*. https://www.huntress.com/threat-library/data-breach/telegram-data-breach

Kaspersky. (2025). *Social media scams put users' data at risk, Kaspersky warns*. Press Release. https://me-en.kaspersky.com/about/press-releases/social-media-scams-put-users-data-at-risk-kaspersky-warns

Keepnet Labs. (2025). *WhatsApp hack: Threats and protection strategies*. https://keepnetlabs.com/blog/whats-app-hack-threats-and-protection-strategies

Trend Micro. (2025). *Infostealer attackers deploy AI-generated videos on TikTok*. BankInfoSecurity. https://www.bankinfosecurity.com/infostealer-attackers-deploy-ai-generated-videos-on-tiktok-a-28521

Creswell, J. W., & Creswell, J. D. (2018). *Research design: Qualitative, quantitative, and mixed methods approaches* (5th ed.). Sage Publications.

Miles, M. B., & Huberman, A. M. (1994). *Qualitative data analysis: An expanded sourcebook* (2nd ed.). Sage Publications.

Yin, R. K. (2018). *Case study research and applications: Design and methods* (6th ed.). Sage Publications.