

## Article

# AI-Powered Social Engineering: Emerging Attack Vectors, Vulnerabilities, and Multi-Layered Defense Strategies <sup>†</sup>

Kely Gonzaga <sup>1</sup>, Sérgio Serra <sup>1</sup>, Marco Gomes <sup>2,3</sup> and Silvestre Malta <sup>1,4,\*</sup>

<sup>1</sup> ESTG, Instituto Politécnico de Viana do Castelo, 4900-347 Viana do Castelo, Portugal; kelygonzaga@ipvc.pt (K.G.); sergioserra@ipvc.pt (S.S.)

<sup>2</sup> ESTG, Polytechnic of Porto, 4610-156 Felgueiras, Portugal; mfg@estg.ipp.pt

<sup>3</sup> CIICESI, Instituto Politécnico do Porto, 4610-156 Felgueiras, Portugal

<sup>4</sup> ADiT-Lab, Instituto Politécnico de Viana do Castelo, 4900-347 Viana do Castelo, Portugal

\* Correspondence: smalta@estg.ipvc.pt

<sup>†</sup> This article is a revised and expanded version of a paper entitled Exploiting Trust: How Artificial Intelligence and Deepfakes Shape Social Engineering Attacks, presented at SASYR 2025—5th Symposium of Applied Science for Young Researchers, Viana do Castelo, Portugal, 2 July 2025.

## Abstract

In the past decade, a growing number of cyberattacks have been reported, enabling unprecedented levels of personalization, automation, and deception. For instance, recent industry surveys have reported sharp increases in unique social engineering attacks within a single month of 2023, coinciding with the public release of ChatGPT-3.5. This trend highlights how Artificial Intelligence (AI)-powered phishing campaigns have become a significant threat to digital ecosystems. The present study provides an integrative analysis of how generative and deepfake technologies have reshaped the landscape of a Social Engineering (SE) attack, categorizing the main attack strategies and examining their psychological, technological, and ethical implications. In addition, to reviewing enabling technologies, our study conducts a comparative analysis of frameworks and analytical models across technical, empirical, and quantitative perspectives that model AI-driven SE operations and their defensive countermeasures. The convergence of these frameworks reveals three core capabilities—realism, personalization, and automation—that systematically amplify attack efficiency. Building on these insights, the study proposes the Unified Model for AI-Driven Social Engineering (UM-AISE), a conceptual framework that integrates these dimensions across the attack lifecycle and employs a theoretical Markov Decision Process (MDP) analysis. This formalization demonstrates how these capabilities can shift the attacker's optimal strategy, offering a formal economic perspective distinct from empirical validation. Finally, the study discusses emerging ethical and regulatory challenges associated with AI-mediated deception, highlighting risks related to opacity, accountability, and large-scale manipulation. Taken together, these elements inform evolving approaches for detection, defense, and governance relevant to researchers, policymakers, and practitioners.



Academic Editor: Paolo Bellavista

Received: 5 January 2026

Revised: 11 February 2026

Accepted: 13 February 2026

Published: 17 February 2026

Copyright: © 2026 by the authors.

Licensee MDPI, Basel, Switzerland.

This article is an open access article distributed under the terms and conditions of the [Creative Commons Attribution \(CC BY\)](https://creativecommons.org/licenses/by/4.0/) license.

**Keywords:** AI-driven social engineering; generative AI; large language models; deepfake attacks; Cybersecurity defenses; psychological manipulation; risk frameworks

## 1. Introduction

Recent advancements in AI have significantly reshaped the digital landscape. It has greatly aided and increased human abilities in numerous areas of expertise, while also dramatically increasing cybersecurity risks, particularly in the domain of SE. As

technological innovation progresses, so do the tools employed by cybercriminals [1], who exploit AI's ability to mimic human behavior [2] to manipulate, infiltrate, and ultimately compromise both personal and institutional security [3].

Traditionally, SE has relied on exploiting human emotions such as trust, fear, curiosity, and urgency. However, with the advent of Generative AI (GenAI), Large Language Models (LLMs), and deepfake technologies, these attacks have evolved from simple psychological manipulations with limited reach to highly automated, scalable, and personalized operations [3,4]. Researchers from Darktrace reported a 135% spike in unique social engineering attacks from January to February 2023, coinciding with the widespread adoption of ChatGPT [5]. Additionally, recent empirical and survey-based evidence suggests that GenAI not only improves linguistic plausibility, but also reshapes the trust surface exploited by attackers, including victims' acceptance of AI-generated identities and profiles [6]. Such shifts in cyber trust and distrust factors reinforce why AI-mediated deception must be analyzed through a socio-behavioral risk lens, as trust dynamics and perceived credibility shape victim decision-making beyond purely technical factors [7].

The pervasive presence of social media platforms and digital communication tools has significantly expanded the attack outreach, making personal information more accessible through Open-Source Intelligence (OSINT) and enabling attackers to craft highly targeted campaigns [4,7]. As emphasized by Alahmed et al. [8] and Balasubramanian et al. [9], generative models now empower attackers to simulate empathy, emotional tone, and linguistic cues that convincingly mirror human communication, blurring the boundary between authentic and synthetic interaction. According to the UK Cyber Security Breaches Survey (Department for Science, Innovation and Technology, "Cyber security breaches survey 2024", Gov.uk, April 2024. Available at: <https://www.gov.uk/government/statistics/cyber-security-breaches-survey-2024/cyber-security-breaches-survey-2024> (accessed on 23 December 2025)), SE is responsible for over 84% of corporate breaches. This number reveals that human-targeted deception, now amplified by AI, is among the most pervasive and costly threats in cybersecurity. In addition, recent enterprise analyses indicate that the primary threat of offensive AI, defined by Mirsky et al. as "the use or abuse of AI to accomplish a malicious task", lies in its ability to enhance social engineering, especially impersonation, and accelerate the initial stages of cyberattacks, such as reconnaissance [10]. Moreover, recent work in the financial crime domain frames GenAI-enabled deception as a rapidly escalating, co-evolutionary "AI vs. AI" arms race, and notes that institutional readiness is lagging behind technological capability [1,11].

In line with this concern, Kurshan et al. cite that the U.S. Treasury has warned that existing risk management frameworks in financial services may not be adequate to cover emerging AI technologies, underscoring a growing governance gap [1]. This gap stems from a fundamental shift in offensive capabilities described by Mirsky et al.: unlike static attack vectors, AI-capable adversaries can now "act with micro-precision, but at macro-scale and with greater speed," targeting individuals with a level of granularity that current defense apparatuses are ill-equipped to counter [10].

A striking illustration of this evolution occurred in 2024, when an employee at a multinational firm was deceived into transferring approximately US\$25 million (HK\$200 million) after participating in a video call where participants, purportedly the company's CFO and other senior staff, were in fact AI-generated impersonations (This incident was widely reported by major news outlets. See, for example: The Guardian (<https://www.theguardian.com/world/2024/feb/05/hong-kong-company-deepfake-video-conference-call-scam>) and CNN (<https://edition.cnn.com/2024/02/04/asia/deepfake-cfo-scam-hong-kong-intl-hnk/index.html>) (accessed on 10 June 2025)). The fraud began with a seemingly routine email requesting a confidential transaction, but

the use of a multi-participant video conference featuring convincing deepfakes eroded the employee's initial suspicion and triggered a chain of financial transfers [12]. This case exemplifies three critical pillars of modern AI-enhanced SE: highly persuasive tailored messaging, real-time generative impersonation, and scale potential. It demonstrates how the blending of voice cloning, identity spoofing, and legitimate corporate communication channels creates a novel attack modality that traditional defenses are poorly equipped to counter. Furthermore, the democratization of AI has significantly lowered the technical barrier to entry [11,13,14]. Recent studies demonstrate that even novice users, with no prior expertise in cyber-offense, can now execute sophisticated phishing campaigns using readily available models like ChatGPT-4o Mini [15].

Given that, despite growing academic attention to cyber threats, substantial gaps persist in research on AI-augmented social engineering. Existing studies focus predominantly on traditional phishing [4], leaving other AI-driven modalities, such as chat-based software, deepfake impersonation and voice cloning, largely unexplored [16,17]. Surveys tend to describe the problem rather than analyze it quantitatively; few address measurable indicators such as the psychological impact of synthetic media or detection accuracy against multimodal forgeries. Moreover, the psychological vulnerabilities exploited by AI-generated persuasive content remain poorly understood [18], and current detection frameworks struggle to address the scale and sophistication of automated, personalized attacks [8]. Additionally, ethical and regulatory responses to AI misuse in SE are underdeveloped [19]. This study addresses these gaps through a systematic literature review that integrates technical, psychological, and policy dimensions. By bridging empirical research and conceptual frameworks, it seeks to transform a fragmented body of knowledge into a coherent understanding of how AI technologies redefine deception at scale. Therefore, a critical synthesis that connects advances in generative AI, psychological persuasion mechanisms, and governance models is essential to guide both research and practice.

In light of this context, the present research seeks to analyze the trajectory of AI-powered SE attacks, consolidate insights from recent studies, and propose a forward-looking framework for detection and defense. Beyond synthesizing the literature, this study introduces the UM-AISE, a conceptual framework supported by a theoretical quantitative demonstration, which integrates technological, psychological, and operational insights to explain how generative AI systems reshape the full attack lifecycle and its defensive counterparts.

The paper is organized as follows: Section 2 presents the systematic literature review, detailing the methodology adopted for article selection and providing a critical analysis of the technological evolution of AI in social engineering—range. This section also evaluates existing frameworks to identify the operational gaps that motivate this study. Section 3 shifts from state-of-the-art analysis to a theoretical contribution, proposing the Unified Model of AI-mediated Social Engineering (UM-AISE) and mathematically validating the attacker's incentive structure through a Markov Decision Process (MDP) analysis. Section 4 addresses the ethical, legal, and regulatory challenges raised by the use of AI in deceptive practices. Finally, Section 5 presents the answers to the research questions, and Section 6 summarizes the key findings and outlines directions for future research and policy development.

## 2. Literature Review

To establish the theoretical and analytical foundations for this study, this section examines the increasing sophistication of AI-driven social engineering. The analysis traces the technological evolution of the field, moving from early Machine Learning (ML) applications to the advanced capabilities of Generative AI and synthetic media. In addition, we evaluate

how existing analytical frameworks have attempted to model these emerging risks. Finally, in Section 2.4, we synthesize these findings to identify specific operational and quantitative gaps in the literature, providing the justification for the construction of the unified model proposed in Section 3.

### 2.1. Methodology

This section describes the methodological procedures used to conduct a systematic literature review on AI-driven social engineering. The approach follows the guidelines of [20], adapted to the specific context of emerging AI-enabled cyber threats. The review focuses on studies published between 2020 and 2025, a timeframe selected to encompass the paradigm shift triggered by the release of GPT-3 and the subsequent proliferation of accessible synthetic media tools [21], covering technical, psychological, and regulatory perspectives relevant to AI-mediated SE attacks.

To organize the objectives of this review, a set of research questions (RQs) was formulated to guide the search strategy, the selection of studies, and the synthesis of evidence. These RQs reflect the technological, psychological, and regulatory dimensions identified as central to AI-mediated SE.

- RQ1—What are the primary methods by which AI is utilized to carry out SE attacks?
- RQ2—How do LLMs, deepfake technologies, and generative AI tools contribute to the efficacy of SE campaigns?
- RQ3—What are the primary psychological and technological vulnerabilities exploited by AI-enabled SE tactics?
- RQ4—What defense mechanisms and mitigation strategies have been proposed?
- RQ5—What ethical and regulatory concerns arise from the malicious use of AI in SE?

These RQs informed the definition of search terms, the boolean query structure, and the inclusion/exclusion criteria used in the screening process. From these RQs, a set of keywords was defined, grouped by domain and combined through logical operators, as shown in Table 1.

**Table 1.** Keyword groups used to construct the search queries.

| Group              | Keywords                                                                                  |
|--------------------|-------------------------------------------------------------------------------------------|
| AI Technologies    | Artificial Intelligence, Machine Learning, Large Language Models, Generative AI, Deepfake |
| Social Engineering | Social Engineering, Phishing, Deception, Cyber Attack, Manipulation                       |

The defined keywords were combined using Boolean operators to refine the scope and improve the precision of search results. This step ensured that retrieved documents addressed the intersection between AI technologies and SE phenomena. All boolean expressions were applied identically across the three databases to ensure reproducibility and consistency of results, using the 2020–2025 publication window as a filter during the search process. Table 2 presents the boolean expressions applied in the searches.

Three major scientific databases were selected for their coverage of high-impact peer-reviewed research in artificial intelligence and cybersecurity: IEEE Xplore, Scopus, and ScienceDirect.

To ensure methodological rigor, three exclusion criteria (EX1–EX3) were applied sequentially. EX1 and EX2 were used during title/abstract screening, whereas EX3 was applied during full-text eligibility assessment (Table 3).

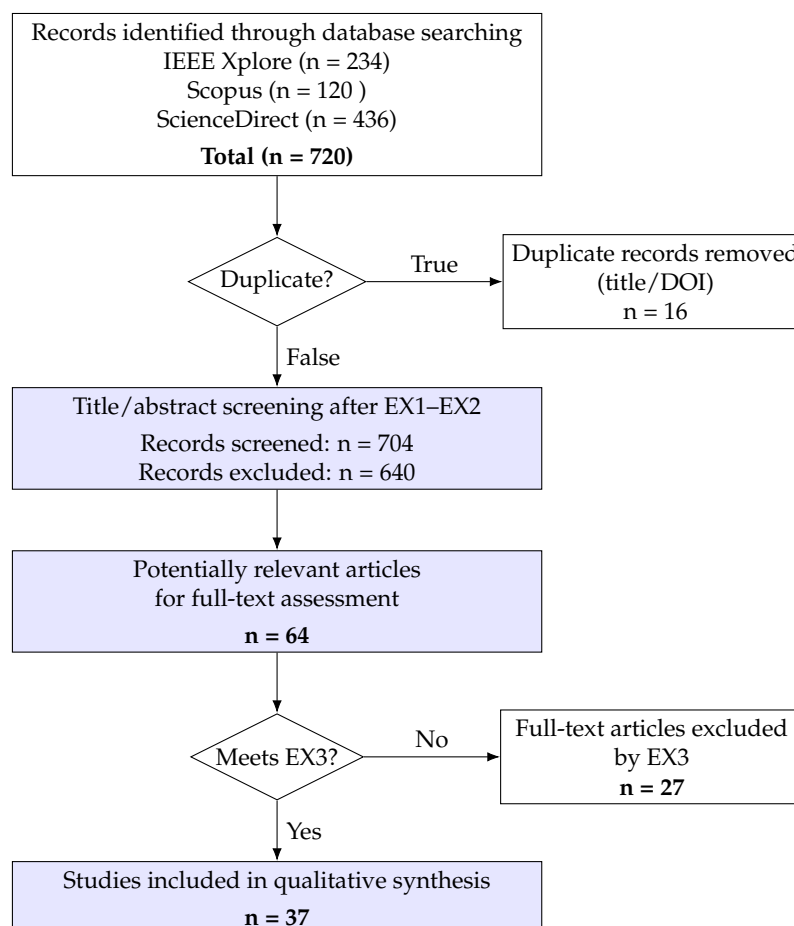
**Table 2.** Boolean Queries Used in the Search.

| OR Group 1 (AI-Related Terms)                                    | AND Operator | OR Group 2 (SE-Related Terms)                        |
|------------------------------------------------------------------|--------------|------------------------------------------------------|
| “Artificial Intelligence” OR “Machine Learning”<br>OR “Deepfake” | AND          | “Social Engineering” OR “Cybersecurity”              |
| “Large Language Models” OR “Generative AI”<br>OR “Deepfake”      | AND          | “Social Engineering” OR “Phishing”<br>OR “Deception” |

**Table 3.** Exclusion Criteria Applied During Screening.

| Code | Description                                                                                                                                                   |
|------|---------------------------------------------------------------------------------------------------------------------------------------------------------------|
| EX1  | Studies not written in English or without full-text availability.                                                                                             |
| EX2  | Studies not directly related to AI-enabled social engineering or its cybersecurity implications.                                                              |
| EX3  | Non-peer-reviewed sources (e.g., news articles and editorials), venues of questionable credibility, or studies for which the full text could not be obtained. |

The searches were conducted between January and February 2026 using the Boolean strings shown above across the selected databases. The database search initially returned 720 records. After removing duplicates and applying the exclusion criteria EX1 and EX2, during title and abstract screening, 64 papers remained eligible for full-text assessment. Following EX3 criteria, 37 studies were retained for qualitative synthesis; additional references were used solely for contextualization. The complete selection process is summarized in Figure 1.

**Figure 1.** Flow diagram summarizing the identification, screening, eligibility assessment, and inclusion stages of the review (see Table 4).

The final set of 37 studies, obtained after full-text eligibility assessment, is summarized in Table 4. These works constitute the analytical foundation for the qualitative synthesis presented in this section, as well as the ethical discussion in Section 4 and the answers to the research questions in Section 5. Although the corpus is deliberately selective, it reflects a quality-driven screening process designed to retain only robust, peer-reviewed evidence within the defined scope of this review.

**Table 4.** Summary of the 37 Primary Studies Included in the Integrative Review.

| Citation | Year | AI Tech       | Focus           | Key Findings                                                                         | Mitigation                                    |
|----------|------|---------------|-----------------|--------------------------------------------------------------------------------------|-----------------------------------------------|
| [22]     | 2025 | Custom GPTs   | App Store Abuse | Analysis of 755k GPTs reveals industrial-scale misuse and evasion tactics.           | Automated vetting; API monitoring.            |
| [14]     | 2025 | GPT-4         | Hybrid Attacks  | “Centaur” attacks (Human + AI) are harder to detect than fully automated ones.       | Training on co-created content.               |
| [11]     | 2025 | LLMs          | Phishing Evol.  | “Scale-Personalization Conflict” theory; AI solves the volume vs. quality trade-off. | Detection of machine-generated patterns.      |
| [7]      | 2025 | Deepfakes     | Trust Modeling  | Game-theoretic model of “Cyber Distrust” dynamics in deepfake interactions.          | Socio-economic risk assessment.               |
| [23]     | 2025 | AI Detection  | Law/Ethics      | Technical detection is insufficient; proposes legal framework (C2PA).                | Provenance tracking; legal accountability.    |
| [24]     | 2025 | GenAI         | TTP Framework   | Comprehensive matrix of Social Engineering Tactics, Techniques, and Procedures.      | Mapping defenses to SE kill chain.            |
| [4]      | 2025 | ML/Blockchain | Hybrid SE       | Hybrid systems improve data integrity against SE; Blockchain for URL storage.        | Blockchain verification; MFA.                 |
| [18]     | 2025 | LLMs          | Psychology      | Maps Cialdini’s principles to LLM prompts; AI automates rapport-building.            | Persuasion-aware classifiers.                 |
| [25]     | 2025 | GenAI         | Workplace       | Analyzes the “dual nature” of GenAI in corporate settings (tool vs. threat).         | Workplace monitoring policies.                |
| [26]     | 2025 | Modeling      | Awareness       | Applies optimal control theory to optimize security training costs.                  | Optimal budget allocation policies.           |
| [9]      | 2025 | GenAI         | Threat Intel    | Surveys GenAI for both offensive malware generation and defensive CTI.               | AI-driven threat intelligence.                |
| [15]     | 2025 | GPT-4o        | Jailbreaking    | Novice users can execute complex phishing campaigns using jailbroken models.         | Robust alignment; safety filters.             |
| [27]     | 2025 | Gemini/QR     | Hybrid Phishing | Demonstrates QR-based “Browser-in-the-Browser” attacks via AI.                       | Visual analysis; QR scanning protection.      |
| [28]     | 2025 | Deepfakes     | Vishing         | Analysis of real-time audiovisual manipulation in video conferencing.                | Liveness detection; organizational protocols. |
| [12]     | 2025 | LLMs          | Scenario Gen.   | Uses LLMs to generate synthetic SE scenarios for resilience training.                | AI-driven awareness simulation.               |

Table 4. Cont.

| Citation | Year | AI Tech      | Focus           | Key Findings                                                                     | Mitigation                                   |
|----------|------|--------------|-----------------|----------------------------------------------------------------------------------|----------------------------------------------|
| [29]     | 2025 | LLM Agents   | Agentic Risks   | Highlights risks of autonomous AI agents executing multi-step attacks.           | Behavior monitoring; access controls.        |
| [30]     | 2025 | LLMs         | Fake CTI        | Study on AI-generated fake Cyber Threat Intelligence (data poisoning).           | Verifiable CTI sources.                      |
| [3]      | 2024 | GenAI        | SE Framework    | Defines Realism, Automation, and Personalization as key amplifiers.              | Watermarking; rate limiting.                 |
| [8]      | 2024 | Chatbots     | Victim Trust    | Empirical test of “Impostor Bias”; specific demographics trust AI profiles more. | Role-based user awareness.                   |
| [31]     | 2024 | LLM Phishing | Victim Data     | Analysis of 25k victims proving personalization increases success rates.         | Behavioral forensics; sentiment analysis.    |
| [1]      | 2024 | GenAI        | Financial Crime | “Co-evolutionary AI” (AI vs. AI) arms race in financial fraud.                   | Multi-modal authentication.                  |
| [2]      | 2024 | AI Malware   | Automation      | AI automates CAPTCHA bypass and malware mutation.                                | Anomaly detection; verbal verification.      |
| [19]     | 2024 | AI Risks     | Ethics/Hack     | AI scales risks in hacking and misinformation; highlights ethical gaps.          | Explainability; ethical AI guidelines.       |
| [32]     | 2024 | Social Bots  | Botnets         | Empirical analysis of “Fox8” botnet; AI mimics human interaction.                | Coordination analysis; platform suspension.  |
| [33]     | 2024 | Chatbots     | Smishing        | “AbuseGPT”: creating smishing campaigns that evade filters.                      | SMS content analysis.                        |
| [34]     | 2024 | LLMs         | Code Gen.       | AI automates the generation of attack code (polymorphic malware).                | AI-based code analysis.                      |
| [35]     | 2024 | Deepfakes    | Detection       | Systematic review of deepfake generation vs. detection techniques.               | Multimodal detection algorithms.             |
| [13]     | 2024 | TTS/LLM      | Voice Cloning   | Case study on Tortoise-TTS and LLMs for creating voice phishing.                 | Audio artifact detection.                    |
| [10]     | 2023 | Offensive AI | Threat Model    | Defined the “AI-Capable Adversary” (micro-precision at macro-scale).             | Moving Target Defense (MTD).                 |
| [16]     | 2023 | GANs         | Deepfakes       | Survey of creation types (face swap, lip-sync) and detection challenges.         | Media forensics; regulatory frameworks.      |
| [36]     | 2023 | Prompt Eng.  | Persuasion      | LLMs manipulated via prompts (urgency/authority) to bypass safety.               | Human-in-the-loop; psychological safeguards. |
| [37]     | 2023 | Deepfakes    | Real-time       | Analysis of real-time deepfake threats in video calls.                           | Challenge-response protocols.                |
| [38]     | 2022 | MDP          | Math Model      | Formalizes SE as a Markov Decision Process (basis for UM-AISE).                  | Increasing operational friction.             |

Table 4. Cont.

| Citation | Year | AI Tech    | Focus              | Key Findings                                                             | Mitigation                              |
|----------|------|------------|--------------------|--------------------------------------------------------------------------|-----------------------------------------|
| [17]     | 2022 | CNNs       | Persuasion         | Early-stage recognition of persuasion in chat-based SE using embeddings. | Semantic analysis of chat logs.         |
| [6]      | 2022 | Deepfakes  | Synthetic Profiles | Empirical evidence that AI-generated faces are perceived as trustworthy. | Identity verification hardening.        |
| [39]     | 2022 | General ML | Crime Evol.        | Classifies malicious AI use; expansion of vulnerabilities.               | Governance; interdisciplinary response. |
| [40]     | 2021 | Deepfakes  | Misinformation     | Technical survey of Deepfake creation/detection risks.                   | Biological signal analysis (blinking).  |

## 2.2. AI in Social Engineering

The technological foundations of AI-based social engineering reflect a gradual and consistent merging of machine learning, language processing, and generative modeling, which together have reshaped digital deception. Early uses of AI in this field relied on fixed rules and limited datasets, but advances in algorithms and computing power have led to systems capable of imitating human reasoning and exploiting linguistic and behavioral cues with remarkable accuracy [8,21]. Thus, this subsection examines how these technologies have been integrated into SE through three main aspects.

### 2.2.1. Evolution and Applications of ML in SE

The use of AI in SE is not a recent phenomenon, although its impact has intensified with the advancement of ML technologies. According to Blauth et al. [39], one of the earliest documented uses of AI in this context was CyberLover, a chatbot developed in 2007 that employed Natural Language Processing (NLP) to simulate chatroom conversations in order to deceive users and collect sensitive information. These early experiments demonstrated the potential of AI to create human-like interactions that exploit emotional and behavioral vulnerabilities.

In the following decade, the development of social bots capable of mimicking believable profiles on social media highlighted the scalability of digital manipulation [24]. These bots were widely used in disinformation campaigns, public opinion manipulation, and electoral interference through strategies such as astroturfing and coordinated retweeting [39].

Recent literature outlines a clear evolutionary trajectory in the weaponization of AI [1,7]. As noted by Blauth et al. [39], the landscape has evolved from simple text-based manipulations and chatbots to complex social bots capable of coordinated campaigns, and finally to the generation of hyper-realistic synthetic media. This progression illustrates how attackers have moved beyond mere efficiency to achieve capabilities that were previously impossible. A notable example of this transition is *DeepLocker*, an experimental malware designed to conceal its malicious intent until the victim is identified by facial recognition, geolocation, or behavioral analysis, illustrating the power of targeted AI-enabled attacks.

This historical progression was systematized by Liu et al. [11] into four distinct evolutionary stages of phishing generation: (1) manual template-based, (2) programmatic heuristic-based, (3) ML-based (employing RNNs and CNNs), and finally (4) the current LLM-driven era. This taxonomy highlights that the transition to ML was not merely an increase in speed, but a structural shift designed to overcome the static, detectable nature of rule-based templates.

Currently, the growing sophistication of SE attacks is largely driven by the use of ML models. Supervised and unsupervised learning techniques are used to craft highly

personalized phishing emails and fake social media profiles, significantly increasing the effectiveness of such campaigns. In supervised learning, models are trained on labeled datasets containing legitimate and fraudulent communications. These models learn to recognize linguistic patterns, sender attributes, and embedded link characteristics, enabling the creation of messages that closely mimic authentic ones [3,14].

The same principle applies to the creation of deceptive social media profiles. Models trained on datasets of genuine and malicious accounts can accurately replicate features such as account creation date, follower count, posting frequency, and writing style, facilitating fraud and manipulation campaigns [8]. Unsupervised learning, on the other hand, enables the detection of behavioral anomalies in social media activity or email traffic without the need for labeled data. This allows for the identification of irregular patterns that may reveal potential vulnerabilities or targets [3].

As Alahmed et al. [8] point out, “AI-generated messages exhibit improved contextual awareness and persuasive language”, demonstrating how attackers leverage ML to extract and analyze large volumes of data from social media platforms, public databases, or data breaches to compose targeted messages tailored to the victim’s interests, roles, or recent activity. This level of personalization marks a significant advancement over generic phishing tactics, allowing for narratives that resonate with the target’s context, instill trust and urgency, and significantly increase the success rate of SE campaigns.

Looking beyond current predictive capabilities, Li and Fung [29] identify a nascent evolutionary leap towards “Agentic Risks”. In this phase, AI systems transition from passive information processing to active execution, where autonomous agents can independently browse the web, utilize tools, and execute multi-stage workflows without human intervention. This shift represents the apex of ML application in SE, where the attacker defines the goal, and the system autonomously navigates the operational steps to achieve it. Consequently, this emergence of self-governing offensive loops provides the empirical ground for the upper bound of attack automation, validating a transition toward a regime of fully autonomous, zero-touch orchestration.

#### 2.2.2. Prompt Engineering and Exploitation in Language Models

While Section 2.2.1 explored how ML models enable personalized and large-scale SE attacks, an equally critical dimension of AI-driven SE lies in the manipulation of language models through prompt engineering [9,13,22]. Models such as ChatGPT, Claude, Gemini, and WormGPT are not only capable of generating coherent and context-sensitive text, but also of sustaining dynamic, adaptive conversations that are often indistinguishable from human interaction [3,33].

Singh et al. [36] demonstrated that well-crafted prompts based on psychological principles such as authority, scarcity, urgency, and reciprocity can cause LLMs to bypass ethical safeguards and produce harmful outputs. As the authors point out, “well-structured psychologically grounded prompts can override LLM safety mechanisms and elicit dangerous outputs”. They also propose a taxonomy that categorizes prompt manipulation strategies according to common behavioral influence techniques exploited by malicious actors.

Furthermore, prompt injection attacks have emerged as a subtle yet powerful method of exploiting vulnerabilities in LLMs. These attacks involve embedding hidden or ambiguous instructions within user inputs, allowing adversaries to bypass model constraints and induce the generation of disinformation, impersonation content, or unintended behaviors [3,13,22]. Balasubramanian et al. highlights that “prompt injection and data poisoning could significantly impact threat detection and the integrity of intelligence workflows” [9].

The sophistication of these attacks is further amplified by the emergence of LLMs specifically architected for offensive purposes, such as WormGPT [7]. As analyzed by Kur-

shan et al. [1], these unrestricted models drive a new phenomenon of ‘GenAI crime waves’, where malware generation, phishing email automation, exploit development and others are effectively distributed on dark web forums for criminal operations, democratizing cyber-offensive capabilities. By removing the technical barriers previously required for such operations, these Jailbreaking-As-A-Service business models allow malicious actors to easily launch complex, multi-stage attacks with unprecedented scale and speed, fundamentally altering the threat landscape [1,7,15]. Not only that, industrialization of AI-SE is further evidenced by the systematic misuse of official AI marketplaces. Shen et al. [22] demonstrate large-scale analysis using the GPTracker framework that identifies thousands of custom agents explicitly designed to facilitate forbidden activities, revealing that builders employ external APIs to bypass platform safety filters, which increases the response rate for illegal queries by an average of 23%. An example of these vulnerabilities within conversational AI platforms is the “AbuseGPT” framework, proposed by Shibli et al. [33]. Their work demonstrates how generative chatbots can be manipulated to autonomously generate smishing (SMS phishing) campaigns that are not only persuasive but also capable of evading linguistic filters by varying the message structure while retaining the malicious intent.

Building on this potential for misuse, recent demonstrations have shown how automation can be weaponized to code complex hybrid attacks. Akram et al. [27] utilized Google Gemini to engineer a ‘QR-based Browser-in-the-Browser (BiTB) phishing attack’. By obfuscating malicious URLs behind dynamic QR codes that launch hyper-realistic fake login windows, this method effectively circumvents traditional text-based detection systems to harvest sensitive user credentials, demonstrating how easily official models can be weaponized for hybrid threats. Furthermore, this automation extends beyond specific phishing vectors. Iturbe et al. [34] demonstrated that LLMs can generate executable attack code across the entire intrusion workflow, producing polymorphic variants that complicate detection. Crucially, they also highlight that this capability possesses a “dual-use nature”: the same automated workflows can be repurposed defensively within Breach-and-Attack Simulation (BAS) frameworks to stress-test security pipelines against rapidly mutating threats. However, this ambiguity creates a semantic blind spot for safety filters; the authors observed that ChatGPT failed to detect malicious intent in approximately 70% of cases, frequently misinterpreting offensive code generation requests as defensive procedures simply because they referenced standard frameworks like MITRE ATT&CK [34].

Similar vulnerabilities were observed by Singh and Namin [18], who warn that a significant concern in AI ethics is the potential for models to unintentionally produce malicious outputs when subjected to carefully crafted prompts that bypass safety constraints. Expanding on this, the authors conducted a scenario-based experiment to evaluate the robustness of LLMs against Cialdini’s principles of persuasion [41]. Their results indicate a high susceptibility to manipulation, with 15 scenarios achieving what they classify as “advanced, socially aware deceptions”, identifying a critical vulnerability when creating a system for sending urgent update alerts. In this highest level of engagement, characterized by “sophisticated, persuasive language” (Stage Three), models demonstrated complex, socially aware manipulations rather than simple compliance. The study revealed that prompts leveraging principles such as “Liking” (via warm, rapport-building language) and “Scarcity” (invoking artificial urgency) were particularly effective at bypassing ethical safeguards in these complex scenarios, generating sophisticated narratives explicitly crafted to establish trust and ensure victim compliance. This demonstrates that models are vulnerable not just to direct commands, but to nuanced emotional rapport and artificial urgency [18].

Another alarming development is the integration of LLMs into interactive chatbots that allow attackers to conduct personalized real-time conversations with their victims. They can dynamically adapt their responses based on user input, increasing engagement, and reducing suspicion. In addition, they can be distributed on a large scale, repeatedly, and coordinated through botnets, a capability that is further amplified when considering their high success rate [3,32]. This ability represents a break from traditional, static phishing attempts and marks the emergence of a new era of dynamic and responsive deception, as it “can supercharge deceptive campaigns, making them highly sophisticated and more challenging to identify and counter” [3].

Webb et al. note that the use of chatbots by attackers prevents them from being revealed by a language barrier by generating messages with proper spelling, grammar, and tone [12]. The enhanced contextual awareness and linguistic sophistication of AI-generated content enables the creation of highly convincing narratives that align with the victim's profile and expectations. As a result, traditional warning signs, such as grammatical errors or unnatural phrasing, are increasingly absent, making users more vulnerable to manipulation [8,11]. This linguistic quality, combined with the dynamic conversational capabilities of LLMs, represents a qualitative shift from static phishing templates to adaptive and interactive deception. Empirically validating this threat, Chrysanthou et al. [31] analyzed a simulation targeting over 25,000 users with LLM-generated content. Their findings confirm the operational viability of these attacks, demonstrating that synthetic text can effectively scale high-volume campaigns and successfully exploit psychological levers, such as authority and curiosity, to trigger compromise. On the other hand, the efficacy of these models is not uniform. Malloy et al. [14] introduce a critical nuance: their behavioral experiments reveal that “hybrid attacks”, co-created by humans and GenAI, pose a greater challenge to detection than those generated by GPT-4 or humans alone. This suggests that the current threat frontier is not purely autonomous, but rather a “centaur” model where AI amplifies human intent with scale and fluency.

To counter these sophisticated manipulations, Tsinganos et al. [17] presented a technical framework for Early-Stage Recognition (ESR), designed to identify the intent of deception during the rapport-building phase. The authors introduced a detection model that uses Convolutional Neural Networks (CNNs) and word embeddings to analyze the semantic dynamics of chat-based social engineering. By explicitly mapping text patterns to persuasion strategies (Cialdini's principles), this approach shifts the defensive focus from static keyword matching to behavioral pattern recognition, aiming to intercept the attack before the malicious payload is delivered. Although limited by the static nature of the training datasets against the infinite variability of GenAI, their model achieved 71.6% accuracy, effectively demonstrating that psychological triggers leave detectable semantic footprints that can be identified before the attack.

In summary, LLMs have catalyzed a new paradigm of SE attacks, transforming it from a manual craft of persuasion into an industrialized, automated discipline. This evolution encompasses not only the semantic refinement of dialogical deception, where linguistic barriers are eroded, but also the operational acceleration of the attack lifecycle through automated code generation and hybrid vector deployment. Their increasing sophistication and scalability require equally adaptive, interdisciplinary countermeasures involving technical safeguards, user-focused education, and regulatory oversight.

### 2.2.3. The Emergence of Generative AI and Synthetic Media

Generative AI, a branch of AI dedicated to creating human-like synthetic content, has greatly improved the potential of SE, being increasingly used for fraud, political disinformation, non-consensual imagery, and harassment, and posing a growing threat to global

information integrity, requiring urgent, coordinated action [23]. Based on algorithms capable of generating text, images, audio, and video that are virtually indistinguishable from authentic creations, its applications span from natural language generation to audiovisual production [8,16].

This capability has been crucial to the advancement of so-called synthetic media, such as deepfakes and voice cloning, which enable the creation of hyper-realistic content that simulates, with impressive fidelity, the appearance, voice, and behavior of real individuals, representing a profound transformation in the threat landscape [3,28,39].

Voice cloning, for instance, employs deep neural networks trained with short voice samples to replicate an individual's tone, rhythm, and vocal inflection. Tools such as Microsoft's Neural Codec Language Model for Text-to-Speech (VALL-E) can create precise voice simulations using as little as three seconds of audio, a technique that has already been exploited in business fraud and simulated audio-based virtual kidnappings [2,16]. These advances have fundamentally challenged the traditional voice biometrics authentication landscape [1]. Kurshan et al. highlight that the security implications are so severe that certain technology providers have opted to withhold the public release of their most sophisticated cloning tools, labeling them as 'too effective' for open distribution due to the high risk of facilitating indistinguishable deception [1].

To operationalize these capabilities, Valdez et al. [13] demonstrated the feasibility of constructing modular architectures that chain LLM outputs directly into audio synthesis pipelines. By integrating a chatbot to drive conversation strategy with a voice cloning system, their experiment confirmed that accessible AI tools can be leveraged to execute "highly convincing pretext attacks". This framework allows for the automation of rapport-building scenarios previously requiring human effort.

In parallel, deepfakes, videos produced using Generative Adversarial Network (Generative Adversarial Network (GAN)), have reached a degree of realism that allows for the precise replication of facial expressions, lip synchronization, and gestures of known individuals. GANs function via two adversarial neural networks: one, the generator, is responsible for producing artificial images, while the other, the discriminator, strives to differentiate between counterfeit and authentic items, leading to a continual enhancement process [2,9,16,40]. As noted by Swathi and Saritha [40], this evolution has been critical in overcoming early flaws, such as irregular eye blinking and unnatural head poses, making modern deepfakes capable of mimicking subtle biological signals that previously served as reliable detection indicators.

Applications such as DALL-E and Stable Diffusion broaden these possibilities by allowing the creation of realistic fake images and novel digital faces, as highlighted in recent forensic frameworks [42]. These capabilities are frequently utilized in the fabrication of fake social media accounts, romance scams, and fraud schemes within corporations [8,16]. The ease of access to these tools enables individuals with minimal or no technical skills to exploit them maliciously, effectively lowering the barrier to entry for cybercrime [2]. Empirical evidence in the literature indicates that users may easily accept and engage with artificially generated profiles, suggesting that synthetic identity cues (e.g., AI-generated faces and profile artifacts) can shift perceived legitimacy and expand the attack outreach beyond message-level plausibility [6].

A further escalation of this technology, discussed by Frankovits and Mirsky [37], is the emergence of real-time deepfakes, which enable live impersonation during voice and video calls, meaning that such channels can no longer be trusted based on content alone. Unlike offline manipulations, real-time settings introduce strong practical constraints for defenders (e.g., latency, device efficiency, compression, and cross-platform deployment), and they make purely artifact-based detection increasingly brittle against

adaptive adversaries [28,37]. Consequently, recent discussion emphasizes the need for a shift from passive forensic inspection toward “active” or out-of-band defenses that authenticate the interaction context rather than relying only on media traces [37]. Empirical validation of this vulnerability was provided by Eberl et al. [28], whose experiments on audiovisual phishing revealed that over 70% of participants were unable to detect real-time manipulated video content in simulated conference calls, “even when familiar with the impersonated individual”, confirming that current deepfake fidelity has effectively surpassed the threshold of unassisted human perception.

The implications of synthetic media are far-reaching. Deepfakes have already been used to bypass facial recognition systems, extort individuals, and manipulate evidence in legal disputes. Documented cases include entire corporate videoconferences composed of falsified avatars, resulting in multimillion-dollar fraudulent transfers. As an example, according to press reports, a case occurred in the United Kingdom where an energy company lost approximately £220,000 after an employee, deceived by a voice clone mimicking the CEO of its parent firm, approved a fraudulent transfer. (Jesse Damiani, “A Voice Deepfake Was Used To Scam A CEO Out Of \$243,000,” *Forbes*, 3 September 2019. Available at: <https://www.forbes.com/sites/jessedamiani/2019/09/03/a-voice-deepfake-was-used-to-scam-a-ceo-out-of-243000/> (accessed on 23 December 2025)) [6]. Economic indicators highlight the alarming scale of this threat, projecting global losses from generative-AI-related fraud to quadruple by 2027, maintaining a staggering annual growth rate of over 30% (e.g., from \$12.3 billion in 2023 to \$40 billion) [1,23], driven in part by the mass proliferation of deepfake technologies on the dark web, now available at extremely affordable prices (“as little as \$20”) [23].

Beyond these economic implications, Abbas and Taeihagh [35] also warn that the democratization of advanced manipulation tools has rendered the digital domain an unreliable locus for truth. Their systematic review reveals that even leading commercial defense mechanisms are easily bypassed; notably, they report that 78% of deepfake content successfully deceived industry-standard APIs, such as Microsoft Azure. This failure extends beyond biometric security to critical infrastructure, where the authors highlight the potential for falsified satellite imagery (e.g., non-existent bridges) to mislead military operations. Consequently, they argue that existing surveys, which often treat detection and generation in isolation, are insufficient, calling instead for integrated frameworks that map the interplay between evolving generation techniques and policy-driven defense.

This systemic vulnerability validates the concerns of Blauth et al. [39], who argue that malicious use of AI-generated audiovisual content threatens not only individual privacy but also the integrity of democratic, judicial, and institutional systems. In this context, deepfake security has devolved into a cycle of adversarial adaptation; as detection metrics improve, generative models are subsequently optimized to neutralize these artifacts or bypass filters through adversarial training [6]. Given that, the growing sophistication of these tools poses an urgent challenge for the development of technical, educational, and regulatory mechanisms that ensure content authenticity, source verification, and protection against manipulation; “AI-versus-AI defense appears to be the only viable solution approach moving forward”, emphasize Romero-Moreno [23].

However, translating these needs into effective countermeasures remains a challenge. The complexity of GenAI models often makes their outputs difficult to interpret, which can reduce trust in AI-assisted security judgments and hinder operational adoption [9]. At the same time, as noted by Balasubramanian et al. [9], the increasing diversity of these cyber threats, aligned with the huge volume of data generated by contemporary digital ecosystems, has reduced the effectiveness of traditional methods in anticipating and mitigating emerging attacks [9]. As a result, the literature highlights a gap in the

defensive domain itself, which emphasizes the urgency to focus research on Generative AI for threat intelligence tasks and security operations, including data integration and real-world validation [9,23,35].

Additionally, despite the frequency and impact of SE attacks, the literature also reports a shortage of practical tools and analytical frameworks capable of providing deeper insight into the tactics, techniques, and procedures (tactics, techniques, and procedures (TTPs)) of multi-stage social engineering, especially beyond purely technological artifacts [24]. This limitation is reflected in widely adopted frameworks: for example, only a small subset of techniques in MITRE ATT&CK explicitly addresses social engineering, as with the Cyber Kill Chain, both frameworks provide limited coverage of SE and offer little detail on rapport-building and iterative attacker–victim dynamics [24]. Therefore, these gaps motivate a closer look at attacker-centric frameworks that go beyond generic cyber matrices and capture the operational sequencing of multi-stage SE campaigns in GenAI-mediated settings, which will pave the way for the unified model in Section 3.

### *2.3. Analytical Frameworks for AI-Powered Social Engineering Attacks*

To further clarify how AI technologies are practically applied within SE, this section provides a comparative examination of illustrative frameworks, based on the technological landscape described in Section 2.2. These models offer different perspectives, technical, empirical, and predictive, on the mechanisms and impacts of AI-driven SE campaigns. The following analysis explores each framework in detail, highlighting their structure, methodology, and contributions to the ongoing discourse on digital deception.

#### *2.3.1. Social Engineering with Generative AI: Framework and Implications*

The Generative AI Social Engineering Framework (GenAI-SE), introduced by Schmitt and Fléchais [3], is a comprehensive technical framework created to assess how generative artificial intelligence enhances the execution and efficacy of SE attacks. This model, arranged across three primary components—namely, the creation of lifelike content, sophisticated targeting with personalization, and the automation of attack infrastructures—facilitates the breakdown and examination of AI’s role in amplifying attack efficiency and minimizing costs at scale.

Regarding lifelike content creation, the authors point out the prominent threat posed by deepfakes, underlining the escalating challenges in identifying these artificial contents. The authors mention that while CNN and Recurrent Neural Networks (RNNs) are utilized for detection, these methods encounter difficulties with generalizability and reliability. Although text generation and neural voice synthesis aren’t explicitly discussed in the article, it is reasonable to deduce that natural language models and Text-To-Speech (TTS) technologies could be employed in this context to mimic corporate emails, automated voice calls, and fabricated video content impersonating figures of authority, especially in highly persuasive spam phishing campaigns.

The advanced targeting and personalization layer is discussed in the context of leveraging large datasets to craft highly targeted campaigns. The authors note that attackers can use AI algorithms to analyze extensive data sources and generate spear phishing attacks with a high degree of personalization. Although the term OSINT is not used, it is reasonable to assume that these datasets include information obtained from social networks, public leaks, and other open sources. Linguistic adaptation techniques based on psychographic or demographic traits—though not detailed in the paper—are compatible with this objective and well documented in related literature. These personalization capabilities are presented as critical factors in the success of targeted campaigns such as spear phishing and whaling.

In the automated attack infrastructure layer, the authors discuss how the adoption of Software as a Service (SaaS) solutions, combined with the gig economy (The gig economy is a labor market characterized by the prevalence of short-term contracts or freelance work. See: Sundararajan, A. (2016). The gig economy is coming. Harvard Business Review. <https://www.stern.nyu.edu/experience-stern/faculty-research/gig-economy-coming-what-will-it-mean-work> (accessed on 10 June 2025) [43]), and the underground cybercrime ecosystem, facilitates the scaling of AI-driven attacks. They caution that even individuals with limited technical expertise can access sophisticated tools and launch large-scale campaigns. Although the technical infrastructure of such automation is not detailed, it can be inferred that it includes delivery bots, integration with communication Application Programming Interfaces (APIs) (email, SMS, VoIP), and distributed systems capable of managing multiple attack instances in parallel.

To measure these effects, the authors introduce two core indicators: Threat Amplification and Cost-Effectiveness. The former refers to the increased success rate and broader reach of attacks, driven by automation and the ability to personalize at scale. The latter concerns the reduction in time, effort, and financial resources required to carry out highly convincing campaigns. Together, these factors demonstrate that generative AI not only enhances attack effectiveness but also makes these tactics more accessible to a wider range of threat actors, including those with limited technical skill or operational capacity. As a result, the authors suggest that spam phishing benefits significantly from automated content generation, while spear phishing and whaling gain the most from personalization and automation, enabling high-precision adaptation and significantly lowering the cost per successful attack.

The study also highlights the convergence between the increasing capabilities of AI models and the decreasing costs associated with their use by adversaries. According to the authors, this phenomenon makes AI-driven attacks not only more powerful but also more accessible, even to low-skill actors—thus increasing the risk of industrialized campaigns within cybercriminal ecosystems supported by SaaS platforms and gig-based services. Regarding countermeasures, the study contends that traditional approaches focused on user training and awareness are insufficient to confront AI-shaped threats. As AI capacity to simulate human interactions increases, it becomes progressively less feasible for even trained users to distinguish legitimate communications from forged ones. The authors critique the assumption that failure to detect fraud stems solely from a lack of training, arguing that such logic unfairly blames the victim and is both circular and ineffective.

As a practical contribution, the authors propose technical and human-centric countermeasures aligned with the three dimensions of the framework. For realistic content generation, they highlight solutions such as deepfake detection using CNNs and RNNs, digital watermarking to authenticate genuine content, and tools for reversing malicious modifications. In the personalization layer, recommendations include media literacy education, privacy-enhancing technologies, and stronger data protection legislation. For the automated infrastructure dimension, strategies such as message rate limiting, multi-factor authentication, and AI-based defensive systems using behavioral analysis and automated threat detection are suggested.

In conclusion, the authors caution that even cybersecurity experts may fall victim to AI-generated phishing attacks due to their ability to manipulate psychological and contextual variables. They emphasize that expecting continuous and flawless vigilance from users is unrealistic given the complexity of modern digital environments, and that awareness-based solutions alone overlook fundamental cognitive and emotional factors. The relentless evolution of AI thus demands new technical and policy approaches to confront the growing sophistication of AI-powered SE threats.

### 2.3.2. SE Framework Powered by Generative AI

The study conducted by Alahmed et al. [8] proposes an analytical model to understand the effectiveness and impact of AI-generated content in SE attacks, with a focus on the victim's experience and the practical implications of generative AI. Although the authors do not formalize the framework as a modular diagram, its conceptual structure can be interpreted along three functional axes: advanced personification, genuine content creation, and automated attack infrastructure, coherently organizing the technical and social elements investigated throughout the study.

In the advanced targeting and personification dimension, the study reports the use of synthetic content, such as cloned voice messages and fake social media profiles with convincing images, based on the narratives collected from the interviewed victims. While the authors do not specify the generation tools used, the described techniques (e.g., voice call simulation and the use of false profile photos) are consistent with technologies such as GANs, neural voice cloning based on Tacotron, and vocoders like WaveGlow, which allow for the emulation of identity characteristics like gender, age, and vocal inflections, enhancing emotionally compelling and realistic interactions. The qualitative data collected indicate that emotional personification and the timing of the interaction were critical elements in the success of the reported attacks.

In the genuine content creation axis, the research shows that the malicious content received by participants (including emails, Telegram, and WhatsApp messages) contained typical elements of emotional persuasion—such as fear induction, urgency cues, and institutional language. Although the technical process behind the generation of such content is not detailed, the attack descriptions suggest the use of language models adjusted to specific contexts. By technical analogy, it is plausible to consider the use of LLMs with few-shot learning and corpora extracted from leaked emails to generate highly contextualized and pragmatic communications. Participants reported difficulty distinguishing fake from legitimate messages, especially when the content included plausible argumentative structure and domain-specific vocabulary.

The automated attack infrastructure dimension was explored through the development of a chatbot that relied on a dataset of 700 spam messages to simulate common social engineering interaction patterns. The study used this setup to conduct usability testing with participants across platforms such as Telegram and Facebook. Although designed for research purposes, the chatbot's architecture reflects capabilities that could also be leveraged in offensive contexts. Thus, its description indirectly serves as a reference model for understanding how automated agents could organize large-scale SE attacks.

The empirical findings of the study were derived using two methods: semi-structured interviews with 40 victims of SE attacks and usability testing with 40 users who interacted with the developed chatbot. The qualitative analysis followed the grounded theory methodology, involving open, axial, and selective coding, with the support of the NVivo-12 software. The data were organized into 33 open codes, subsequently grouped into six categories that provided an empirical taxonomy of AI-mediated SE mechanisms, with an emphasis on users' subjective experience and the perceived authenticity of the content.

In the usability experiment, the authors applied the System Usability Scale (SUS) to measure the chatbot's effectiveness and user acceptance in simulating and detecting attacks. The results revealed significant variations between occupational groups: unemployed users showed the highest levels of trust and acceptance towards AI-generated content, which the authors link to lower exposure and expertise in social engineering, while employed users demonstrated greater resistance, reflecting their familiarity with cyber threats, which tends to improve critical awareness and skepticism about suspicious messages. Students

displayed an intermediate level of acceptance, positioned between the unemployed and employed groups.

Overall, the study shows that generative AI-based attacks exploit both technical and psychological vulnerabilities. The authors emphasize that emotional personalization, the absence of tailored preventive measures, and the difficulty in detecting synthetic content were decisive factors in the success of the attacks. They recommend the development of AI-based defensive tools, ongoing educational campaigns, and user-centered approaches as means to mitigate emerging risks.

In summary, the investigative model of Alahmed et al. [8] contributes to the field by integrating qualitative and quantitative methods to examine, from the victim's perspective, the real-world impact of AI techniques in SE. Their approach underscores the importance of understanding the human factors involved, offering a robust empirical foundation for the development of countermeasures that are responsive to user behavior and the operational strategies of attackers in AI-mediated environments.

#### 2.4. The Influence of AI on SE

The joint analysis of the cited studies reveals a significant methodological convergence in how AI is integrated into SE attack models. Despite formal differences in the proposed frameworks, a clear structural correspondence emerges among the core components of each approach. This convergence centers on three elements: realistic content generation, contextual message personalization, and automated or scalable execution.

Analyzing the evolutionary trajectory of AI-enabled crime as categorized by Blauth et al. [39], it reveals a shift from simple automation to highly sophisticated, adaptive attacks, which is aligned with the cited frameworks. To visualize this, we structure this evolution into three developmental phases that mirror the functional dimensions identified by Schmitt and Fléchais [3] and Alahmed et al. [8]:

- Phase 1: Corresponding to the initial industrialization of cybercrime, this phase focuses on expanding the attack surface through digitization and task automation (e.g., bulk email distribution). As noted by Blauth et al. [39], this reduces the marginal cost of reaching targets.
- Phase 2: Contextual Personalization: Advancing beyond generic attacks, this phase leverages ML and LLM to craft highly tailored narratives that help attackers transition victims from a neutral state to a trusted state more efficiently.
- Phase 3: Adaptive Synthesis: The current frontier involves the use of autonomous agents capable of adapting strategies in real-time. This phase uses synthetic media (deepfakes) to bypass skepticism, representing the emerging threat landscape highlighted in recent AI crime surveys.

This evolutionary trajectory is theoretically substantiated by Liu et al. [11], who propose the “Scale-Personalization Conflict Framework”. They argue that the historical development of social engineering has been driven by a persistent trade-off between attack volume and tailored effectiveness. The advent of GenAIs marks a revolutionary turning point by resolving this conflict, enabling, for the first time, high-precision personalization at a scalable low marginal cost. While Blauth et al. [39] study provides the temporal context, identifying the shift from simple automation to adaptive synthesis, confirming that the threat landscape is not static but follows a predictable trajectory of increasing complexity, along with the frameworks analyzed, they contribute to a distinct methodological lens that directs us toward a unified understanding. Since Schmitt and Fléchais [3] offer us a systematic decomposition of these attacks, highlighting how generative AI amplifies specific attack vectors, and Alahmed et al. [8] complement the technical view with a user-

centered empirical approach, emphasizing the perceptual and emotional mechanisms, such as the perception of empathy and authority, that increase the success rate of these attacks.

This triangulation reflects a consolidated trend in the literature: AI-mediated SE attacks are increasingly modeled as complex sociotechnical systems whose effectiveness derives from the synergy between linguistic sophistication, victim-specific adaptation, and large-scale orchestration. It also confirms that advances in AI not only increase the persuasive power of malicious content, but also reduce technical and operational barriers for attackers across different skill levels.

In addition, these studies underscore that traditional mitigation approaches are insufficient in the face of highly personalized and contextually relevant attacks. The need for integrated countermeasures that combine behavior-based automated detection, restrictions on personal data exposure, and explainable algorithmic defenses is consistently emphasized across the reviewed literature. Furthermore, as highlighted by Qin et al., since “approximately 95% of security incidents are attributed to human errors”, recent work also models awareness and training as a cost-effective policy instrument, proposing structured approaches to maximize risk reduction under budget constraints rather than treating training as a generic checklist measure [26]. Complementarily, the authors propose using GenAI defensively to generate realistic synthetic social engineering scenarios for exercises and awareness programs, enabling scalable simulation aligned with evolving attacker narratives [26].

These perspectives indicate a profound reshaping of the threat landscape. The boundary between human interaction and automated action is increasingly blurred. To address this complexity, we identify the need for a Unified Model that integrates these dimensions (Realism, Personalization, Automation) and quantifies their impact on attacker behavior, because despite the contributions, three critical gaps remain unaddressed in the current literature, which this study aims to fill:

1. The Operational Gap: While existing studies describe “what” tools are used and “who” is targeted [3,8,39], they lack a granular explanation of “how” these attacks unfold. Our model bridges this by mapping AI affordances directly onto the attack lifecycle phases (Reconnaissance, Engagement, and Exploitation) integrated into a Markov Decision Process (MDP) flow.
2. The Quantitative Proof Gap: Although frameworks mention “cost-effectiveness” primarily in a qualitative manner [3], there is a lack of mathematical proof to quantify it. Given that, we address this through an Analytical Demonstration using the Bellman Equation, with conservative parameters to prove that even modest AI enhancements can flip an attacker’s rational decision from “Do Not Attack” to “Attack”.
3. The Attacker Rationality Gap: Previous models often treat the attacker as a passive or generic entity [3]. Echoing Liu et al.’s call to analyze the “economic drivers” [11], our model introduces the attacker as a rational economic agent. By identifying the automation plateau ( $\sigma > 4$ ), we demonstrate a strategic shift where rational actors prioritize quality (Realism) over quantity (Volume)—a trade-off not deeply explored in the analyzed studies.

Ultimately, the impact of AI on SE extends beyond technical considerations, touching on ethical, cognitive, and regulatory domains. The ability of AI to manipulate perceptions, emotions, and judgments on an emerging industrial scale, as evidenced by LLM-powered botnets (e.g., fox8) that show coordinated and scalable malicious activity [32], requires a repositioning of cybersecurity policies, with particular emphasis on developer accountability, AI model governance, algorithmic transparency, and effective security awareness training.

### 3. Toward a Unified Model of AI-Mediated Social Engineering

To address the operational, quantitative, and strategic gaps identified in the previous section, we propose the Unified Model of AI-mediated Social Engineering (UM-AISE). This framework unifies the three core AI-driven affordances into a structured system that maps technical capabilities directly to the stages of an attack. By integrating these dimensions, the model moves beyond the descriptive nature of the existing literature to provide a mechanistic and measurable understanding of how AI reshapes the attacker's decision-making process.

#### 3.1. Conceptual Architecture

The UM-AISE conceptualizes AI-driven deception as a triadic system where each dimension amplifies a specific phase of the attack, addressing the operational gap by defining the “how” of the execution:

- **Realism ( $\rho$ ):** Refers to the ability of generative systems (e.g., LLMs, GANs, TTS, vocoders) to create lifelike and multimodal content. It enhances the engagement phase, increasing trust, emotional resonance, and message plausibility.
- **Personalization ( $\theta$ ):** Encompasses psycholinguistic profiling, behavioral analytics, and OSINT-based data fusion. It dominates the reconnaissance phase, improving target selection and tailoring persuasion vectors.
- **Automation ( $\sigma$ ):** Represents the orchestration of AI agents, SaaS infrastructures, and botnets that support scalable delivery and adaptive iteration. It optimizes the exploitation phase, reducing cost and operational latency while increasing campaign throughput.

These three components interact synergistically, defining a multidimensional space in which any AI-mediated attack can be positioned. The intersection of high realism, deep personalization, and full automation represents the highest-risk zone, characterized by agentic, self-optimizing attacks capable of continuous adaptation as illustrated in Figure 2.

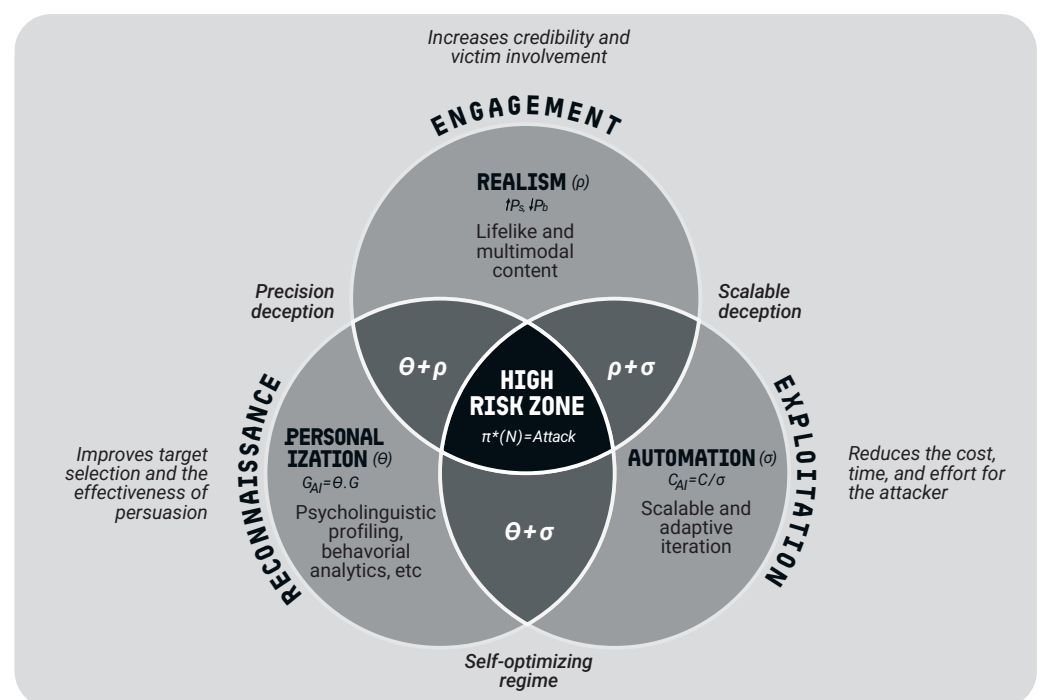


**Figure 2.** Unified Model of AI-mediated Social Engineering (UM-AISE).

The framework maps the core dimensions ( $\theta$ ,  $\sigma$ ,  $\rho$ ) to their respective impacts on the attack's economic variables: AI-augmented gain ( $G_{AI} = \theta G$ ), optimized operational

cost ( $C_{AI} = C/\sigma$ ), and success/blocking probabilities ( $p_s, p_b$ ), which collectively shift the optimal policy ( $\pi^*(N)$ ) toward the High-Risk Zone.

### 3.2. Analytical Demonstration: The Quantitative MDP Approach

To examine how the three core affordances of UM-AISE, the “High-Risk Zone”, reshape attacker incentives, we integrate the quantitative risk assessment framework by Abri et al. [38], modeling the attacker’s decision-making as a Markov Decision Process (MDP). Rather than empirically validating the model with live attack data (which would require field data that the literature does not yet provide), we apply an analytical MDP approach showing how, under theoretical assumptions of rational behavior, plausible perturbations induced by AI can shift the attacker’s optimal policy.

Formally, the attacker–victim interaction is modeled as a tuple  $\mathcal{M} = (S, A, P, R, \gamma)$  [38], where:

- States ( $S$ ): The set of possible states the attacker occupies relative to the victim, defined as:  $S = \{Neutral, Trusted, Challenged, Blocked\}$ .
- Actions ( $A$ ): The set of actions available to the attacker, defined as:  $A = \{Cooperative, Deceptive, Reset\}$ .
- Transitions ( $P$ ):  $P(s'|s, a)$  is the probability of moving to state  $s'$  from state  $s$  after taking action  $a$ .
- Rewards ( $R$ ):  $R(s'|s, a)$  is the net reward received by the attacker. In the simplest form, it can be expressed as the *Gross Gain* ( $G$ ) (value of access/data) minus the *Operational Cost* ( $C$ ) of the action.
- Discount Factor ( $\gamma$ ): Represents the attacker’s preference for immediate versus future rewards.

Following the formalism established by Abri et al. [38], the States ( $S$ ) and Actions ( $A$ ) are defined as follows:

**Neutral ( $N$ ):** The initial state where no prior interaction exists. The victim has not yet formed an opinion or suspicion.

**Trusted ( $T$ ):** A state of successful deception where the victim perceives the attacker as a legitimate entity, facilitating the exploitation of resources or information.

**Challenged ( $C$ ):** A state of heightened skepticism. The victim suspects foul play but has not yet taken defensive action, allowing the attacker a narrow window to salvage or abandon the operation.

**Blocked ( $B$ ):** An absorbing terminal state where the victim has identified the threat and terminated all communication, resulting in zero further gains for the current session.

**Cooperative:** Legitimate interaction used primarily to build rapport or maintain the current state without immediate malicious intent.

**Deceptive:** The core offensive action (e.g., sending a phishing link or a deepfake audio request). It carries the highest potential reward but also the highest risk of transition to the *Blocked* state.

over from the *Neutral* state, incurring a repositioning cost ( $C_R$ ).

The interaction between these elements is governed by the attacker’s objective to find an optimal policy  $\pi^*$ , which maximizes the expected cumulative reward. This is formally calculated using the Bellman optimality equation [38]:

$$V^*(s) = \max_{a \in A} \underbrace{\sum_{s' \in S} P(s, a, s') [R(s, a, s') + \gamma V^*(s')]}_{\text{Expected Action Value (Q)}}. \quad (1)$$

To interpret this equation within the context of an attack decision, we have:

- The summation ( $\Sigma$ ): Calculates the expected utility of executing a specific action (e.g., “Deceptive”) across all possible outcomes.
- The maximization (max) captures the attacker’s rational choice among competing actions (e.g., Deceptive vs. Reset).
- The resulting  $V^*$  represents the best achievable long-term value at state  $s$ , and  $\pi^*$  is the action that attains it.

Transitioning from global theory to our proposed strategic scenarios, we define a decision threshold based on the attacker’s entry-level incentives. This allows us to isolate how specific AI-induced changes influence the initial choice to engage. To this end, we evaluate two distinct operational settings: (i) the Manual Baseline, which represents the traditional, human-led social engineering environment, serving as a control scenario to demonstrate why attackers are often deterred under high operational costs; and (ii) the AI-Augmented Scenario, which integrates the UM-AISE affordances to show how even conservative gains in automation and realism can flip the optimal policy toward aggressive engagement.

This dual-scenario approach is critical to quantify the “strategic tipping point” where AI transforms a non-viable criminal enterprise into an incentive-compatible one.

### 3.2.1. Decision Condition (Attack vs. Do Not Attack)

To make the analysis transparent, we focus on the key decision at the *Neutral* state: attacking is optimal if the expected action-value of deception exceeds the best non-attack alternative:

$$Q(N, \text{Deceptive}) > Q(N, \text{Reset}). \quad (2)$$

### Baseline (Manual) One-Step Stylization

Consistent with Abri et al.’s state semantics, from  $N$  under Deceptive, we consider a compact transition approximation:

$N \rightarrow T$  with probability  $p_s$ ,  
 $N \rightarrow B$  with probability  $p_b$ ,  
 and return to  $N$  with probability  $1 - p_s - p_b$ .

Then:

$$Q_{\text{manual}}(N, \text{Deceptive}) = p_s \left[ (G - C) + \gamma V(T) \right] + p_b \left[ (-C) + \gamma V(B) \right] + (1 - p_s - p_b) \left[ (-C) + \gamma V(N) \right] \quad (3)$$

and

$$Q(N, \text{Reset}) = -C_R + \gamma V(N), \quad (4)$$

where  $C_R$  is the operational cost of resetting/repositioning (For the purpose of this comparative instantiation, we normalize the continuation value  $V(N) = 0$  to isolate the impact of immediate action costs and future state rewards.).

### UM-AISE as Structured Perturbations

UM-AISE affects the incentive structure by modulating (i) operational cost, (ii) expected gross gain per target, and (iii) state transition probabilities:

1. Automation ( $\sigma \geq 1$ ): reduces marginal operational cost:

$$C_{\text{AI}} = \frac{C}{\sigma}. \quad (5)$$

2. Personalization ( $\theta \geq 1$ ) increases expected gross gain per target:

$$G_{AI} = \theta G. \quad (6)$$

3. Realism ( $\rho \geq 0$ ): shifts transition probabilities toward  $T$  and away from  $B$  (rather than assuming a full inversion):

$$p_s^{AI} = p_s + \rho \Delta_s, \quad p_b^{AI} = p_b - \rho \Delta_b, \quad (7)$$

with  $\Delta_s, \Delta_b > 0$  denoting the maximum plausible shifts under improved realism.

Substituting these terms yields an AI-augmented action-value:

$$Q_{AI}(N, \text{Deceptive}) = Q_{\text{manual}}(N, \text{Deceptive}) \text{ with } (G, C, p_s, p_b) \rightarrow (\theta G, C/\sigma, p_s^{AI}, p_b^{AI}). \quad (8)$$

This makes explicit that UM-AISE can flip the optimal policy by moving the inequality in (2) from false to true under conservative (non-extreme) parameter shifts.

### 3.2.2. Model Calibration and Numerical Illustration

To demonstrate the analytical capabilities of the UM-AISE, this section instantiates the proposed MDP to quantify the structural shift in the attacker's optimal policy  $\pi^*$ . By transitioning from a traditional baseline to an AI-augmented environment, we aim to identify the specific thresholds where deception becomes the dominant strategy within the High-Risk Zone.

#### Parameter Calibration

To ground these variables in real-world scenarios and provide a concrete basis for the model's calibration, we propose the operational scales detailed in Table 5. This categorization allows for the classification of any AI-mediated attack based on its technical sophistication and strategic objective, while the subsequent MDP demonstration of AI-mediated perturbations instantiates a deliberately conservative point within this broader space.

**Table 5.** Operational Scaling of UM-AISE Dimensions.

| Dimension                    | Value/Range | Technical Description                                     |
|------------------------------|-------------|-----------------------------------------------------------|
| Realism ( $\rho$ )           | 0.0–0.3     | Generic text, obvious linguistic errors.                  |
|                              | 0.4–0.6     | Coherent content, no visual/vocal personalization.        |
|                              | 0.7–0.8     | Cloned audio or high-quality synthetic imagery.           |
|                              | 0.9–1.0     | Real-time, indistinguishable deepfake video.              |
| Personalization ( $\theta$ ) | 1.00–1.09   | Mass spam (basic name/email personalization).             |
|                              | 1.10–1.19   | Publicly available data (company, job title).             |
|                              | 1.20–1.29   | Specific context (active projects, professional circles). |
|                              | 1.30–1.40   | Privileged information (internal communications/whaling). |
| Automation ( $\sigma$ )      | 1           | Single manual attack.                                     |
|                              | 2–5         | Semi-automated campaigns (Strategic Pivot Zone).          |
|                              | 6–10+       | Botnet orchestration and Crime-as-a-Service (CaaS).       |

The parameter ranges adopted in the UM-AISE are calibrated based on stylized facts derived from the reviewed literature to ensure face validity. The Realism index ( $\rho \in [0, 1]$ ) acts as a normalized probability of deception success, mapping the human perception findings from Alahmed et al. [8] into the model's transition probabilities ( $p_s, p_b$ ). The Personalization multiplier ( $\theta \in [1.0, 1.4]$ ) represents a conservative 40% increase in expected gross gain ( $G$ ) due to high-precision targeting, such as spear phishing and whaling, as discussed by Schmitt and Fléchais [3]. Lastly, the Automation scale ( $\sigma \in [1, 10]$ ) captures

the initial exponential decay of operational costs ( $C/\sigma$ ), identifying the “Strategic Pivot Zone” where the marginal utility of volume is surpassed by the utility of content quality.

Following the operational scales defined in Table 5, we now instantiate the MDP parameters to examine the transition from manual to AI-augmented social engineering. We preserve the qualitative transition schema of Abri et al. [38] and adopt conservative numerical ranges consistent with their manual-attack baseline. This calibration ensures that the observed policy shifts result from structural AI-mediated affordances rather than aggressive parameter assumptions.

### Manual Baseline Scenario

Using Abri et al.’s structure [38], we first consider the baseline parameters, which correspond to the lowest tier of our scales ( $\sigma = 1$ ):

$$p_s = 0.15, \quad p_b = 0.30, \quad G = 10, \quad C = 5, \quad C_R = 1, \quad V_T = +100, \quad V_B = -50, \quad \gamma = 0.9.$$

Substituting these values into Equations (3)–(4), we obtain:

$$Q_{\text{manual}}(N, \text{Deceptive}) < Q(N, \text{Reset}),$$

which implies that the optimal policy under manual social engineering conditions is:

$$\pi^*(N) = \text{Do Not Attack}.$$

This result reproduces the core insight of Abri et al. [38]: when operational costs are high and success probabilities are modest, rational attackers are often deterred from engaging in deception.

### AI-Augmented Scenario (Conservative Perturbations)

We next introduce AI-mediated perturbations consistent with current empirical evidence on automation, personalization, and realism, while deliberately avoiding extreme assumptions. Although recent literature broadly characterizes Generative AI as a significant force multiplier, enhancing both the operational scale and the deceptive quality of social engineering [3], precise quantification of these gains remains an open research challenge. Consequently, the coefficients adopted here do not attempt to model the maximum capabilities of state-of-the-art models. Instead, they represent a “conservative lower bound”.

Specifically, we apply:

$$\begin{aligned} \sigma &= 3 && \text{(moderate automation-driven cost reduction),} \\ \theta &= 1.30 && \text{(30\% increase in expected gain via personalization),} \\ \Delta_s &= 0.10, \quad \Delta_b = 0.05, \quad \rho = 1. \end{aligned}$$

These values yield:

$$p_s^{\text{AI}} = 0.25, \quad p_b^{\text{AI}} = 0.25, \quad C_{\text{AI}} \approx 1.67, \quad G_{\text{AI}} = 13.$$

Substituting into Equation (8), we obtain:

$$Q_{\text{AI}}(N, \text{Deceptive}) > Q(N, \text{Reset}),$$

implying that the optimal policy shifts to:

$$\pi^*(N) = \text{Attack}.$$

By demonstrating that the optimal policy inverts even under these modest perturbations, UM-AISE moves the attacker into the High-Risk Zone, where deception becomes incentive-compatible. This decision flow is synthesized in Figure 3.

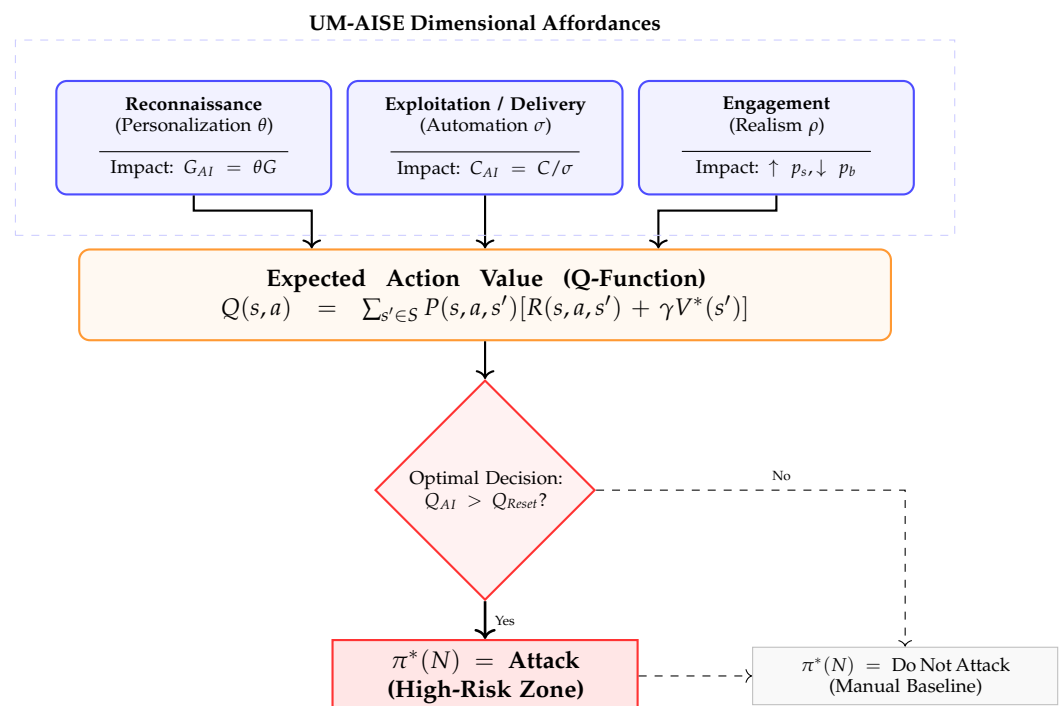


**Figure 3.** Analytical flowchart of the UM-AISE: AI affordances modulate MDP parameters, leading to the inversion of the attacker’s optimal policy.

### 3.2.3. Sensitivity Analysis

To provide a complete view of the model’s strategic landscape, Figure 4 presents the unified decision space of the UM-AISE model, mapping the strategic interaction between automation ( $\sigma$ ), personalization ( $\theta$ ) and realism ( $\rho$ ), and their impact on the attacker’s decision utility ( $Q_{AI} - Q_{Reset}$ ).

The High-Risk Zone (red area) represents the parameter space where AI-driven affordances make deception the attacker’s optimal policy. Conversely, the Manual Baseline (blue area) indicates the region where operational costs and detection risks outweigh potential gains, leading to an “Do Not Attack” decision. The dashed “Decision Frontier” marks the critical threshold where AI convergence shifts the attacker’s rational choice.

Notably, the visual plateau observed in the automation axis ( $\sigma > 4$ ) reflects the law of diminishing marginal returns inherent to the cost function  $C/\sigma$ . This mathematical property suggests a clear “quality vs. quantity” trade-off. The visual gradient of the heatmap shows that the decision utility increases much more rapidly along the horizontal axis (AI Quality: Realism and Personalization) than along the vertical axis (Automation Scale). This implies that once a baseline of efficiency is met, it is strategically more advantageous for a rational attacker to invest in increasing the realism of a deepfake or the precision of a psycholinguistic profile than to simply increase the volume of messages sent.

In summary, the UM-AISE model demonstrates that the convergence of these dimensions does not merely add new tools to the attacker’s arsenal but fundamentally reshapes the risk landscape, where AI mediates a self-optimizing environment where deceptive actions become the default optimal policy  $\pi^*$ .

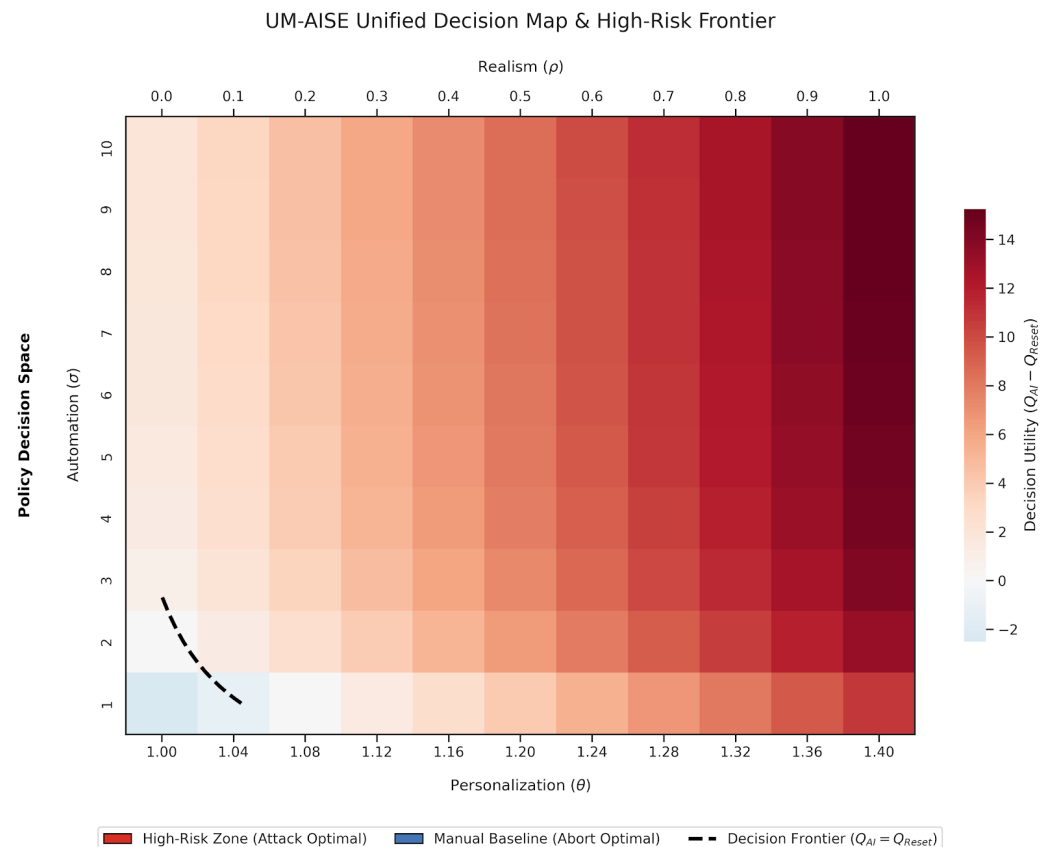


**Figure 4.** Heatmap of the UM-AISE Decision Space. The gradient represents the attacker’s Decision Utility ( $Q_{AI} - Q_{Reset}$ ). The red area (High-Risk Zone) illustrates the region where the attack becomes economically profitable ( $Q_{AI} > Q_{Reset}$ ) even under the conservative parameter. The dashed line marks the strategic “Decision Frontier,” indicating the tipping point where AI affordances shift the optimal policy from ‘Do Not Attack’ (blue region) to ‘Attack’.

However, we emphasize that this analysis was intentionally stylized. It aggregates multi-stage attacks into a single decision period, treats UM-AISE affordances as separable dimensions, and does not model adaptive defenders or multi-agent botnets. The numerical values are illustrative and grounded in the literature, but not derived from field-measured attack telemetry. Consequently, the contribution is not empirical validation, but a theoretically grounded demonstration that AI-mediated affordances can shift rational attacker behavior under conservative assumptions.

#### 4. Constraints and Ethical Implications

The growing use of AI in digital environments extends beyond technical and operational vulnerabilities, broadening the scope of unresolved ethical and legal dilemmas. In this context, recent literature identifies three primary dimensions in which these risks materialize: the opacity of AI systems and lack of accountability, the large-scale manipulation of human perception, and the instrumentalization of AI in crimes that are difficult to trace.

A significant portion of AI systems based on supervised ML and deep neural networks operate as “black boxes,” generating inferences without offering transparent explanations for their decisions [16,42]. This issue is strongly emphasized by [39,42], who identify algorithmic opacity as a first-order ethical challenge, as it obstructs audits of system errors, biases, or embedded abuse. According to the authors, this lack of transparency can be deliberately exploited by malicious actors who manipulate decision-making systems or embed harmful instructions in ways that are practically undetectable. Furthermore, the absence of clear mechanisms for attributing responsibility in automated decision-making

creates a legal void that fosters impunity—particularly in contexts involving profiling, behavioral surveillance, or the processing of sensitive data.

The ethics of authenticity is also being deeply affected by synthetic content generation technologies. According to [16], the capacity of generative architectures to produce audio and video content that is virtually indistinguishable from real recordings makes these tools powerful instruments for fraud, blackmail, disinformation, and organizational sabotage. The authors emphasize that such technologies allow for the creation of highly convincing digital evidence, capable of influencing judicial decisions, deceiving financial systems, or damaging the reputations of individuals and institutions. The situation is further exacerbated by the high false-negative rates of current deepfake detection systems, fueling a vicious cycle in which the malicious use of AI advances faster than available countermeasures.

Furthermore, the integrity of defensive operations itself is at risk, as demonstrated by Huang et al. [30]. The authors warn of the emergence of AI-generated fake Cyber Threat Intelligence (CTI), where LLMs can mimic the precise technical jargon and structure of legitimate security advisories to generate synthetic reports that evade detection by state-of-the-art classifiers. By flooding intelligence channels with these plausible but fabricated indicators, attackers can execute “data poisoning” attacks, effectively blinding security analysts and causing automated defense systems to misclassify threats. This creates an environment where even the “truth” used for defense becomes a vector of manipulation, undermining the foundational trust required for collaborative cybersecurity.

In the realm of biometrics and authentication, ref. [2] demonstrate that AI-based voice cloning technologies can replicate the tone, cadence, and rhythm of an individual’s voice using just a few seconds of audio. This enables the generation of highly convincing fraudulent voice commands. The authors report cases in which companies were deceived into authorizing bank transfers based on cloned executive voices. Beyond direct financial risks, such fabricated voices have been used in fake emergency scenarios, emotionally pressing victims to act under duress. These practices not only exploit weaknesses in biometric systems but also undermine public trust in voice-based authentication, highlighting the urgent need for more resilient and adaptive verification mechanisms. Consequently, Eberl et al. [28] argue that technical countermeasures alone are insufficient to address these biometric threats. Their analysis advocates for a hybrid framework where legal statutes must be explicitly updated to criminalize identity manipulation, acknowledging that algorithmic detection cannot keep pace with generative evolution.

Another critical concern involves the misuse of personal data for offensive purposes, particularly in the context of AI-driven SE attacks. As pointed out by Anwar and Perez [19], AI algorithms have been employed to collect and process vast amounts of data, often without individuals’ awareness or consent, to build highly accurate psychosocial profiles. These profiles are subsequently used to tailor attacks such as phishing campaigns, emotional manipulation, or behavioral fraud. The violation of informational self-determination is clear, especially given that the data in question may be sourced from social media, public records, or corporate leaks, and processed without any meaningful governance structure. Algorithmic segmentation of users based on sensitive characteristics such as ethnicity, political orientation, or economic status opens the door to discriminatory practices, political manipulation, or targeted coercion, further reinforcing the power asymmetries between technology operators and affected individuals. This erosion of privacy extends to the corporate sphere, described by Diro et al. [25] as a complex landscape where GenAI systems exhibit a “dual nature”, acting simultaneously as potential victims of exploitation and perpetrators of security breaches. The authors warn that the unmanaged integration of GenAI in workplaces leads employees to “inadvertently disclose sensitive information”

into public models. Consequently, the reactive implementation of AI-driven workplace monitoring to mitigate these risks creates a critical tension, where the pursuit of data security through surveillance risks violating the very employee privacy it aims to protect.

Therefore, the ethical challenges associated with AI do not lie solely in its direct use for generating fraudulent content but also in the systemic structures that facilitate opaque, abusive, and unaccountable practices. What emerges is the outline of a digital ecosystem in which automated decisions, hyper-realistic simulations, and hidden inferences become ubiquitous, undermining core guarantees such as privacy, information authenticity, and individual autonomy. Addressing these challenges demands more than technical fixes; it requires the development of a comprehensive regulatory framework capable of responding to the complex moral and legal dilemmas posed by AI in high-risk environments.

## 5. Research Questions Answered

Based on the systematic analysis performed in Sections 2.2 and 2.3, and the ethical synthesis in Section 4, this section presents the answers to the research questions. The findings are integrated through the lens of the proposed UM-AISE (Section 3), and the identified research gaps.

- RQ1: What are the primary methods by which AI is utilized to carry out SE attacks? Our review identifies that AI is not merely adding methods and new tools but converging distinct technological vectors that map directly onto critical phases of the Cyber Kill Chain lifecycle. Within the logic of our proposed UM-AISE, these methods serve as the operational drivers for the model's core variables, operating through three consolidated mechanisms: (i) *Automated profiling*, used by attackers to scrape and analyze OSINT data, identifying high-value targets based on behavioral patterns (Reconnaissance) [8]; (ii) *Generative mimicry*, using GANs to synthesize hyper-realistic deepfakes that bypass sensory verification (Engagement) [16,40]; and finally (iii) *Contextual text generation*, using LLMs to power malicious chatbots and automate dynamic narratives at scale (Delivery/Exploitation) [3,36]. This convergence enables the transition from generic “spray-and-pray” tactics to precision-targeted campaigns fueled by the industrialized abuse of AI ecosystems, including the systematic exploitation of legitimate AI infrastructures (e.g., marketplaces and APIs) [22] and the deployment of commoditized “dark” services (e.g., Jailbreaking-as-a-Service) [15], which serve as the operational backbone for the attack, extending capabilities far beyond simple content generation.

Furthermore, the emergence of fully autonomous agentic chains [29] do not simply generate content but actively navigate the reconnaissance-exploitation loop without human intervention, effectively closing the operational gap between intent and execution.

Within the context of our research, these mechanisms provide the operational infrastructure that fuels the UM-AISE dimensions (Personalization, Realism, and Automation) whose specific impact on attack efficacy is analyzed in RQ2.

- RQ2: How do LLMs, deepfake technologies, and generative AI tools contribute to the efficacy of SE campaigns?

The analysis indicates that these technologies contribute to efficacy fundamentally by altering the economic and operational logic of attacks. As formalized in our UM-AISE framework (Section 3), AI tools amplify efficacy by perturbing specific variables in the attacker's decision process:

- Realism ( $\rho$ ): Generative tools reduce the “suspicion trigger” in victims by eliminating artifacts (e.g., unnatural voice) that typically signal fraud. This modifies the state transition probabilities, increasing the likelihood of reaching the *Trusted* state ( $p_s^{AI}$ ) while minimizing the risk of being *Blocked* ( $p_b^{AI}$ ).
- Personalization ( $\theta$ ): LLMs enable the mass-customization of narratives based on victim profiling, acting as a multiplier on the attack’s utility ( $\theta G$ ), significantly increasing the Expected Gross Gain per interaction compared to manual engineering.
- Automation ( $\sigma$ ): This dimension transcends simple task repetition to encompass the orchestration of agentic workflows and supply chain capabilities (as identified in RQ1). By leveraging industrialized infrastructure (e.g., MaaS), attackers drastically reduce operational latency and effort. This functions as a cost divisor ( $C/\sigma$ ), creating a scenario in which, as demonstrated in our MDP analysis, the incentive to attack becomes positive (the optimal policy  $\pi^*(N) = \text{Attack}$ ) even with moderate probability of success.

Crucially, our analysis confirms that AI resolves the historical “Scale-Personalization Conflict” of Liu et al. [11]. Furthermore, the MDP model demonstrates that efficacy is not merely about volume; the “Automation Plateau” indicates that once operational costs are minimized ( $\sigma > 4$ ), the strategic advantage shifts decisively toward hyper-realism ( $\rho$ ), making quality of deception the new dominant variable over pure quantity.

- RQ3: What are the primary psychological and technological vulnerabilities exploited by AI-enabled SE tactics?

Our analysis reveals that psychological and technological vulnerabilities have ceased to be isolated vectors, converging instead on the *systemic exploitation of trust* through hyper-realistic human mimicry.

The primary psychological vulnerability lies in the human inability to distinguish AI-generated fidelity from reality, a phenomenon identified as the exploitation of “Impostor Bias”. In this scenario, the technical perfection of deepfakes and voice cloning neutralizes the victim’s natural skepticism, leading to a fundamental trust calibration failure. Victims do not merely fail to detect artifacts, but actively misclassify synthetic identities as legitimate [6], a vulnerability exacerbated by the cognitive load of real-time synthetic interactions where immediacy precludes verification [28]. This exploitation is further amplified by *Personalization* ( $\theta$ ), as defined in our UM-AISE model, which leverages the vast availability of real-world data from social networks and OSINT to craft narratives so precise they eliminate traditional “suspicion triggers.” Technologically, this convergence can effectively bypass the practical protections expected from Zero Trust deployments. Specifically, the vulnerability extends to the collapse of biometric trust surfaces: voice and facial recognition systems, previously relied upon as immutable proofs of identity, are now susceptible to digital injection attacks. This renders standard ‘liveness detection’ increasingly brittle against generative mimicry, effectively bypassing technical filters designed to intercept automated scripts [3,4,36]. Ultimately, as formalized in our framework, this synergy between truthful data and ultra-realistic mimicry drastically reduces the probability of the attack being blocked ( $p_b^{AI}$ ), transforming *trust*, once a pillar of social cohesion, into a dynamic risk variable exploited by modern adversaries.

- RQ4: What defense mechanisms and mitigation strategies have been proposed?

Our synthesis indicates that defending against AI-enabled SE requires the *strategic amplification* of existing defensive capabilities, evolving from static automation to a dynamic “AI-on-AI” counter-strategy. Since the primary vulnerability is the systemic exploitation of human trust (as identified in RQ3), and given that our UM-AISE/MDP

model demonstrates how specific variables dictate the attacker's decision, we propose that effective mitigation must utilize the adversary's own tools to overcome the limitations of manual oversight and legacy protocols. Instead of relying on analysts to track threats using standard tools, this approach provides an *Augmented Intelligence* layer: using AI to automate the heavy lifting of detection and artifact analysis, thus freeing human to focus on high-level contextual judgments.

First, to counteract Realism ( $\rho$ ), we propose a defensive focus on *Detection and Attribution*. This involves employing AI-powered systems to identify consistent patterns resulting from the model architecture, code, training data, and operational parameters (fingerprints). This field of Digital Image Forensics (DIF) applies scientific techniques and deep learning algorithms to determine the veracity of digital media, serving as a crucial line of defense by verifying the authenticity of synthetic content [44,45], a task that has become impossible for unassisted human perception. However, it faces limitations in real-time scenarios. Therefore, early-stage persuasion detection [17] and Out-of-Band (OOB) authentication [37] have shown promise. By using NLP to identify linguistic patterns of rapport-building (intent) rather than just technical payloads, and validating interaction context via secondary channels, defenders can intervene during the *Engagement* phase, effectively increasing the blocking probability ( $p_b$ ) before trust is established.

To mitigate *Personalization* ( $\theta$ ), our analysis proposes a strategic shift. First, strategies focused solely on *Data Privacy and Information Protection*, including stricter regulations to limit OSINT availability, face inherent limitations. Although existing regulations and directives are sophisticated, the challenge remains bridging the gap between "paper compliance" and actual technical enforcement to limit data sprawl. At the ecosystem level, standardization and cross-sector collaboration are more effective than isolated mandates. A unified operational front, spanning media, civil society and industry, is crucial to creating a dynamic and adaptive ecosystem capable of disrupting financial incentives and closing the jurisdictional loopholes that attackers exploit. In addition, moving beyond the reactive level, the ultimate mitigation lies in the proactive tier: *Data Starvation and Obfuscation*. This involves using AI to actively scan and "clean" OSINT exposure, or employing *adversarial poisoning defense*, injecting noise into public profiles, to corrupt the attacker's dataset. By degrading the quality of the input data, we effectively drive the personalization variable ( $\theta$ ) toward zero before the attack is even launched.

Finally, mitigation must alter the *attacker's economic logic* by increasing operational cost and reducing potential rewards to neutralize threats posed by Automation ( $\sigma$ ). This is achieved through *Systemic Hardening*, such as execution isolation and phishing-resistant authentication (e.g., passkeys), which increase entry barriers and prevent lateral movement. Furthermore, AI-on-AAI defense must treat Cyber Threat Intelligence (CTI) as an attack surface [30]. If the defender's automation is fed with synthetic, plausible yet false indicators, the entire pipeline becomes a self-sabotaging loop: the system "defends" against ghosts while real campaigns pass through. Thus, CTI integrity controls, such as provenance, validation, and anti-poisoning gates, are not auxiliary safeguards, but a prerequisite for sustaining any scalable defensive posture under  $\sigma$ -driven adversaries.

Ultimately, because humans remain the primary target in AI-enabled SE, keeping a human-in-the-loop as a strategic control is rational. Consequently, AI-specific awareness training to spot artifacts and persuasion cues—combined with model-level resilience and collaborative governance—helps restore the blocking probability ( $p_b$ ) and makes large-scale AI-mediated attacks operationally non-viable.

- RQ5: What ethical and regulatory concerns arise from the malicious use of AI in SE? Our analysis reveals that ethical and regulatory challenges bifurcate into two distinct frontiers: corporate accountability regarding development and active exploitation within the criminal sphere. This distinction separates civil responsibility from the challenges of systemic impunity.

The first frontier concerns the paradox of AI fuel and systemic opacity. While literature emphasizes *Personalization* ( $\theta$ ) through social networks, we argue that the deeper risk lies in the unprecedented “stock” of digital existence (health, behavioral, and economic data) amassed by manufacturers. This dynamic is increasingly amplified by the rapid acceleration of GenAI integration within organizations, as AI is no longer a peripheral tool for casual users or educational contexts but a default layer in day-to-day professional activity, quietly expanding the volume, velocity, and sensitivity of data flowing into AI-mediated systems. In practice, this creates an ethical double bind in workplace settings: employees may inadvertently disclose sensitive organizational information into public AI models [25], while reactive monitoring policies can create a security-by-surveillance tension that threatens employee privacy and rights. Given that, without data sovereignty, this transforms AI into a “weapon of surgical precision”. Beyond scale, the central governance failure is opacity: we contend that “black box” architectures are often a profit-driven choice. By prioritizing intellectual property and rapid deployment over auditability, manufacturers externalize social risk, leading to the *weaponization* of systems whose autonomous evolution remains uncontrolled even by their creators. In this context, transparency is not only desirable but foundational: beyond data protection, explainability is essential to build trust and enable accountability in deepfake detection and moderation, requiring XAI-style reporting that provides accessible reasons for automated decisions (e.g., Romero Moreno [23] highlights the textual/visual inconsistencies behind a deepfake classification and how salient features contribute to it).

The second frontier addresses the criminal democratization of deception. The Dark Web’s “AI-as-a-Service” model thrives by exploiting the non-auditable nature of legitimate tools to bypass safety filters. This intersection of rapid evolution and lack of oversight creates a legal void, where undetectable instructions and automated profiling foster systemic impunity. Even worse, the industrialization of abuse is no longer limited to underground ecosystems, but increasingly extends to legitimate AI marketplaces and agent repositories [22], where malicious capabilities can be packaged, distributed, and scaled. Consequently, the discussion must transition from abstract ethics to compulsory technical controls and “forensic-ready” architectures. Such measures are vital to counteract the scaling of *Automation* ( $\sigma$ ), *Realism* ( $\rho$ ), and *Personalization* ( $\theta$ ) factors now available to low-skilled actors. This also includes protecting the integrity of defensive pipelines themselves (as identified in RQ4), since AI-generated artifacts can contaminate intelligence and reporting layers, degrading automated defenses and further widening the accountability gap.

Ultimately, as supported by [39], if regulation fails to bridge the gap between corporate incentives and social safety, future defenses against AI-enabled deception will become operationally and economically unsustainable.

To consolidate the findings detailed above, Table 6 presents a structured synthesis of the primary scientific contributions derived from each Research Question. This overview demonstrates the trajectory of the study: from the granular identification of AI convergence in attack vectors to the formulation of the UM-AISE framework, and finally, to the proposal of restorative defensive strategies and governance models necessary to address the identified operational and legal voids.

**Table 6.** Summary of Scientific Contributions by Research Question.

| RQ  | Focus Domain            | Scientific Contribution                                                                                                                                                                                                                                                                                                                                                                                            |
|-----|-------------------------|--------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| RQ1 | Attack Vectors          | Taxonomy of Convergence: Maps distinct AI vectors (Profiling, Mimicry, Generation) directly onto the <i>Cyber Kill Chain</i> , identifying them as the foundational infrastructure that fuels the UM-AISE dimensions ( $\rho$ , $\theta$ , $\sigma$ ), while extending the operational backbone to industrialized AI-ecosystem abuse (marketplaces/APIs, dark services) and the fully autonomous agentic chains.   |
| RQ2 | Economic Dynamics       | UM-AISE Quantification: Quantifies “efficacy” via the proposed mathematical triad ( $\rho$ ), ( $\theta$ ), and ( $\sigma$ ), demonstrating (analytically) the structural inversion of the attacker’s cost-benefit analysis ( $Q_{AI} > Q_{Reset}$ ).                                                                                                                                                              |
| RQ3 | Cognitive Vulnerability | Systemic Trust Exploitation: Identifies “Impostor Bias” as the psychological flaw where AI fidelity neutralizes biological skepticism, amplified by real-time cognitive load and the collapse of biometric trust surfaces, and argues that this mimicry circumvents Zero Trust technical filters, drastically reducing the block probability ( $p_b^{AI}$ ).                                                       |
| RQ4 | Defense Strategy        | Restorative “AI-on-AI” Matrix: Proposes the strategic amplification of defense via Augmented Intelligence to neutralize $\rho$ , $\theta$ , $\sigma$ , including early-stage intent detection, OOB Authentication and CTI integrity controls (provenance, validation, anti-poisoning gates) as prerequisites for scalable defense.                                                                                 |
| RQ5 | Ethics & Governance     | Bifurcated Accountability Model: Offers a novel distinction between “Corporate Opacity” (data sovereignty issues amplified by rapid GenAI workplace integration and its ethical double bind) and “Criminal Democratization” (AI-as-a-Service and marketplace/agent-repository abuse), framing the need for forensic-ready, transparent (XAI-grounded) accountability mechanisms and pipeline integrity safeguards. |

## 6. Conclusions

Generative AI is widely discussed in the recent literature as a transformative force in social engineering, reframing it not merely as a technological upgrade, but as a potential economic and structural shift in attacker capabilities. Our synthesis suggests that AI does not merely enhance existing tactics; it may introduce a discontinuity in the attack lifecycle by amplifying Realism ( $\rho$ ), Personalization ( $\theta$ ), and Automation ( $\sigma$ ) in combination. This convergence allows attackers to shift strategy at the “Automation Plateau”, where marginal returns on scale diminish and the strategic advantage shifts toward realism and personalization, keeping the operational cost divisor low while increasing the potential reward, potentially shifting the threat from distinct, high-risk attempts to a continuous, high-yield algorithmic process.

The reviewed studies converge on three key pillars ( $\rho$ ,  $\theta$ ,  $\sigma$ ), which, integrated into our UM-AISE model, support the incentive compatibility of mass-customized deception. By mapping technological capabilities onto psychological vulnerabilities, specifically the erosion of systemic human trust, the model moves beyond a simple taxonomy. As highlighted in our analysis, emotional personalization and communicative realism are critical success factors, fueled by what we define as the “paradox of AI fuel”, the involuntary hoarding of private data enabling surgical precision in modern attacks.

A central contribution of this review is the articulation of the UM-AISE model and its formalization through an MDP-based analytical demonstration, which shows, under explicit assumptions, how capability convergence can shift the attacker’s optimal strategy ( $\pi^*$ ) toward more aggressive engagement. Rather than constituting empirical validation, the model is intended as an analytical lens to clarify incentives and to help prioritize defensive levers; its predictive validity remains contingent on future empirical and forensic validation.

However, despite technical advancements, the literature and our synthesis consistently caution that purely technical countermeasures are insufficient. AI’s capacity to simulate

human interaction introduces profound ethical, cognitive, and regulatory challenges. While the proposed restorative defenses can enable resistance, they function within a compromised ecosystem characterized by development opacity and unrestricted commoditization of crime. This disconnect creates a *legal void* where the sophistication of attacks outpaces the capacity for attribution. Furthermore, the emergence of new methodologies, such as “Centaur” attacks and fully autonomous agentic risks, suggest that future threats may penetrate defensive layers with minimal human involvement, potentially undermining traditional reactive postures. Therefore, effective cybersecurity policies must shift beyond abstract compliance toward enforceable auditability, ensuring that defender sovereignty is supported by governance mechanisms capable of piercing anonymity and disrupting the cycle of systemic impunity.

Although this review brings together significant evidence and multiple perspectives, several limitations point to crucial directions for future research.

First, we identify a critical empirical gap: the scarcity of quantitative data demonstrating the real-world impact of AI compared to manual social engineering. Despite widespread consensus on the escalating threat, there are no definitive forensic reports or longitudinal studies quantifying the magnitude ( $\Delta$ ) of efficiency gains or the measurable increase in success rates provided by automated campaigns. Consequently, the risk metrics discussed, such as the MDP-based model, remain largely theoretical. Future work must prioritize the collection of primary data and controlled simulations to validate whether our proposed variables ( $\rho, \theta, \sigma$ ) correlate with actual incident rates and cost-efficiency patterns observed in the wild.

Second, most of the literature analyzed is geographically concentrated in North America and Europe and predominantly published in English. This linguistic and regional bias limits the visibility of AI-mediated SE phenomena emerging in other socio-technical contexts. Expanding cross-regional and multilingual data collection could reveal distinct cultural and regulatory dynamics influencing attack vectors and user susceptibility.

Finally, because AI architectures evolve at an exponential rate, the taxonomies provided here represent a snapshot of a moving target. The rise of agentic systems and multimodal models suggests that future defensive strategies must remain as iterative and flexible as the threats they aim to mitigate. In sum, continuous reassessment and the development of international standards for data sovereignty will be essential to ensure that the defensive cost of AI does not become economically unsustainable for society.

**Author Contributions:** K.G. and S.S.: Conceptualization of this study, Methodology, Writing. K.G., S.S., M.G. and S.M.: Writing—review and editing. All authors have read and agreed to the published version of the manuscript.

**Funding:** This work is supported by The Applied Digital Transformation Laboratory (ADiT-LAB), through the Portuguese Foundation for Science and Technology (FCT—Fundação para a Ciência e a Tecnologia) within project: UIDP/06121/2025 and DOI identifier <https://doi.org/10.54499/UID/06121/2025>. This work was also supported by “CyberPRAISE—Cybersecurity research for PRivAte, Intelligent and truStable solutions”—NORTE2030-FEDER-01820300.

**Institutional Review Board Statement:** Not applicable.

**Informed Consent Statement:** Not applicable.

**Data Availability Statement:** The original contributions presented in this study are included in the article. Further inquiries can be directed to the corresponding author.

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

## Abbreviations

The following abbreviations are used in this manuscript:

|          |                                                            |
|----------|------------------------------------------------------------|
| ML       | Machine Learning                                           |
| AI       | Artificial Intelligence                                    |
| SE       | Social Engineering                                         |
| NLP      | Natural Language Processing                                |
| LLM      | Large Language Model                                       |
| OSINT    | Open-Source Intelligence                                   |
| GAN      | Generative Adversarial Network                             |
| GenAI-SE | Generative AI Social Engineering Framework                 |
| GenAI    | Generative AI                                              |
| BEC      | Business Email Compromise                                  |
| VALL-E   | Microsoft's Neural Codec Language Model for Text-to-Speech |
| SUS      | System Usability Scale                                     |
| IC3      | Internet Crime Complaint Center                            |
| ANN      | Artificial Neural Networks                                 |
| CNN      | Convolutional Neural Networks                              |
| RNN      | Recurrent Neural Network                                   |
| TTP      | tactics, techniques, and procedures                        |
| TTS      | Text-To-Speech                                             |
| SaaS     | Software as a Service                                      |
| DNN      | Deep Neural Network                                        |
| MFA      | Multi-Factor Authentication                                |
| MDP      | Markov Decision Process                                    |
| RL       | Reinforcement Learning                                     |
| API      | Application Programming Interface                          |
| UM-AISE  | Unified Model for AI-Driven Social Engineering             |
| CTI      | Cyber Threat Intelligence                                  |

## References

1. Kurshan, E.; Mehta, D.; Balch, T. AI versus AI in Financial Crimes & Detection: GenAI Crime Waves to Co-Evolutionary AI. In Proceedings of the 5th ACM International Conference on AI in Finance (ICAIF '24), New York, NY, USA, 14–17 November 2024. [\[CrossRef\]](#)
2. Kamruzzaman, A.S.; Thakur, K.; Mahbub, S. AI Tools Building Cybercrime & Defenses. In Proceedings of the 2024 International Conference on Artificial Intelligence, Computer, Data Sciences and Applications (ACDSA), Victoria, Seychelles, 1–2 February 2024; pp. 1–5. [\[CrossRef\]](#)
3. Schmitt, M.; Fléchais, I. Digital deception: Generative artificial intelligence in social engineering and phishing. *Artif. Intell. Rev.* **2024**, *57*, 324. [\[CrossRef\]](#)
4. Rathod, T.; Jadav, N.K.; Tanwar, S.; Alabdulatif, A.; Garg, D.; Singh, A. A comprehensive survey on social engineering attacks, countermeasures, case study, and research challenges. *Inf. Process. Manag.* **2025**, *62*, 103928. [\[CrossRef\]](#)
5. Nageab, W.M.; Alrasheed, R.; Khalifa, M. Cybersecurity in the Era of Artificial Intelligence: Risks and Solutions. In *Proceedings of the 2024 ASU International Conference in Emerging Technologies for Sustainability and Intelligent Systems (ICETIS)*; IEEE: New York, NY, USA, 2024; pp. 240–245. [\[CrossRef\]](#)
6. Mink, J.; Luo, L.; Barbosa, N.M.; Figueira, O.; Wang, Y.; Wang, G. DeepPhish: Understanding User Trust Towards Artificially Generated Profiles in Online Social Networks. In Proceedings of the 31st USENIX Security Symposium (USENIX Security 22), Boston, MA, USA, 10–12 August 2022; pp. 1669–1686.
7. Ahvanooy, M.T.; Mazurczyk, W.; Wang, Z.; Zhao, J. A novel framework for assessing determinant risk factors on cyber (dis)trust behaviors of netizens in deepfakes. *Eng. Appl. Artif. Intell.* **2025**, *159*, 111319. [\[CrossRef\]](#)
8. Alahmed, Y.; Abadla, R.; Al Ansari, M.J. 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)*; IEEE: New York, NY, USA, 2024. [\[CrossRef\]](#)

9. Balasubramanian, P.; Liyana, S.; Sankaran, H.; Sivaramakrishnan, S.; Pusuluri, S.; Pirttikangas, S.; Peltonen, E. Generative AI for cyber threat intelligence: Applications, challenges, and analysis of real-world case studies. *Artif. Intell. Rev.* **2025**, *58*, 336. [\[CrossRef\]](#)
10. Mirsky, Y.; Demontis, A.; Kotak, J.; Shankar, R.; Deng, G.; Liu, Y.; Zhang, X.; Pintor, M.; Lee, W.; Elovici, Y.; et al. The Threat of Offensive AI to Organizations. *Comput. Secur.* **2023**, *124*, 103006. [\[CrossRef\]](#)
11. Liu, Z.; Chen, Y.; He, Y.; Wang, Z.; Lu, H.; Zhang, X.; Wu, J. An Arms Race in the Inbox: A Systematic Review of Phishing Generation and the Rise of LLMs. In *Proceedings of the 2025 IEEE 10th International Conference on Data Science in Cyberspace (DSC)*; IEEE: New York, NY, USA, 2025; pp. 1–8. [\[CrossRef\]](#)
12. Webb, J.; Abri, F.; Akther, S. Synthetic Social Engineering Scenario Generation Using LLMs for Awareness-Based Attack Resilience. *IEEE Access* **2025**, *13*, 174856–174870. [\[CrossRef\]](#)
13. Valdez, H.P.D.; Abri, F.; Webb, J.; Austin, T.H. Exploring the Use and Misuse of Large Language Models. *Information* **2025**, *16*, 758. [\[CrossRef\]](#)
14. Malloy, T.; Ferreira, M.; Fang, F.; Gonzalez, C. Training Users Against Human and GPT-4 Generated Social Engineering Attacks. In *Proceedings of the HCI for Cybersecurity, Privacy and Trust. HCII 2025. Lecture Notes in Computer Science*; Moallem, A., Ed.; Springer Nature: Cham, Switzerland, 2025; Volume 15814, pp. 47–65. [\[CrossRef\]](#)
15. Mishra, R.; Varshney, G.; Singh, S. Jailbreaking Generative AI: Empowering Novices to Conduct Phishing Attacks. In *Proceedings of the 2025 55th Annual IEEE/IFIP International Conference on Dependable Systems and Networks—Supplemental Volume (DSN-S)*; IEEE: New York, NY, USA, 2025; pp. 1–6. [\[CrossRef\]](#)
16. Masood, M.; Nawaz, M.; Malik, K.M.; Javed, A.; Irtaza, A.; Malik, H. Deepfakes generation and detection: State-of-the-art, open challenges, countermeasures, and way forward. *Appl. Intell.* **2023**, *53*, 3974–4026. [\[CrossRef\]](#)
17. Tsinganos, N.; Mavridis, I.; Gritzalis, D. Utilizing Convolutional Neural Networks and Word Embeddings for Early-Stage Recognition of Persuasion in Chat-Based Social Engineering Attacks. *IEEE Access* **2022**, *10*, 108529–108548. [\[CrossRef\]](#)
18. Singh, S.U.; Namin, A.S. The influence of persuasive techniques on large language models: A scenario-based study. *Comput. Hum. Behav. Artif. Humans* **2025**, *6*, 100197. [\[CrossRef\]](#)
19. Anwar, S.; Perez, A. Risks and Ethical Concerns in Cyber Security with Advancements of Artificial Intelligence—A Systematic Review. In *Proceedings of the IEEE Frontiers in Education Conference (FIE)*; IEEE: New York, NY, USA, 2024. [\[CrossRef\]](#)
20. Kitchenham, B.; Charters, S. *Guidelines for Performing Systematic Literature Reviews in Software Engineering*; Technical Report EBSE-2007-01; Keele University: Newcastle, UK, 2007.
21. Brown, T.B.; Mann, B.; Ryder, N.; Subbiah, M.; Kaplan, J.; Dhariwal, P.; Neelakantan, A.; Shyam, P.; Sastry, G.; Askell, A.; et al. Language models are few-shot learners. In *Proceedings of the 34th International Conference on Neural Information Processing Systems*, Red Hook, NY, USA, 6–12 December 2020; Number Article 159 in NIPS '20; pp. 1877–1901.
22. Shen, X.; Shen, Y.; Backes, M.; Zhang, Y. GPTracker: A Large-Scale Measurement of Misused GPTs. In *Proceedings of the Proceedings of the 2025 IEEE Symposium on Security and Privacy (SP)*; IEEE: New York, NY, USA, 2025; pp. 336–354. [\[CrossRef\]](#)
23. Romero-Moreno, F. Deepfake detection in generative AI: A legal framework proposal to protect human rights. *Comput. Law Secur. Rev.* **2025**, *58*, 106162. [\[CrossRef\]](#)
24. Nowakowski, W. Social Engineering Analysis Framework: A Comprehensive Playbook for Human Hacking. *IEEE Access* **2025**, *13*, 18827–18849. [\[CrossRef\]](#)
25. Diro, A.; Kaisar, S.; Saini, A.; Fatima, S.; Hiep, P.C.; Erba, F. Workplace security and privacy implications in the GenAI age: A survey. *J. Inf. Secur. Appl.* **2025**, *89*, 103960. [\[CrossRef\]](#)
26. Qin, Y.; Yang, X.; Yang, L.X.; Huang, K. Mitigating Social Engineering Attacks Through Cost-Effective Security Awareness Training Policy. *IEEE Trans. Netw. Sci. Eng.* **2025**, *12*, 3145–3158. [\[CrossRef\]](#)
27. Akram, M.W.; Sood, K.; Ul Hassan, M.; Subba, B. Exemplifying Emerging Phishing: QR-Based Browser-in-the-Browser (BiTB) Attack. *IEEE Netw. Lett.* **2025**, *7*, 274–278. [\[CrossRef\]](#)
28. Eberl, L.; Engländer, L.M.; Löhle, C.; Jahnecke, D.; Baris, A.; Seljaci, D.; Shamsi, S.; Günther, J. Phishing and Identity Manipulation through Audiovisual Channels. In *Proceedings of the Open Identity Summit 2025*, Neubiberg, Germany, 22–23 May 2025; pp. 101–112. [\[CrossRef\]](#)
29. Li, M.Q.; Fung, B.C. Security concerns for Large Language Models: A survey. *J. Inf. Secur. Appl.* **2025**, *95*, 104284. [\[CrossRef\]](#)
30. Huang, H.; Sun, N.; Tani, M.; Zhang, Y.; Jiang, J.; Jha, S. Can LLM-generated misinformation be detected: A study on Cyber Threat Intelligence. *Future Gener. Comput. Syst.* **2025**, *173*, 107877. [\[CrossRef\]](#)
31. Chrysanthou, A.; Pantis, Y.; Patsakis, C. The anatomy of deception: Measuring technical and human factors of a large-scale phishing campaign. *Comput. Secur.* **2024**, *140*, 103780. [\[CrossRef\]](#)
32. Yang, K.C.; Menczer, F. Anatomy of an AI-powered malicious social botnet. *J. Quant. Descr. Digit. Media* **2024**, *4*. [\[CrossRef\]](#)
33. Shibli, A.M.; Pritom, M.M.A.; Gupta, M. AbuseGPT: Abuse of Generative AI ChatBots to Create Smishing Campaigns. In *Proceedings of the 2024 12th International Symposium on Digital Forensics and Security (ISDFS)*; IEEE: New York, NY, USA, 2024; pp. 1–6. [\[CrossRef\]](#)

34. Iturbe, E.; Llorente-Vazquez, O.; Rego, A.; Rios, E.; Toledo, N. Unleashing offensive artificial intelligence: Automated attack technique code generation. *Comput. Secur.* **2024**, *147*, 104077. [[CrossRef](#)]
35. Abbas, F.; Taeihagh, A. Unmasking deepfakes: A systematic review of deepfake detection and generation techniques using artificial intelligence. *Expert Syst. Appl.* **2024**, *252*, 124260. [[CrossRef](#)]
36. Singh, S.; Abri, F.; Namin, A.S. Exploiting Large Language Models (LLMs) through Deception Techniques and Persuasion Principles. In Proceedings of the 2023 IEEE International Conference on Big Data (BigData), Sorrento, Italy, 15–18 December 2023; pp. 2508–2517. [[CrossRef](#)]
37. Frankovits, G.; Mirsky, Y. Discussion Paper: The Threat of Real Time Deepfakes. In *Proceedings of the 2023 ACM Asia Conference on Computer and Communications Security (AsiaCCS '23)*; ACM: New York, NY, USA, 2023; pp. 1–5. [[CrossRef](#)]
38. Abri, F.; Zheng, J.; Namin, A.S.; Jones, K.S. Markov Decision Process for Modeling Social Engineering Attacks and Finding Optimal Attack Strategies. *IEEE Access* **2022**, *10*, 109949–109968. [[CrossRef](#)]
39. Blauth, T.F.; Gstrein, O.J.; Zwitter, A. Artificial intelligence crime: An overview of malicious use and abuse of AI. *IEEE Access* **2022**, *10*, 77110–77122. [[CrossRef](#)]
40. Swathi, P.; Saritha, S. DeepFake Creation and Detection: A Survey. In *Proceedings of the 2021 Third International Conference on Inventive Research in Computing Applications (ICIRCA)*; IEEE: New York, NY, USA, 2021; pp. 1–6. [[CrossRef](#)]
41. Cialdini, R.B.; Goldstein, N.J. Social influence: Compliance and conformity. *Annu. Rev. Psychol.* **2004**, *55*, 591–621. [[CrossRef](#)]
42. Amerini, I.; Barni, M.; Battiato, S.; Bestagini, P.; Boato, G.; Bruni, V.; Caldelli, R.; De Natale, F.; De Nicola, R.; Guarnera, L.; et al. Deepfake Media Forensics: Status and Future Challenges. *J. Imaging* **2025**, *11*, 73. [[CrossRef](#)]
43. Sundararajan, A. The Gig Economy Is Coming. Harvard Business Review. 2016. Available online: <https://www.stern.nyu.edu/experience-stern/faculty-research/gig-economy-coming-what-will-it-mean-work> (accessed on 10 June 2025).
44. Tampubolon, M. Digital Face Forgery and the Role of Digital Forensics. *Int. J. Semiot. Law* **2024**, *37*, 753–767. [[CrossRef](#)]
45. Verma, S.; Chauhan, R.; Rawat, R.; Pratibha. Deep learning algorithm for digital image forensics. In Proceedings of the 2024 International Conference on Automation and Computation (AUTOCOM), Dehradun, India, 14–16 March 2024; pp. 631–634. [[CrossRef](#)]

**Disclaimer/Publisher's Note:** The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.