Review

Social Engineering Attacks: Trends, Psychological Triggers, and AIdriven Prevention

Shashank Tiwari *

Posted Date: 9 October 2025

doi: 10.20944/preprints202510.0663.v1

Keywords: social engineering; phishing; spear phishing; vishing; smishing; business email compromise (BEC); pretexting; psychological triggers; cognitive biases; Cialdini persuasion principles; deepfakes & voice cloning; MFA fatigue; user & entity behavior analytics (UEBA); AIdriven detection; machine learning; natural language processing (NLP); anomaly detection; prompt injection & adversarial attacks; explainable AI (XAI); privacy & ethics; security awareness training; human factors; large language models (LLMs); generative AI

 doi:10.20944/preprints202510.0663.v1

Review

Social Engineering Attacks: Trends, Psychological Triggers, and AIdriven Prevention

Shashank Tiwari

MCA (Cyber Security) Scholar, SGT University, INDIA; shashank6889@gmail.com

Abstract

Social engineering is one the most common and potent attack techniques in cybersecurity, by which humans are deceived instead of computers, to breach a system and data. This article examines the recent evolution of social engineering attacks, the psychological factors that make them effective, and how advances in artificial intelligence (AI) are helping to combat them. Recent examples, such as the 2020 Twitter breach and the 2022 entrapment of Uber, showcase how adversaries are now combining multichannel vector tactics phishing, vishing, smishing and deep fake based impersonation with reconnaissance from social media and opensource data into sophisticated pretexts that defy credulity. The psychological principles applied were: authority, urgency, fear, trust/familiarity, reciprocation, social proof and commitment (paralleling existing theories of persuasion and cognitive bias). In this environment, AI and machine learning have become critical defensive weapons. The progress achieved is driven by AI powered email filtering, phishing URL detection, user and entity behavior analytics (UEBA), voice and chat scam detection, as well as adaptive phishing simulation for user enablement. Yet, AI based security systems also come with limitations such as adversarial evasion approaches , false positive/negative rates, privacy considerations and ethical issues related to the behavioral tracking itself. Attackers are more and more exploiting generative AI to produce hyper custom lures, generating what becomes an evolutionary arms race between offensive and defensive AI. This survey highlights the importance of having a sociotechnical approach that fuses psychological motivators with explainable and privacy preserving AI. Research agenda As a future research direction, emphasis should be placed to interdisciplinary collaboration, adversarial robust models and user centric security design in order to fight against the emerging threat of social engineering.

Keywords: social engineering; phishing; spear phishing; vishing; smishing; business email compromise (BEC); pretexting; psychological triggers; cognitive biases; Cialdini persuasion principles; deepfakes & voice cloning; MFA fatigue; user & entity behavior analytics (UEBA); AIdriven detection; machine learning; natural language processing (NLP); anomaly detection; prompt injection & adversarial attacks; explainable AI (XAI); privacy & ethics; security awareness training; human factors; large language models (LLMs); generative AI

Introduction

Social engineering is the manipulation of people to access private information or systems. It's the art of deceit so they can lie to you and get you to do what exposes your data, or even the company as a whole. Unlike technical exploits that attack software or hardware, social engineering focuses on the "human element," exploiting human cognitive biases and emotional responses to get around cyber defenses. The strategy is not new; throughout history there exist various examples of a distortion to gain an advantage (e.g. the mythological Trojan Horse in the form of an early "con" trick). During the contemporary period, social engineering within the domain of information security has become more prevalent in the latter half of the 20th century. In particular, the exploits of hacker Kevin Mitnick during the 1980s–1990s, who notoriously conned phone company employees into releasing access to their systems, did much to popularize such terminology in the realm of computer

 doi:10.20944/preprints202510.0663.v1

security. Mitnick's 2002 book, The Art of Deception, also in some ways underscored that if you trust people rather than the machines you are using to guard your information and work product, even the best technical protection in the world can be defiled. Social engineering attacks were already effective during the early computer era such as the 2000 "ILOVEYOU" email worm that tricked people into reading a fake love letter and inflicted $15 billion in damage or the 1999 Melissa virus, which rode into corporate infrastructures through an opening provided by a phony "important message" from someone they knew. It was not accomplished using advanced malware, but by tricking users into opening malicious attachments – a practice that is still widespread today.

We flash forward to the present digital world and social engineering is frighteningly ubiquitous. Verizon's Data Breach Investigations Report (DBIR) reveals 82% of data breaches in 2021 that had included a human related element, such as social engineering and misuse or errors by end users. It has been reported that phishing a popular social engineering attack tactic accounts for the starting point of nearly all cyberattacks (one Microsoft analysis found 91% of attacks start with a phishing email). In a sense, attackers have also learned from anyone involved in information defense industry and recognize that it is often easier to "hack a person" than it is to hack through the security of a well secured system. The COVID19 pandemic accentuated the risks of social engineering: 2020's sudden push to remote work opened up new possibilities for scammers, who could pose as IT support people or prey on fears facilitated by the knowledge that many employees were (and in some cases still are) sequestered outside the protective bubble of corporate IT supervision. In this era of pervasively interconnected world and the deluge of information, social engineering attacks have constantly expanded its scale and complexity; thus, an immediate focus for cyber security professionals and researchers. In the next few sections (Sects. 3–6), we present a comprehensive review of contemporary social engineering evolutions, insights into the psychological triggers that can give partial or total success to these techniques, AIbased means designed to prevent and mitigate such attacks as well as different ethical issues and challenges of using AI strategies against organizations in this field.

Recent Trends in Social Engineering Attacks

Tactics Influx: Social engineering tactics have evolved beyond traditional "Nigerian prince" email scams and widespread phishing campaigns originating in the early 2000s. Attackers are now using an array of tricks across various channels. Email phishing is still rampant, however… but along with it we're increasingly seeing spear phishing that's highly targeted (i.e., personalized messages sent to individuals or roles) and whaling (attacking high value targets like CEOs), often with personal data taken from social media used to craft get baits And let's not forget the plethora of information sharing and gathering tools out there. Smishing (SMS/text phishing) and vishing (voice phishing through phone calls) are increasing as users get wise to suspicious emails. Indeed, voice call scams have become even more convincing thanks to AI generated speech impersonation: in one recent instance, by imitating the voice of real people such as CEOs or relatives and conning victims over the phone. Attackers have expanded their channels as well, now using messaging apps and social networks as vectors for impersonation schemes. Per an analysis of 2025 incidents by Unit 42 (Palo Alto Networks), over one third of social engineering attacks are nonmailable vectors – such as malicious browser popups (spoofed security alerts), software update requests with malicious payloads, and impersonation of a help desk via chat or call. These tactics prey on users across platforms, not just the inbox.

Pretexting and Impersonation: One of the defining trends is the sophistication of pretexting, in which malicious actors fabricate complex backstories that make them appear legitimate. Today's attackers frequently do some amount of recon on their targets (LinkedIn, data breaches, etc.) to create another realistic sounding pretext. They may be pretending to be an IT support worker, a vendor or even a colleague in trouble. A well documented recent example is the 2020 Twitter hack, in which teenagers successfully posed as members of Twitter's IT department. By telephoning employees and saying there were VPN problems (plausible, given the amount of remote working taking place), they fooled staff into logging onto a bogus VPN page, thus giving up their usernames and MFA codes.

 doi:10.20944/preprints202510.0663.v1

This social engineering toehold enabled the hackers to break into Twitter's internal admin tools and use them to hijack scores of high profile accounts (including those of Elon Musk and Barack Obama) for posting cryptocurrency scam messages. The Twitter hack provides a stark example of the efficacy of social engineering as the first phase in an attack: function (1) was social engineering to get onto the network, then functions (2) and (3) were technical exploitation and malicious use of popped accounts. Similarly, in 2022 rideshare company Uber was breached after an employee account managed by a contractor was compromised using social engineering. According to The Guardian, the attacker brought the contractor's password on the dark web, attempted numerous logins, causing a deluge of MFA push notifications and when one finally got approved after a tired user set it through, he gained entry and continued onto sensitive internal systems. This "MFA fatigue" approach – repeatedly asking a user to login until they give in has become a popular social engineering method of breaking down multifactor authentication barriers.

Channels and Campaigns: Email phishing is still the most prevalent social engineering attack. Recent phishing campaigns are also taking advantage of current events and urgent topics waves of COVID19 phishing emails in 2020, for instance, enticed the recipients with discussions about pandemic news or vaccine appointments. BEC scams have become a multibillion dollar problem, involving a perpetrator posing as a company executive or trusted business partner in order to request that an employee make wire transfers of funds. The FBI Internet Crime Complaint Centre has received thousands of BEC complaints each year, resulting in worldwide losses of billions of dollars. An example of a typical BEC scenario could be an attacker impersonating a CEO's email address, and sending the finance department an instruction to transfer money into what must now surely – no matter how large the transaction – be the actual supplier's (the sender), bank account. The other cycle in use is "baiting" and quid pro quo: attackers with infected USB drives on which is written "Confidential," or an offer of a small return for information or access to something. The ubiquitous Internet of Things (IOT) and digital footprints we leave on social networks create even more fuel for the fire for modern day Social Engineers. The overabundance of personal info floating out on the Web, coupled with weak privacy laws, makes it a "field day" for attackers when collecting reconnaissance data on their targets, as one security pro sums up. Attackers cobble together phone numbers, workplaces and even real time location data from social media allowing for some highly persuasive approaches.

Case Studies and Notable Incidents: Aside from the aforementioned Twitter and Uber hacks, plenty of other breaches of late underscore ling social engineering. In 2023, a notorious case of social engineering saw ghost hackers pulling off lucrative deepfake voice calls to imitate the director of a business and approve an illicit bank transfer – epitomizing how AI is being combined with social engineering & voice verification checks in order to subvert the digital defenses businesses and consumers rely upon. "Another example is the 2022 breach of a big video game company through compromise of internal Slack channels by posing as IT support, who tricked someone into entering their credentials into a false login prompt indicating how communication tools within corporates can be exploited for social engineering," they added. According to the EU's cybersecurity agency (ENISA), between 2023 and 2024, there was a rise in attacks such as "callback" phishing campaigns where an attacker sends out instructions via email for victims to call a number which is run by the attacker under some pretext or another, inviting Realtime social engineering over the telephone. Fall In – Fake Web Browser Alerts (Popups which say "Your PC is infected, call us now!" etc.) have ensnared tons of users into scam call canters, leveraging a combination of technical trickery and social engineering. The Unit 42 (2025) release highlights that social engineering "is now one of the most reliable, scalable and impactful intrusion techniques" they've found during their review of incidents across which it represented 36% of cases. Curiously, they noticed an increase in "high touch" social engineering handson interaction with victims through voice or live chat a type of cyberattack that allowed attackers to circumvent security and make their way into computer systems while breaching them in real time (eg convincing a help desk to reset a password), often without deploying malware. Attack groups are getting better at integrating with enterprise processes and hiding within the weeds;

 doi:10.20944/preprints202510.0663.v1

we saw one attacker who moved from the first victim in under 40 minutes from gaining an initial foothold, to being a domain admin purely through advanced social manoeuvring and identity abuse. On the whole, recent patterns have seen social engineering attacks becoming more pinpointed, tech fuelled and multichannel. Bad actors leverage whatever the victim is most likely to believe – whether that's an email, phone call, text or even personal communication. The unifying theme is the desire to leverage human vulnerabilities as a way of circumventing technical measures.

Psychological Triggers and Human Vulnerabilities

What is so effective about social engineering attacks? The answer has to do with playing the game of human psychology. Attackers write lures to invoke automatic, conditioned behavior pressing our cognitive "buttons" as it were. Social engineers consistently take advantage of cognitive biases, emotions and social pressures that can make even informed users drop their guard. Like any other social engineering, one of its strong points is how it capitalizes on feelings and convictions. Common psychological triggers include:

Urgency and Fear : Instilling some sense of time pressure and or fear of consequences is a classic approach. Common language in phishing emails includes "your account will be closed in the next 24 hours if you fail to act" or "unauthorized activity has been detected –please verify now!" The induced panic short-circuits rational scrutiny; the targeted rush to comply before they have time to think. In the midst of the COVID19 crisis, many scam messages used fear ("You've been exposed to the virus, click here for urgent instructions") to induce instant clicks. Urgency is effective because when our brains are stressed, they often default to swift action, as the attackers know all too well. So they rely on the fear of punishment when some threatening message (an impersonation, a fake legal notice or security alert) claims to come from the government. An academic research about phishing tactics verifies that those depicting time context and exploiting fear appeals generate victim compliance rate increase due to instinctive reflex.

Authority and Trust: We're hardwired to obey an authority image as well as trusting someone we believe is credible. This inclination is abused by social engineers, who pose as authorities bosses or IT administrators at your company, bank officials, government officers to take advantage of the characteristic. An earlier generation of experimental social psychology, by Stanley Milgram in the 1960s, showed that we obeyed authorities when we did not want to the power of authority on behavior. This is exploited by phishing scams that send emails with spoofed CEOs or law enforcement letterheads, for example. One of the most common forms is that BEC/CEO fraud we mentioned earlier, where an email "from" the CEO orders a worker to wire money now – combining high status with urgency. People are loathe to challenge requests that appear to come from a higherup. Attackers also take advantage of trust by establishing rapport or garnering legitimacy (assailants talk from some of your public personal information to seem friendly devils, commoner imps). With an email or phone call that seems to originate from within one's company or a trusted service provider, the instinct is to believe it. Social proof (the assumption that "others are doing it, so this isn't wrong") can enhance trust e.g., a phishing page may contain fake reviews or cc multiple people to imply agreement. Consensus/social proof is one of the key persuasion principles described by psychologist Robert Cialdini; attackers use it against us, insinuating that an action is normal or popular among our peers.

Reciprocity & Curiosity: There is a generally held desire in people to reciprocate favors – the principle of reciprocity. Attackers take advantage of this with a modest gift or act of kindness to disarm victims. For example, they could send an unsolicited email containing a free coupon or a "useful" PDF report and then ask the victim to register or provide information in return. Just the act of paying any attention at all (such as a friendly personal email) can trigger an insidious pressure to be reciprocated in some way by clicking or replying. Then there is the strong feeling of curiosity, given that malicious emails typically have interesting or exciting subjects (such as "Take a look at this photo of you!" or "Amended Salary Attached") which entice people into satisfying their own

 doi:10.20944/preprints202510.0663.v1

curiosity by clicking on a link or opening an attachment. Likewise, when it comes to baiting attacks using USB drives or fake ads, the curiosity can get the better of victims.

Liking and Similarity: It is easier to become persuaded by those who we like or identify with. Social engineers can mirror a coworker's style or act friendly to increase liking. They may allude to shared interests or echo the recipient's language in an attempt at bonding. Occasionally an online scam is the long con, bringing along a would-be lover to help him with his scheme. Brand recognition is another consideration – an email that appears to mimic how a popular brand looks and feels (including logos and appropriate language) gains instant trust with the recipients via behalf of brand recognition and fondness. This is the very reason why pages used in phishing campaigns are crafted to visually mimic actual login pages of banks, mail providers and whatnot – they make us trust those brands.

Consistency and Commitment: We all desire to be consistent with our commitments or past decisions. An attacker could exploit this by obtaining a modest initial concession and then demanding more. Many would start by getting the victim to respond to an inquiry of no consequence, a survey say (to establish dialogue), and later bran move on with requests for confidential information. When people say "yes "to a small offer, they're then more likely to want to keep saying yes to that same offer (it's part of our need for consistency) – 'foot in the door'. Cialdini also included commitment when someone has taken a stand or role (e.g., "helpful employee"), we humans are very likely to live up to the implications of that role.

Many researches and models have examined these psychological triggers in social engineering. Indeed, Rosana Montañez and coauthors (2020) have put forward a way of looking at how social engineers trick basic human cognitive functions – be it the attention system (distracting targets away from heeding warning signs), the memory systems (relying on familiarity or context cues to snatch up available space in our tight working memories your brain is willing to trust for retrieving patterns on demand under stress), or various decision making shortcuts – so as to increase their odds of success. And a detailed 2025 review by Stylianou et al. measured the effectiveness of different persuasion approaches in phishing. They discovered that "Authority, Commitment & Consistency, and Reciprocity" among others were the most successful tactics in increasing victim compliance rates within experimental paradigms. Curiously, the same study found that in groups with a low baseline level of compliance group pressure (they called it "Majority Size", effectively social proof to induce agreement) was very effective. "Essentially, if a communication message says an action is popular or supported by many others already doing it well, even sceptics can be persuaded.

Note, the attackers stack these triggers together, for more effect. For example, a phishing email could leverage authority ("This is the IT department"), and urgency ("your account will be locked today") and reciprocity ("click here for a free security scan"). Together, this can overwhelm the target's critical abilities. On top of that, social engineers also utilize what Microsoft Security Team refers to as the "shortcuts" taken during human decision making. People are overloaded with information and routine tasks, so we depend on mental shortcuts (heuristics) to make expedient decisions such as whether or not to trust a legitimate seeming email or follow an order from one's boss. It takes advantage of the way we respond to situations even when our brains are switched off. So basically, these human traits and weaknesses – trust, fear, greed curiosity, compassion or complacency amongst others are the actual "attack surface" that is being exploited here. While people feel emotions and have cognitive biases, social engineers are going to try to prey upon them. This is why security training that teaches people to recognize when those psychological buttons are being pushed is so important. And, it adds, it's a reminder that solely technical defenses aren't ever going to be sufficient – recognizing and factoring in the human element is key.

AI-Driven Prevention and Detection Mechanisms

With the rising complex nature of social engineering attacks, defenders have been (sensibly) using Artificial Intelligence (AI) and Machine Learning (ML) to fight such threats. AIbased tools promise to be able to sift through enormous data sets, find faint patterns, and do it all much faster

 doi:10.20944/preprints202510.0663.v1

than a human analyst. This part focuses on how the AI/ML can help in the identification, prevention and mitigation of social engineering attacks, actual tools and frameworks and recently published works from 20202025.

ML to the Rescue : The Phishing Email Detection One of its earliest AI applications in this sector is email security. Contemporary email providers and secure email gateway vendors utilise machine learning models to accurately detect phishing and spam emails. These models can consider attributes of an email message like headers and sender reputation, as well as content (such a language patterns) to identify malicious messages that might slip past traditional rule based filters. For instance, even Google's Gmail employs AI based filters that scan each and every one of the emails traffic that is heading inside – according to Google its AI led defences block 99.9% of spam, phishing and malware emails while it sends around 15 billion unwanted emails into the junk box per day. This is a reminder of the magnitude problem for phish that would be all but unmanageable by humans, and shows how effective AI can be at combatting it. Approaches such as NLP allow detectors to identify phishing emails by phrases contained in them (for example: Recognizing what make a 'phishing lure' in text). Machine learning can also be used to identify phishing URLs by dissecting features of the link (length, presence of abnormal tokens, etc.) or even rendering the page in a sandbox and checking if it's a fake login page. Academic research in recent years has investigated the use of deep learning models (such as recurrent neural networks or transformers) for enhancing phishing email classification, which can include detecting spear phishing attacks that leverage language more contextually and with a target in mind. These models learn tiny differences between, let's say an email that appears to be sent from a CEO but is not (differences in writing style or metadata) and a true communication. The result: Much more adaptive detection that can spot new phishing tactics that perhaps a signature based approach would miss. Then, some enterprise AI systems engage in anomaly detection based on email behaviors for example, if an employee's account abruptly begins emailing different recipients at odd hours (a hint the account is compromised and being operated by a phisher), the system may alert or intervene automatically.

User Behavior Analytics and Anomaly Detection: AI is also being applied to build models of normal user behavior and identify anomalies that could signal social engineering led compromise, rather than just scanning communication content. This is in the realm of User and Entity Behavior Analytics (UEBA). Machine learning models train off logs of user activity (logins, patterns, transactions), and profile what a usual day is for regular users or roles. If an account that usually doesn't do much begins moving 10GBs of data at 2 a.m., or if the credentials of an employee who normally works in an office are used to access your network from another country, these systems will raise a red flag. These kinds of anomalies can be indicative of an attacker who was able to leverage stolen credentials from a phishing or some other social engineering attack. By detecting misuse early on, AI powered monitoring can help limit the damage (for example by automatically freezing the account or prompting for reauthentication if suspicious activity is detected). ITDR isn't purely AI, it's more akin to AI training or apprenticeship– where artificial intelligence watches how creds are used then mentions "this looks like fishy activity" when appropriate. Unit 42 now advises organizations to incorporate behavioral analytics and ITDR proactively in order to catch misuse of credentials and lateral movement that frequently occur subsequent to social engineering compromise. The advance notice that these technologist give defenders could be enough to get them a foothold for responding before the contents of this messaging is used against the user, effectively taming the attack.

Malicious Call & Chat Detection: One other frontier that AI/ML can be used is to detect social engineering in voice calls and chat conversations. As vishing and deepfake audio attacks are on the rise, various researchers have proposed voice analysis algorithms to identify voice spoofing or anomalous speech. AI might analyze patterns in people's vocal tone and sentence structure while they're on the phone to catch calls that are likely scam attempts. For example, there are experimental solutions such as the "Antisocial Engineering Tool (ASsET)" that has been developed to detect telephone scams by examining the semantic information exchanged in a conversation. ASsET and like methods use NLP strategies to identify "scam signatures" – particular patterns of dialogue or

 doi:10.20944/preprints202510.0663.v1

sets of phrases that are distinct to social engineering calls (akin to how AV products have malware signatures). With enough input (spoken or transcription) of dialogue from phone scams and clustering/classification algorithms on a transcript, such systems could potentially flag when a call is reading from a known scam script (eg faux IRS agents, tech support). This is a fertile field for research, and commercial call filtering apps are starting to employ AI to block robocalls (and suspected vishing) attempts. Similarly, chatbots and messaging apps have started using AI content filters to detect any suspicious or fake message in live chats. For instance, some banking chat services incorporate AI that can sense if a user is reading from a script; possibly following instructions of the scammer (i.e strongarm type coercion or abnormally abnormal requests are its basis).

AI for Social Engineering Prevention: Detection and prevention go together. AI is also being worked into to help users become more aware and empowered. One interesting trend is that even the training themselves are now AI based; systems, like ant phishing training services, can automatically use machine learning to devise a targeted phishing email test or in some cases just generate a more realistic phishing simulation via AIs. Organisations can build up resilience by inoculating their users via AI generated fake phishing emails, coupled with feedback. Some email clients, with AI assistants and all, can also proactively alert you within a message. For example, an AI might read an email on a user's behalf, and provide a warning that says: "This email requests a password and exhibits characteristics that resemble known phishing; use caution." ("But as I'll explain later, attackers are also attempting to manipulate AI assistants directly themselves a trend called prompt injection to evade these defenses.") At a larger level, AI can facilitate the maintenance of blacklists and reputation systems in a more dynamic way. Services are using machine learning to scour the web for phishing sites, fake social media profiles and malicious content and take them down or warn users before acting.

Tools, Frameworks and Developments: In the academic domain all over the literature a lot of research frameworks and prototypes exist. In 2021, A neural networks (CNNBiLSTM) and attention based hybrid model RFCNNBiLSTM was used to detect voice phishing in Korean, obtaining an accuracy of >99% on experiments. Another emerging trend that's poised to change the game in defensive strategies is the use of large language models (LLMs) and generative AI. Yet with AI, the same class of technology attackers utilize to create fake content is also available for defenders, making it possible to more effectively analyse and screen that pretend stuff. For example, researchers are exploring LLMs that could understand the intent behind an email or message; an AI might "comprehend" that a piece of text is meant to scare the recipient into clicking a link and label it as likely phishing. AI is also being used to monitor transactions for example, among companies that have started using AI algorithms at banks to detect and block suspicious transactions (based on patterns of known scam techniques and unusual amounts, for instance) or rampant scams while happening in real time (such as when fraudsters socially engineer their prey into moving money).

Some specific applications and products include: email security AI (such as machine learning models in Microsoft 365 Defender that can spot BEC attempts by comparing the writing style of incoming email with historical messages to examine metadata), browser based AI protections (web browsers using machine learning to flag deceptive sites or fake login forms), and AI assisted authentication (systems analyzing logins for risk scores could, upon spotting a login from an unrecognised device, trigger extra verification based on what a ML model reckons is risky activity). In industry, vendors such as Abnormal Security and Barracuda Sentinel advertise AI-based detection of social engineering (particularly BEC) through the use of communication pattern modelling. In the meantime more and opensource projects and academic tools (e.g., such as ASsET) are elbowing their way into areas such as telephone scam detection.

Lastly, AI is being tapped to handle incident response to social engineering. Upon detecting an attack, each of these components uses AI for playbooks in order to automate actions such as quarantining phishing emails on all mailboxes, resetting owned accounts, or even conducting conversation with the attacker (thought there is an AI chatbot of course) to stall them and gain information. These responses, already automated thanks to artificial intelligence's analytical speed,

 doi:10.20944/preprints202510.0663.v1

can contain some of the threat that originates in social engineering before it becomes a potentially greater breach. At a high level, AI and ML are key new vector's for defence tools against social engineering from filtering phishing communications at scale, to watching user behavior for strange activities, to assisting in user training and awareness. With attacks on the rise in both volume and complexity, however, AI's ability to adapt and learn provides a significant advantage when it comes to outmanoeuvring social engineers.

Challenges and Ethical Considerations

Although it is clear that AIbased solutions provide an added combatant to fell, there are many challenges and ethical concerns introduced with AI in the fight against social engineering. These concerns have to be taken into consideration so that prevention rules are effective, fair and constitutional.

False Positives and Negatives: A limitations of AI based detection systems also comes into play in practice: the need for high accuracy. If an AI system is too permissive, it will generate false negatives failing to flag phishing emails or malicious calls that go on to reach the victims. Alternatively, if it is too strict, then it will generate false positives flagging or blocking legitimate communications and activities. Both outcomes carry costs. A false negative could result in a compromise (e.g.phishing email bypasses gateways, infecting an end system machine with ransomware). False positives can be disruptive to business and damage trust in security controls (such as trapping critical messages or locking out an executive for "anomalous" behavior), as the poster notes. Certainly, AI detectors leave something to be desired. Attackers adapt continuously to how they are being caught, and they sometimes use AI themselves for producing increasingly humanlike phishing content that ML models get a reputation from telling real text. At the same time, many of these content detectors do fail some or much of the time – one study found extremely high rates of false positives when screening specifically for AI written text. This concept can be applied to email security; an overzealous AI could flag a carefully crafted but authentic business email as "suspicious" because it's formal and contains data (characteristics that it was trained on for phishing). So trade off sensitivity can be quite pesky: defenders have to continually calibrate their models and add feedback loops so they don't make errors. And when AI does make decisions (to cut of a message, lock an account) transparency is crucial – users and admins needs to know why so they can put things right if it's wrong (this plays into the wider issue about being able to explain why your AI made a certain decision in security).

Adversarial Attacks on AI: There seems to be a cat and mouse game where attacks try to spoof or even attack the AI defenses. One such emerging threat is adversarial examples, which alter input in subtle ways to fool ML models. A phishing email, for instance, might contain intentional spelling errors or esoteric Unicode characters that humans would skim over but a machine learning classifier could be easily thrown off of. We've learned about attackers who embed bad instructions in specific content that's designed to trick AI systems the Gmail "prompt injection" exploit found five years ago is a good example. In that instance, attackers put hidden text inside emails (including white on white text) that would be read by AI summarization tools. The AI would then spit out a manipulated synopsis for the user, essentially making an honest AI assistant an unwilling co-conspirator informing the victim of a lie (such as, "Your account is in danger, call this number now"). This new attack demonstrates that AI defenses can themselves be tricked using clever questioning of the input provided to the AI. As opponents employ AI (such as using generative AI to create more realistic feeling phishing content at scale), defenders' AI faces an arms race in which both sides have adaptive instruments. What's more, phishing attempts that feature a high degree of customization ("spear phishing 2.0") can now employ AI to replicate a person's writing style, further minimizing the potential for detection. Security teams need to have that adversarial resilience for their AI models to be resilient against manipulation and updating continuously so they can identify an attack content from the AI.

 doi:10.20944/preprints202510.0663.v1

Privacy Issues: Much of the AI based solutions need to analyse a vast number of users data i.e., communication, behavior logs, biometrics etc. This raises significant privacy issues. For example, an AI system could be required to read through all messages including personal or sensitive ones like incriminating emails in order to detect them. Behavior analytics can also mean tracking employees' keystrokes or web activity, or tracking their physical locations (to identify anomalies), and that can make them feel watched. There are dangers that explained data could be abused or stolen, especially as the AI systems become targets for hackers. We need strong data protection and governance so that security AI doesn't ride roughshod over people's privacy rights. An ethics in AI approach would involve data minimisation (only looking at what's needed), anonymising where possible and providing clear information for users about what is being monitored. In some places, going deep on communications can bump up against regulation (some countries have tight privacy rules that might consider reading employees' emails without the employee's knowledge or a valid legal reason a no no). Companies need to weigh their security requirements with their responsibilities around privacy – AI tools for prevention must always be law abiding and ethically governed. The ethically responsible use of behavioral data is especially thorny: if an AI identifies someone as a potential insider threat or susceptible to social engineering, what do you do with that information? With care and representation, it runs the risk of unfair profiling or punishment on the basis of AI judgements which may be incorrect. Transparency and human scrutiny are important – decisions that have a large impact on someone (say accusing them of being compromised or limiting their access) ought not to be made solely by AI without human review.

Bias and Fairness: Models using AI can unconsciously perpetuate biases rooted in whatever data they were trained. In the case of social engineering detection, this might mean that the AI is better for some languages than others (it may have been trained mostly on English phishing emails and is less likely to catch them in, say, Japanese). There might even be instances where certain types of user behavior (which are prevalent in some cultural settings) unjustly get labeled as "risky" from an universal model. Fairness and non-discrimination is the ethical obligation. For instance, if an AI tracking tool is inclined to raise a red flag over foreign logins, it could also disproportionately impact those employees who travel frequently for work or may be based beyond the borders of the country where they are employed with potential implications on their ability to do their jobs and on public perceptions. The AI's judgments should be constantly evaluated to make certain it is judging real threats, not the accidental characteristics of specific users or groups.

Dependency and Duty : A related issue is the danger of becoming too dependent on AI. Organizations could have a false sense of security ("the AI will catch it all") and therefore miss opportunities for user education and other controls. In practice, we should hedge our bets: AI as an assistive tool rather than a crutch. There is also a question of responsibility when an AI messes up – if a machine learning filter misses a phishing email and costs you a major data breach, who do you look to for responsible? This gets to some of the broader AI ethics topics around explainability and liability. It's a reminder about how AI in cybersecurity should be used to complement human analysts, not replace them altogether. Human intuition and contextual awareness is very much needed today to catch those edge cases or make judgment calls on alerts that the AI generates.

Behavior Manipulation and Autonomy: On the other side of the coin, AI implemented defensive deception also triggers ethical concerns. Also take "honeypot" avatars or chatbot bots pretending to be victims to interact with scammers – that is, in effect, tricking the tricksters, which may possibly fall into a legal Gray area if you haven't been given permission. When even well meaning nudges (such as a warning popup from an AI "This link looks dangerous, you sure? if not handled with care could be criticized as manipulative design. We need to make sure we don't veer into unduly clouding users' autonomy, or contributing to what's called "alert fatigue" responding too often that it dilutes trust in the alerts.

To sum up, AI offers new and powerful tools to counter social engineering, but it is not a magic bullet. Technical constraints think false positives and negatives, as well adversarial exploits require constant refining, in which the algorithm is married to human judgment. When it comes to deploying

 doi:10.20944/preprints202510.0663.v1

AI for security reasons, ethical questions – around privacy, transparency, fairness and accountability – need to be at the heart of efforts. We need a multidisciplinary perspective, including ethicists and lawyers, to develop guidelines for the responsible use of AI in cybersecurity. Users need to be informed about how their data is being used and secured, while "privacy by design" should be a best practice for security AI systems in organizations. Ultimately, the objective is to make security stronger without compromising our foundational principles privacy, trust and fairness) that security exists to protect. By facing these problems headon and taking a proactive stance, we can leverage the power of AI in combating social engineering effectively and ethically.

Conclusions

Social engineering is and has been (though that doesn't seem to be stopping it any time soon) one of the most powerful threats present in modern cyber security – an issue based not on technical flaws, but human psychology and behavior. Attacks have evolved from simple email scams to incredibly sophisticated multicurie campaigns that prey on our trust, fears and cognitive shortcuts. The examination of trends in the recent past uncovered that humans are still the "weakest link": Phishing emails, scam phone calls, impersonation on social media – adversaries have perfected the phishing of people to get around even the strongest technical defenses. Meanwhile, defensive players are trying to catch up by using more sophisticated tactics and technology. AI and machine learning are now leading the charge in preventing and detecting social engineering efforts, ultimately processing billions of messages, while simultaneously analyzing user behavior in real time to identify any signs of compromise. AI powered systems have already proven themselves to be taking miraculous action – from stopping well over 99% of attempts at phishing to catching out suspicious behavior that suggests someone somehow got fooled – and they're closing the attack window dramatically.

But this race is not finished. Attackers are also using AI themselves (e.g., to create barely detectable phishing lures or deepfake voices) in an ever escalating arms race between offensive and defensive AI. Furthermore, AI creates intricate ethical and operational risks false alarms, privacy concerns, and adversarial resiliency that the security community will need to steer though carefully. There is an increasing realization that fighting social engineering is not just a technical challenge, but a sociotechnical one. Thus future research needs to be interdisciplinary. One avenue for further research is more extensive integration of behavioral science into cybersecurity: learning the subtle influences that lead people to fall prey to scams, and designing interventions that can inoculate users from manipulation. For instance, they might develop adaptive training systems that tailor cybersecurity education to an employee's cognitive profile or history of susceptibility (while also respecting privacy). There's also a lot of work to do in research around explainable AI in security such that decisions made by automated systems (such as where an email or user account is flagged) are clear and can be verified by analysts – this will build trust in the tools using AI, and help tune them too.

Interdisciplinary research does, in fact seem a promising way forward as one paper states it, we need to shake ourselves out of the coma of "Dicksonianism" and combine forces between IT engineering and psychology (see e.g. [46] for further discussion on this). This might lead to suggestions about what kind of personality types are associated with an invitation for phishing, or how group dynamic within organizations can be exploited to enhance a group's watchfulness. Another one is AI for simulating social engineering attacks (in a controlled environment) to keep testing and improving human and AI defences. For example, generating extremely realistic phishing simulations with generative AI as a way of better training users could be very effective kind of like creating a "vaccine", where benign simulations teach us how to build resistance against the real thing.

On the defense technology, we may investigate next how to better detect the AI empowered attacks. With this boundary becoming less clear between human authored and machineproduced phishing content, researchers are working on creating models that can recognize the slight differences in AI generated messages or deepfakes. Coordination between academia, industry and organizations,

 doi:10.20944/preprints202510.0663.v1

such as CERTs (Computer Emergency Response Teams), will be crucial in sharing threat intelligence to new social engineering tactics and how AI can help combat them. Further, there will be ethical considerations setting up controls around the use of employee data for threat detection that protects privacy and ensuring AI tools don't create biased or unfair outcomes. Examples include AI ethics guidelines and regulatory adherence (the coming European Union AI Act, for instance) on what role AI can responsibly play within cybersecurity themes.

In summary, social engineering attacks flourish due to mankind's never ending relationship with the traits of trust and error yet when humans are alert and intelligent machines join, a powerful defense forms. There's a common saying in cybersecurity: "humans are the weakest link." It turns out, however, that with increased understanding and AI augmentation, humans can also be the strongest line of defense. The fight against social engineering is first and foremost a fight for the "hearts and minds" of users one that cannot be won without acknowledging doubts about human nature as much as we indulge certainties in networks and algorithms. It's up to all of us, by keeping investigating the psychology of deception, developing better and AI driven defenses and instilling a healthy scepticism into our corporate culture. Of course this arms race will never be over, and attackers will evolve, but the gap between attacks and defenses can be reduced by continued research as long as it is done in an ethical manner. As the Unit 42 report put it so succinctly, adversaries aren't just hacking systems; "they're hacking people" – and that is why our mission should be to bolster the human element – via technology as well as training. In a comprehensive framework, in addition to preventing intruders from getting beyond our firewalls, our future cyber defenses should enable each and every user to become his own best firewall against social engineering attacks.

Acknowledgments: The author would like to thank the cybersecurity community for the incident reports, datasets, and research that informed this review. Continued knowledge sharing between industry and academia is vital in the fight against social engineering threats.

References

1. Choi, T. (2022, May 31). *Key findings from the 2022 Verizon Data Breach Investigations Report underscore the role of the human element in data breaches* [Blog post]. Proofpoint. proofpoint.comproofpoint.com
2. Dziwa, A. A. (2022). *How Social Engineering Bypasses Technical Controls*. *ISACA Journal, 2022*(5). isaca.orgisaca.org
3. Gatefy. (2021, March 19). *Social engineering history in the age of computers and the internet* [Blog post]. Gatefy Blog. gatefy.comgatefy.com
4. Hoplon InfoSec. (2025). *Gmail Phishing AI Prompt Injection: The 2025 Guide to Spotting and Stopping AIDriven Email Attacks* [Blog post]. Hoplon Security. hoploninfosec.comhoploninfosec.com
5. Microsoft Security Team. (2020, June 30). *The psychology of social engineeringthe "soft" side of cybercrime* [Blog post]. Microsoft Security Blog. microsoft.commicrosoft.com
6. New York State Department of Financial Services. (2020). *Report on the Twitter Cybersecurity Incident of July 15, 2020*. NYS DFS. dfs.ny.govdfs.ny.gov
7. Palo Alto Networks Unit 42. (2025). *2025 Unit 42 Incident Response Report: Social Engineering Edition*. Palo Alto Networks. unit42.paloaltonetworks.comunit42.paloaltonetworks.com
8. Stylianou, I., Bountakas, P., Zarras, A., & Xenakis, C. (2025). Suspicious minds: Psychological techniques correlated with online phishing attacks. *Computers in Human Behavior Reports, 4*, Article 100694. zenodo.orgzenodo.org
9. Traviss, M. (2024, March 21). *Humans aren't prepared for AI phishing – neither is AI*. *Innovation News Network*. innovationnewsnetwork.cominnovationnewsnetwork.com
10. Almaslukh, A. (2024, October). *AI could empower and proliferate social engineering cyberattacks*. World Economic Forum. https://www.weforum.org/stories/2024/10/aiagentsincybersecurity
11. Ferreira, A. (2019). Persuasion: How phishing emails can influence users and decisionmaking. *Computers & Security, 87*, 101592. https://doi.org/10.1016/j.cose.2019.101592

 doi:10.20944/preprints202510.0663.v1

12. Hogan Lovells. (2024). *Confronting social engineering in the age of artificial intelligence*. Hogan Lovells Insights. https://www.hoganlovells.com
13. Hossain, I., Puppala, S., Talukder, S., & Alam, M. J. (2025). AIintheloop: Privacy preserving realtime scam detection and conversational scambaiting. arXiv:2509.05362.
14. Kumarage, T., Johnson, C., Adams, J., Kirchner, M., Hoogs, A., & Hirschberg, J. (2025). Personalized social engineering attacks and detection with LLM agents. arXiv:2503.15552.
15. Matz, S. C., Kosinski, M., Nave, G., & Stillwell, D. (2017). Psychological targeting as an effective approach to digital persuasion. *PNAS*, *114*(48), 12714–12719. https://doi.org/10.1073/pnas.1710966114
16. Ozen, I., Subramani, K., Vadrevu, P., & Perdisci, R. (2024). SENet: Visual detection of online social engineering campaigns. arXiv:2401.05569.
17. Schmitt, M., & Flechais, I. (2023). Digital deception: Generative artificial intelligence in social engineering and phishing. arXiv:2310.13715.
18. Wang, Z., Ren, Y., Zhu, H., & Sun, L. (2022). Threat detection for social engineering attacks using machine learning. arXiv:2203.07933.
19. Wantenaar, L. (2022). Social engineering and persuasion in cyberattacks. *Journal of Cybersecurity*, *6*(2). https://doi.org/10.1093/cybsec/tyac002