CYBERSECURITY Social Engineering / Corporate Security / Human Vulnerability / AI-Augmented Attacks

The Human Firewall: The Analysis of Social Engineering weaknesses in the Contemporary Business Society

You can invest millions of hardening your network perimeter, implementing zero-trust architecture and running the most advanced endpoint detection in the industry - and one employee who clicks the wrong email on a Tuesday afternoon can bring it all down. The most serious weakness within any organisation lies not within the code. It lies in the hands of human beings who utilize it. And in the era of social engineering with AI enhancement that vulnerability has never been so well used.

By **Usman Siddiqui** | Agentic AI Engineer | ML & DL Expert | CS Student | Published: June. 2026
Specialisation: Cybersecurity, Agentic AI, Machine Learning, Deep Learning, Social Engineering Defence Systems

An Email That Cost $25 Million

At the beginning of 2024, a multinational firm employee in Hong Kong was on a video call and it seemed to be with his Chief Financial Officer of the company when he received what he thought was a video call. The CFO, with his face, voice, mannerisms, all of it, was on screen together with what appeared to be other executives of the organisation. They had the employee approve a set of fund transfers amounting to 25 million USD. He complied. All the individuals on that call were deepfakes. None of the people in that video were real. This worker was chatting with a room of AI-generated fakers, and he had no clue until the cash was stolen and the actual CFO called in regarding a totally different subject.

This could not be done by a nation-state actor, with billions of resources. It was the creation of a criminal organisation that had access to publicly available deepfake tools, some social engineering tradecraft, and a basic knowledge of the process of corporate authorisation. The technical difficulty of carrying out this attack was truly astounding. The effectiveness of psychology was truly amazing. And the unpleasant reality is that this very attack (or a slight variation on it) could be executed tomorrow against thousands of organisations worldwide, including right here in Pakistan, using tools that can be found on the shelf and be used by anyone with the technical know-how to do it.

Introducing the current social engineering environment. My aim in writing this article was to chart it out in as candid and as full a way as possible - to describe what social engineering is, both psychologically and technically, how it works so reliably even against advanced organisations, how AI and machine learning are enhancing the arsenal of the attacker, and what a serious, evidence-based human firewall programme actually looks like, when built to operate rather than to pass a compliance checkbox. The issue can be resolved. But first we must be clear what it is.understand it clearly first.

"The worst weakness in any organisation is never in the code. It is in the human beings that utilise it and no patch has ever been issued to human psychology."

Defining Social Engineering: The Art of Hacking Humans

The term social engineering in the context of cybersecurity means influencing the psychology of humans to make them do things or reveal information that they would not have otherwise done or revealed with the intention of obtaining unauthorised access to systems, data, or financial resources. It is fundamentally an information security confidence trick. The attacker is not penetrating the wall; he/she is persuading somebody inside the building to open the door. And they have been doing this with great success since the advent of corporate environments.

The difference between the modern social engineering and the old-fashioned con artistry is the accuracy, magnitude, and timeliness that technology achieves. An older con man may have a couple of scams every year. A social engineering attack today can be executed on thousand or more highly personalised, each attack is specific to the psychological profile, working environment and known relationships of the target, and the cost per attack is approaching zero. The maths around this are highly unfavourable to the defenders. You must prevent each and every assault. All that the attacker needs to do is to succeed once.

In machine learning terms - the prism I apply to this as one who works at the interface of AI systems and security - social engineering is nothing but an adversarial attack on the human classification system. People are continually engaged in quick binary decisions: legitimate or suspicious, trustworthy or fraudulent, safe or dangerous. Our system of classification was developed to accommodate an essentially different social context, in which the information sources were scarce, identity could be checked through the sense of sight and hearing, and the effects of a misclassification were localized and manageable. Attackers take advantage of the discrepancy between our heuristics as we evolve and the environment in which we actually exist. They design inputs that are tailored to induce false positives in our trust classifier.

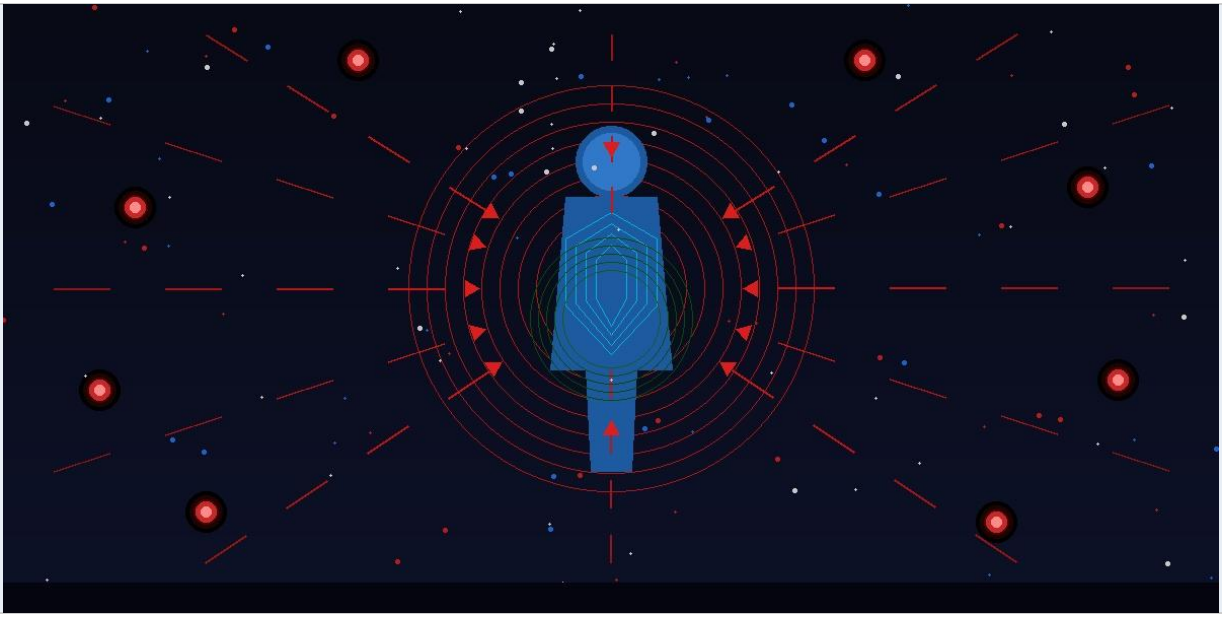

Figure 1: The Human Firewall under siege — a visualisation of the multi-directional social engineering attack surface facing every individual in a modern corporate environment. Red vectors represent active threat channels; the human at the centre is simultaneously the most critical defensive layer and the most frequently exploited vulnerability. Credit: Usman Siddiqui / PakTech Today

The Attack Taxonomy: Every Weapon in the Social Engineer's Arsenal

The first step to prevent social engineering is to be familiar with its entire taxonomy. Security awareness programmes which concentrate solely on phishing, the most widely discussed, but by no means the sole technique, leave organisations in a very vulnerable place in regards to the wider attack surface. There are at least eight categories of attacks that are defined as modern social engineering, each exploiting various vulnerabilities in psychology and various facets of corporate operating conditions.

Phishing and its Variants. Phishing is the most common and economically harmful type of social engineering, but the word has now adopted a much broader attack surface than the mass-email campaigns of the early 2000s that were crudely executed. Spear phishing focuses on specific individuals using highly personalised content based on open-source intelligence (OSINT) collection on social media or professional networks and publicly available databases. Whaling attacks senior executives and makes decisions tailored to their positions and power. Vishing performs the same manipulation on voice calls, typically making use of caller ID spoofing to masquerade as colleagues or trusted institutions. Smishing uses the SMS platform, and it takes advantage of the increased level of trust that most people have in text messages as opposed to emails. The defensive response needed is different in each of the variants, yet the psychological engine is the same: exploiting trust, urgency, and authority.

Pretexting. Pretexting is the practice of providing a fabricated situation, or pretext, in order to obtain information or access on a target. One of the most typical examples is an attacker calling the IT support and posing as a high-ranking executive that has been locked out of their account and requires urgent access to be reinstated. The attacker has conducted sufficient research to be aware of the name of the executive, his/her department, and perhaps the name of the manager. They take advantage of the natural disposition of the IT support technician to be supportive, the power of the executive character, and the pressure of time presented by an artificial emergency to circumvent verification mechanisms that are specifically designed to frustrate such attempts. The pretexting attacks work especially well with organisations that have strong service cultures whereby employees are trained to be responsive rather than cautious.

Business Email Compromise (BEC). BEC is among the most expensive types of social engineering attack worldwide, with a reported loss of more than 2.9 billion in the United States alone in 2023, according to the FBI Internet Crime Complaint Center. The main characteristics of a BEC attack are that either an attacker steals the legitimate executive email account or generates an authentic impersonation of one and directs the finance or accounting departments to execute fraudulent wire transfers or reassign payroll. The brilliance behind BEC is that it takes advantage of the authoritative systems within organisations that are legitimate in organisations - the hierarchies that enable organisations to operate effectively are used as attack vectors when paired with impersonated executive accounts.

Baiting and Quid Pro Quo. Baiting is about receiving something enticing, such as a free USB drive in a car park, a download that is too good to be true, access to premium content, in exchange to an action that will provide access to the system or install malware. Research has time and again demonstrated that an alarming number of individuals are willing to stick an unverified USB drive they have discovered into a work computer in spite of the general security awareness training to the contrary. Quid pro quo attacks provide a service in exchange of information - a typical example is an attacker impersonating IT support and offering to fix a reported issue in exchange of the user providing his or her login credentials. They both take advantage of fundamental human reciprocity instincts which lie beneath the threshold of conscious thought.

Tailgating and Physical Intrusion. Social engineering is not always screen and phone based. Physical social engineering - escorting a person inside a secured building, posing as a maintenance technician or delivery man, wearing a counterfeit visitor badge - is very efficient because most physical security is based on the same psychological premises that digital attacks are based on. Individuals do not wish to be uncivilized. They do not

want to confront somebody who appears like they do. They do not want to be mistaken about one who appears to be confident and official. Such social norms, which are actually prosocial in most situations, are a liability in high-security settings where any unverified entry can be a violation. The attacker is not going to breach the wall. They are persuading someone inside the company to open the door. And they have been doing that successfully since corporate surroundings were created.

> ***"The attacker is not going to breach the wall. They are persuading someone inside the company to open the door. And they have been doing that successfully since corporate surroundings were created."***

The Psychology Engine: Why These Attacks Work Every Time

To get to the point of knowing why social engineering occurs, one must look deeper than just a superficial explanation of the attacks themselves and into cognition science of human decision-making. It is not because these attacks are made against intelligent and educated people who are security conscious. It is the foreseeable, written functioning of psychological laws that are profoundly inscribed in the human cognition. The original study of influence by Robert Cialdini has been used in the foundational research on influence and six major principles of influence highlighted by the author are reciprocity, commitment, social proof, authority, liking, and scarcity that social engineers are able to use with remarkable accuracy. All the above needs to be understood in the context of a corporate attack scenario to ensure any person who is constructing a serious defence understands each of them.

One principle that is most abused in the corporate world is authority. When a message is framed as an announcement by the CEO, a regulatory agency, a senior partner or an IT security team, then most individuals are under the visceral urge to follow instructions that override the process of critical assessment. This is not a personality defect, but rather the adaptive product of social systems, which actually would be more adaptive when the majority of the population obey legitimate authority the majority of the time. The issue is that our system of authority detection uses superficial signals which include an email address that sounds official, a tone of voice that sounds certain, a logo that seems to be valid, and so on, instead of cryptographically authenticated identity. Until not so long ago, that sufficed. In the era of AI-generated content, it is disastrously inadequate. The cognitive shortcuts that are the most widely used to avoid thorough analysis are urgency and scarcity. When an email message mentions that your account will be locked in 30 minutes unless you verify your credentials, it is precisely designed to induce the fight-or-flight response, which disables the prefrontal cortex, the slow and deliberate and analytical area of the brain, and activates the fast and reactive limbic system. Security researchers have recorded that most highly trained people make much poorer security decisions when they are put under artificial time pressure. Attackers know this. Urgent cultivation is not a chance attribute of phishing attacks, but a design principle.

Social proof uses the very human trait of seeking to use the behaviour of others as a reference point to our own behaviour, specifically in new or uncertain situations. When an attacker is able to plausibly say that all other members of your department have already gone through this verification process or that your coworker has signed this transfer this morning, s/he is using a cognitive shortcut which is useful in the real world social context but proves counterproductive in the context of a counterfeit social proof. Social proof attacks are especially effective in organisations that have high collaborative cultures, the type of cultures that facilitate effective teamwork. The same traits that make one a good team player, are the ones that make them more vulnerable to manipulated social context.

The most pernicious of the principles is liking, since it is the most difficult to notice at work. We will obey requests of people we like much more dramatically and we like like-minded people, those who compliment us

and familiar people. An advanced social engineer takes weeks or months to develop a legitimate connection with a target - by interacting in a real way on the professional network, praising his work, making himself an ally and friend in the workplace - before striking. When the request arrives, the guard of the target is actually down as the relationship is a real one. One of the most frequent high-value corporate espionage cases is the investment in the cultivation of liking prior to the attack.

AI-Augmented Social Engineering: When the Threat Learns to Learn

Assuming that traditional social engineering is a scalpel, AI-enhanced social engineering is a precision-guided missile. The same technologies of large language models, deep learning, and generative AI that are generating massive value in legitimate applications are also generating an attack surface that is qualitatively new to anything corporate security has needed to deal with in the past. Since I deal with these models on a daily basis, I can confidently say that we are in the initial stages of a threat escalation cycle that the security community has yet to properly price.

Poor language quality has been the most accurate indicator of a phishing attack, which large language models have removed. The phishing emails with grammatical errors and clumsily translated sentences that security awareness training has taken years of educating employees to be aware of is no longer a current artefact. An LLM-generated phishing email in modern times is no different in terms of the quality of the language used as compared to a real communication written by a native speaker with full professional context. Worse, having access to the publicly visible social media posts of the target, his/her LinkedIn profile, and any publicly available communications, an LLM can create hyper-personalised phishing messages making references to actual projects, actual colleagues, actual organisational environment, and the specific language usage and communication style of individuals that the target actually knows. The generic Dear Customer phishing attack has been transformed into Hey Sarah, could you please re-read this document before the Thursday deadline - based on what we talked about during the Q3 planning session.

Phishing attacks have been exponentially increased by voice cloning technology. Using as little as three seconds of audio recording, which can be found in any publicly available video, podcast, or recorded talk, commercially offered voice cloning technologies can render a synthetic replica of the voice of an individual that not only passes the test of casual listening but also passes more rigorous tests over time as well. This implies that the I heard the voice of my boss hence I knew it was real heuristic that employees have been using long enough to verify phone calls is no longer effective. It is all the more true with even stronger force with the video deepfakes of the type that can be employed in the $25 million attack that I have described at the beginning of this paper. With the ever-growing quality of the generative models and the ever-decreasing inference costs, the price of generating a believable real-time deepfake video call will reach zero in the next two to three years. The future of this threat is agentic AI attack systems.

Where today AI-enhanced social engineering still involves human operators to instigate and operate individual attacks, agentic systems are capable of operating social engineering campaigns at multi-stage, continuously, and autonomously, establishing relationships over time, performing OSINT reconnaissance, personalising attack content, and adjusting strategy to target responses. These systems are already being coordinated and implemented by advanced threat actors, and the disparity between the capabilities of nation-states and the commercially offered threat environment is shrinking at an alarming rate. Planning defensive posture based on the threat landscape of 2022, security teams are not prepared to the threat landscape of 2025 and beyond.

> ***"Using just three seconds of audio recordings, commercially available voice cloning applications can create a realistic voice, which, when heard by a***

> ***casual listener, passes the test of not knowing it is synthetic. The heuristic of hearing voice and knowing it is real ceases to be dependable."***

The Corporate Vulnerability Landscape: Where Organisations Are Most Exposed

Not every employee and not every organisational function has equal social engineering vulnerability and the vulnerability landscape in a particular organisation must be comprehended in order to intelligently allocate defensive resources. In its annual State of the Phish report, Proofpoint consistently reports finance, HR, and IT as the top three most at risk departments, not due to the lack of intelligence or security-awareness in the employees of these departments, but because they are the departments that have the highest number of high-value assets (financial systems, employee information, system access) and operational processes requiring regular communication with third parties.

New workers constitute an unequivocal vulnerability window. During the initial 90 days of employment, a person lacks an effective intuitive map of organisational communication norms, is not well acquainted with all with whom they deal to discern voice or email impersonation, and is under the social pressure to be helpful and responsive, which defines the new-hire experience. Attackers will target organisations with open positions, and based on publicly accessible information on job postings and LinkedIn updates, attackers will target newcomers as they can be compromised prior to their effective onboarding and integration.

The top executives, on the contrary, are the most susceptible targets despite the fact that they have the highest intensive security awareness training. The reason is that their time pressure is quite real and very intense, their volumes of communication are enormous, and the consequences of being unhelpful or unresponsive, even to a strange request, seem to them greater than to a junior employee who can put in a call to get permission to do so. This dynamic is what attackers capitalize on: whaling attacks are explicitly constructed in such a way as to make use of the executive authority themself and time pressure against them, making urgent requests sound like a necessity to a system that is being presented with a significant bureaucratic waste to someone used to operating at a certain speed.

Distanced and blended work areas have greatly increased the corporate attack area in ways that most organisations have not entirely come to terms with. When people work at home, the informal checks that are created by physical co-location such as you can walk over to the desk of a colleague and check, you see your boss in the corridor and confirm the meeting, are lost. Remote workers are increasingly relying on digital communications to do all interactions, more separated due to informal social intelligence that aids in detecting unusual behaviour, and more prone to working across time zones and communication channels in a manner that unusual requests are harder to stand out against a definite background of normal behaviour.

Key Findings: The Damage in Numbers

The operational and monetary harm due to social engineering attacks is appalling, and the trend lines are heading towards this entirely wrong direction. According to the 2024 Data Breach Investigations Report by Verizon, the human factor, in the form of social engineering, human error, and abuse, was involved in 68 percent of the data breaches worldwide. In the IBM Cost of a Data Breach Report 2024, the overall cost of a data breach was reported to be highest than ever at an average of $4.88 million USD USD, with the breach involving social engineering attacks being more expensive than the average cost, as it has a dwell time that is considerably longer than the average, and is likely the result of an initial credential-based compromise. In the United States alone, the FBI IC3 report reported a total of more than 12.5 billion in cybercrime losses, with BEC attacks representing the highest percentage of nearly 3 billion.

The situation is alarming to the Pakistani corporate world, in particular. It is claimed that reported phishing and BEC incidents in Pakistan have been on the rise every year according to the Pakistan Telecommunication Authority and CERT-PK, but the real numbers are thought to be hugely underreported because of the stigma attached to reporting a successful social engineering attack. Its fast-growing fintech industry, its vast banking industry with large investments in digital transformation, and its outsourcing industry dealing with sensitive data and financial transactions on behalf of international clients are all high-value targets posing a social engineering risk on an unprecedented scale that most organisations in the sector have not properly staffed to address.

In addition to the financial indicators, the strategic loss of the social engineering attacks may be more lasting than the direct financial loss. Whenever an attack leads to the loss of intellectual property, client information, or strategic plans, the loss of competitive advantage and client confidence can be existential to organisations whose business concept is based on the image of reliability and safety. Some of the big outsourcing deals in the technology industry in Pakistan have been canceled due to security breaches with social engineering being a factor in that- a trend that ought to be a wakeup call to an industry that relies on international trust to make a living.

Building the Real Human Firewall: What Actually Works

This is where most security awareness articles fail, and I want to be very clear as to why. The standard business model of the human factor in cybersecurity is a yearly online training, quarterly phishing test, policy paper no-one reads, a poster in the break room, is not a human firewall programme. It is theatre of security. It meets audit demands and provides organisations with something to refer to when the post-breach inspection they were bound to experience demands what was being done to train the employees. However the study is categorical that these methods result in little lasting behaviour change and a phenomenally insufficient enhancement in real attack detection rates. To construct a real human firewall, a radically different strategy that is based on behavioural science instead of compliance needs is required.

Ongoing, Continuous Training on an Annual Dump of Compliance. The learning science evidence is clear: repeated learning that is spaced and includes frequent short sessions leads to markedly superior knowledge retention and behaviour change compared to infrequent long sessions. A good human firewall programme presents security education in brief, regular, contextually specific bursts - a five minute simulation based on a recent threat, a five-minute team discussion about a recent industry event, a Friday weekly roundup of new pattern of attacks in the wild. Some platforms such as KnowBe4 and Proofpoint Security AwarenessTraining have gone that way, but it is up to the organisation to commit to using it on a regular basis and not just to tick the box of compliance annually.

Psychological Inoculation, Not Information Transfer. The best training in social engineering defence, does not merely inform employees what attacks should look like - it immunises them against the psychological processes that attacks use. The study of inoculation theory, which was initially applied to the idea of propaganda resistance, indicates that exposing people to diluted versions of manipulative strategies, such as explaining how urgency is created, displaying how authority cues are created, illustrating how social proof is created, etc., produces resistance to the manipulative strategies in real-life context. Training to make employees aware of the emotion of being manipulated, as opposed to the content of manipulation, is far more readily applicable to new attack situations than that which simply teaches employees to be aware of particular formats of attacks.

A Culture of Checking but not Punishing. Implied punishment of challenging authority or delaying an urgent request is one of the most harmful cultural processes in most organisations. When an employee requests further confirmation of a request made by someone purporting to be the CEO, or who insists on a wire transfer request awaiting proper authorisation, he or she is in the right frame of mind security-wise. They also face a risk of being perceived as obstructionist, paranoid, or inadequately responsive in most organisations. Those leaders who are

serious about human firewall effectiveness have to establish overt, recurrent, culturally internalized structures of permission to precisely this sort of security-induced friction. It is not a training problem but a leadership behaviour problem. It must be modelled from the top.

Technical Controls which minimize the impact of Human error. There will be no training programme of this quality that will have zero failure rates when having to face an opponent that is as good and well-equipped as modern social engineers are. The human firewall is a vital defence point, but it has to be supported with technical controls to reduce the blast radius of successful attacks. Credential theft attacks become much less consequential by using multi-factor authentication, especially hardware security keys and passkey-based authentication. Authentication standards (SPF, DKIM and DMARC) make it harder to send spoofed emails. Privileged access management means that even a fully compromised employee account cannot be used to get a direct access to the crown jewels of the organisation without going through extra verification procedures. Zero-trust network architecture restricts subsequent movement once some initial compromise has occurred. Such technical controls do not substitute the human firewall; they cause it to be resilient to the failures even the most effective human firewalls will sometimes suffer.

AI-Powered Defence Systems. The very AI that is enabling attackers can be and should be used in defence. The phishing and BEC attempts can be fine-tuned by machine learning models trained on vast amounts of legitimate and malicious communications to be detected much more accurately and consistently than rule filters or human reviewers alone. Behavioural analytics systems are able to define thresholds of normal user behaviour, and indicate unusual behaviour - an employee who suddenly starts to access vast quantities of sensitive documents when he is not in his regular working hours, a wire transfer request that arrives via a strange channel at a strange time - that are not immediately visible to human observation at scale. Deepfake detection devices, though in their infancy, are improving fast and they are starting to be incorporated in the enterprise communication system. Attacker versus defender AI arms race is a reality, and organisations that are not increasing their defensive capabilities with AI are disarming themselves unilaterally in a conflict where the other side is constantly increasing.

> ***"An annual training and a poster in the break room is not a human firewall programme, it is security theatre, and the research is clear: it generates little lasting behaviour change.upgrading."***

Red Teaming the Human Firewall: Measuring What You Think You Have

What you cannot measure cannot be controlled, and the social engineering exposure of a human firewall is highly measurable in case an organisation has the heart to be sincere in measuring it. Social engineering red team exercises are controlled, authorised efforts to attack the employees with realistic social engineering attacks, which offer objective information about the actual vulnerability stance of the human layer of an organisation. Not the score of the employees on awareness quizzes but the actual performance of the employees when they are faced with a complex, realistic attack under circumstances that simulate the real adversarial situations.

An effective social engineering red team exercise does not only involve sending out a phishing email and monitoring the number of clicks. It encompasses pretexting calls to IT and finance, physical access attempts at office locations, targeted spear phishing campaigns based on OSINT collected on individual high-value targets, vishing attacks on individual executives, and, more recently, deepfake video and voice attacks, which challenge defenses in the AI-enhanced threat environment. Measures of interest are not merely whether an employee was deceived, but how the employee was able to escalate the incident once having become suspicious, whether the employee reported a successful deception quickly, and whether the process and technical controls that were

applicable at the time constrained the damage that a real attacker could have inflicted using the same initial foothold.

The fact that the culture of red team exercises should be framed in terms of organisational learning, and not the blame of a person, is of critical importance. The organisational response to an employee failing a social engineering simulation is to realise why the attack was believable and to offer immediate contextual learning that is relevant to that particular situation and to investigate whether the process or technical controls should have detected what the individual failed to detect. Labeling and humiliating those employees who fail simulations (a much more common practice than it ought to be) has the precisely opposite effect: employees grow more focused on being caught not passing a test and less focused on actually learning to judge security.

What's Next: The Escalating Arms Race and Pakistan's Imperative

The future of social engineering as a threat actor is evident and not promising. These technical obstacles to the implementation of highly advanced, AI-enhanced social engineering attacks will keep declining as generative AI models will be increasingly competent, more affordable, and more available. Attacker personalisation and psychological accuracy of attacks will only get better as attackers have access to higher quality OSINT data feeds and more effective means of processing and exploiting them. With the development of completely autonomous agentic attack systems, the human operator bottleneck, which is now the main limiting factor in the amount and complexity of attack, will gradually be eliminated. We are on a path to a world where all employees of all ranks of all organisations are a constantly targeted surface to highly sophisticated, personally-targeted, social engineering attacks, at a scale and velocity that cannot be effectively defended by human alone.

The proper reaction to this course is not hopelessness, but immediacy and investment. The gap in security awareness is a real phenomenon among the corporate sector in Pakistan and it is increasing. It is the organisations that will invest today in actual behavioural change programmes, AI-enhanced defensive tooling, verification culture, and frequent red team measurement, that will create a meaningful and lasting edge over those that remain stuck on compliance-driven theatre. To the technology professionals of Pakistan, the engineers, the security practitioners, the ML specialists who know the capabilities of the attacks as well as the defence technologies, this is a real and an expanding professional opportunity. The crossroads of AI and cybersecurity, most notably in social engineering defence, is one of the most impactful current fields of applied practice in the field.

The human firewall does exist. It can be constructed, and constructed properly. However, it must be addressed as the multidisciplinary, multidimensional, ever-present security layer that it is, rather than a second-thought that the majority of organisations make it. The assailants have already invested to get to know more about your people than you do. Whether your organisation will be prepared to invest the same time in learning, training, and securing them. With the technology of one click on a Tuesday afternoon and millions in security investment being undone, the answer to such a question is no longer a choice.

About the Author

Usman Siddiqui is an Agentic AI Engineer and Machine Learning and Deep Learning expert currently pursuing his degree in Computer Science. He works at the convergence of artificial intelligence and cybersecurity, with a particular focus on social engineering defence systems, AI-augmented threat modelling, and the application of deep learning to anomaly detection and behavioural analytics in enterprise security environments. His work spans both the offensive and defensive dimensions of the AI-security intersection, giving him a practitioner's understanding of how the same technical capabilities can be weaponised and defended against. He is committed to elevating the cybersecurity conversation in Pakistan's technology community to match the sophistication of the threats the sector faces.

Keywords: Social Engineering; Human Firewall; Cybersecurity; Phishing; BEC; Deepfake Attacks; AI-Augmented Threats; Corporate Security; Behavioural Science; Security Awareness; Pakistan Cybersecurity; Zero Trust

References

[1] Verizon (2024) *2024 Data Breach Investigations Report,* Verizon Business. www.verizon.com/dbir

[2] IBM Security (2024) *Cost of a Data Breach Report 2024,* IBM Corporation. www.ibm.com/security/data-breach

[3] FBI Internet Crime Complaint Center (2024) *Internet Crime Report 2023,* Federal Bureau of Investigation. www.ic3.gov

[4] Cialdini, R. B. (2021) *Influence: The Psychology of Persuasion, revised edn.,* Harper Business.

[5] Proofpoint (2024) *State of the Phish 2024: An In-Depth Look at User Awareness, Vulnerability and Resilience,* Proofpoint Inc. www.proofpoint.com