

Emotional Malware: A Review of Emotion-Driven Cyber Threats, Psychological Manipulation, and AI-Enabled Social Engineering

1. **Shaily Shah** (shahshaily2310@gmail.com), M.Sc. Cyber Security Student,

Department of Biochemistry and Forensic Science, Gujarat University, Ahmedabad, Gujarat, INDIA.

2. **Dr. Kiran Dodiya** (kirandodiya01@gmail.com), Teaching and Research Associate (TRA),

Department of Biochemistry and Forensic Science, Gujarat University, Ahmedabad, Gujarat, INDIA

3. **Dr. Parvesh Sharma**(times.parvesh@gmail.com), Assistant Professor

Department of Biochemistry and Forensic Science, Gujarat University, Ahmedabad, Gujarat, INDIA

4.**Dr. Kapil Kumar** (kkforensic@gmail.com), Professor & Coordinator,

Department of Biochemistry and Forensic Science, Gujarat University, Ahmedabad, Gujarat, INDIA

Abstract

Social Engineering and psychological hacking target the cognitive shortcomings of individuals, making them the weakest link in the field of cybersecurity. Traditional computer hacking and cyber threats target the software, networking, and systems; however, more contemporary threats are concerned with the manipulation of human psychology and emotions. This paper proposes a new idea and term, “Emotional Malware”. Through a review of the literature, this study (i) identifies cyber threats and attacks done using AI models, (ii) describes a complete framework with psychological theories and manipulation techniques and decisions and how the algorithm would work, (iii) reviews and evaluates existing studies and practical methods for detecting and counteracting those attacks using explainable AI (XAI), behavioral analysis, and multimodal deepfake detection and zero-trust architecture, and (iv) analyzes the damage and impact that it has on society and individuals. The following case studies (2023 to 2026) describe and define emotional malware and describe the method and type of the cyberattacks that occurred.

Keywords

Emotional Malware, Social Engineering, Artificial Intelligence, Psychological Manipulation, Deepfake Detection, Explainable AI(XAI), Behavioural analysis, Cognitive Bias, Human-Centered Cybersecurity, Decision-making, Human emotions.

Review Methodology

This review methodology outlines the methods used for collecting data and sets rules for sources and terms used and database search.

Databases covered: IEEE Xplore, Elsevier, SpringerLink, ACM Digital Library, and Google Scholar (DOI confirmation). Search terms: Terms used in the intersection of three groups: (i) emotional manipulation and social engineering, (ii) behavioral analysis and decision-making, (iii) explainable AI, (iv) generative AI, (v) deepfakes and psychologically manipulative content.

Publication window: Give preference to work published during the year 2022 and thereafter. The earlier work includes the foundations of cognitive bias and social engineering and consists of only necessary frameworks and concepts that are needed to understand the concepts at a basic level.

Inclusion criteria: Inclusion of peer-reviewed articles along with proceedings and established archives such as arXiv, cyberattack trends and studies on social engineering. Sources must include four of the following areas: psychological manipulation and social engineering, behavioral analysis, generative AI and LLMs, and deepfakes. This review paper will also include case studies, their effects and operational principles.

Exclusion criteria: This review paper excluded opinion-based sources, working mechanisms and case studies without proof, and theoretically based sources that lack integrity and are unverified by any trustworthy sources. Some content only provides a technical analysis of malware and does not include any psychological aspect.

Screening: First, titles and abstracts were screened for relevance. 45 articles were finally included in the synthesis.

Positioning statement: As per reviewer instructions, “Emotional Malware” will not be a newly established malware classification or technical standard that CISOs, security vendors, or technology standards organizations choose to adopt. The author proposes four elements as a framework for this concept. Emotional malware integrates attack types such as phishing, deepfakes, social engineering, and psychological manipulation. The framework proposes psychological mechanisms—biases, trust calibration, loss aversion, and the example-fear heuristic. Emotional malware occurs with or is integrated with the mechanisms within Cognitive bias, HCI, affective computing; psychology—Prospect theory; and Detection & governance—XAI, provenance, zero trust, legal/ethical.

1. Introduction

The recent virtual world developments, pushed by digital technology, cloud, social networks and AI, have irrevocably changed the rules of the game in cybersecurity. Until recently, malware targeted the system. Now, the digital threat is in the human mind. Until recently, hackers wanted to get unauthorised access to software, operating systems, and the network, but hackers figured out that any security system has "one critical link" that is related to a human's emotions and the decision-making process. Consequently, today hackers turn to psychological vulnerabilities. With the appearance of artificial intelligence, social engineering is more dangerous because now hackers can personalize their attacks even more to target a specific individual and control automated campaigns that manipulate emotions and eventually go around the existing security systems. The combination of artificial intelligence and psychological manipulation dramatically changes the nature of existing threats. The purpose of our paper is to construct the Emotional Malware theory. By this, we mean the "cyber threats that exclusively rely on the manipulation of human emotions and on social engineering and AI deception, and do not depend on the use of malware" (Gonzaga et al., 2026).

1.1 Research Gap and Contributions of This Review

There has been a surge of research on social engineering, phishing, deepfakes, AI-assisted scams, and psychological manipulation over the last few years. Often, research fails to conclude that the merging of generative AI and human emotion significantly increases the threat of social engineering for attackers. “Emotional Malware” framework provides a framework for the interplay of behavioral manipulation and automated threat-delivery systems. In contrast to the existing research that investigates phishing, social engineering, deepfake attacks, or AI-enhanced deception in isolation, this research integrates these closely interrelated areas into a comprehensive framework that positions human emotion at the core of contemporary cyber-assaults. Technical papers describe the mechanisms of malware and LLMs, as well as deepfake audio and video technologies. Behavioral Papers focus on cognitive bias, vulnerability, and decision-making; here we integrate both analyses, technical and psychological, as there is no unified framework that integrates psychological and AI-related analysis. (Pedersen et al., 2025)

2. Literature Review

2.1 From Technical to Human-Centered Threat Models

Early cybersecurity research I reviewed treated social engineering as an afterthought when compared to malware research. This was an early limitation of the research. The primary focus of this paper is the use of psychology. It includes the emotions of trust, curiosity, urgency, greed, etc. An early restriction of AI was that attackers had to find information manually from the digital world, learn many different languages, and early tools could not provide adequate answers

because rigid rules constrained them. As a result, more refined chatbots could not respond to more complex requests or attacks, and even though attackers could not launch wide-scale attacks, the damages were minor; therefore, these attacks were rarely studied. Earlier studies described ways to perform attacks, but could not describe why a particular emotion is associated with a specific attack and how this attack would be successful against a particular profile. Therefore, this review paper describes not only types of attacks that are technical and psychological in nature, but also describes the types of emotions and their relationship to decision-based attacks from a rationalistic perspective. This research by Zaoui et al. (2024) describes the social engineering domain of different types of attacks. It describes the environment, the situation, the medium, and the stages, along with the approaches, and therefore classifies attacks not only from a technical perspective but also from a non-technical view, from which the defense mechanisms may be understood. (Zaoui et al., 2024a)

2.2 AI-Enabled Social Engineering: Convergence of Technology and Psychology

Early cybersecurity research did not pay much mind to social engineering relative to malware. As a consequence, social engineering was far from being a priority for early cyber research. The purpose of this research paper is to present effective psychological mechanisms. Emotions such as trust, curiosity, and greed, for example, are included in this research. An earlier challenge regarding AI was that before an attacker could launch an attack against a victim, he had to manually research different ways to collect victim information from digital media and had to learn several different languages. Earlier tools were not able to provide an adequate solution because they were built by programmers who defined their rules; therefore, if a victim asked a question that was considered unusual, older chatbots were not able to provide an adequate reply; moreover, attackers had difficulties scaling an attack, resulting in minor damage. The originality of this research paper lies in the fact that it describes not only technical or psychological attack types, but also descriptions of the relationship between emotions and decision theory. An important contribution by Zaoui et al. (2024) and team focuses on a range of social engineering attacks. This paper describes the attack environment and situational context, media, stages, and the strategies used. Therefore, it categorizes defenses not only based on technical sophistication but also based on a non-technical perspective.

2.3 Cognitive Bias, Behavioural Cybersecurity, and Decision-Making

This topic focuses on the attacker's perspective versus the victim's perspective. This topic describes how the attacker relies on the flaws that exist within the human mind. "Optimism Bias" describes how users often do not perceive how attacks become more efficient and swift. This bias only proves to be true when they begin to learn the meaning of certain terms and understand how to gain security awareness. The phishing attacks of today will eliminate all the "overt phishing cues". This is done when AI generates highly personalized and realistic phishing emails. The strength of this study is that it will be released based on a review of methodological rigour and practical verification and will possess a solid theoretical foundation. The results of this study will therefore be reliable. Optimism Bias- which attack technology will be used will depend on the victim's manipulation and psychological bias. This bias is not part of any conscious risk assessment and occurs within the subconscious mind.

Users' optimism bias leads them to believe that this type of scam would never happen to them. So, in some cases, security awareness training is not effective. AI defeats training demonstrates how generative AI fails to detect techniques used in security awareness cue-based training that employees have learned in the past. In the past, training was cue-based, meaning that employees would check for mistakes based on pre-established instructions. Generative AI creates personalized, mistake-free responses and messages that are almost impossible to distinguish from original messages. (Waqas et al. 2023a) When technical cues were not present, the security team relied on behavioral analysis and decision-making. Still, due to a higher cognitive load and cognitive biases, the minds of the individuals made wrong decisions that resulted in the complete loss of the impact of the training. (Owen et al. 2024)

2.4 Deepfake Detection: Technical Progress and Persistent Gaps

Though detection methods are progressing rapidly, they still need to be analyzed for loopholes. State-of-the-art detectors still leave 45 to 50 percent of real-world deepfakes undetected (AUC performance). This gap in detection will not be temporary but structural and will be built into some methods. To eliminate this, we have Explainable AI (XAI). By using XAI, detection methods will be explained, but since it does not prove that it is appropriate to use at an advanced level, it will affect raw detection accuracy. Users can rely on technical detection, but will have to consider human behavior as

well. Security machine learning models will be achieved only with the help of advanced software and psychology. Explainable AI is the most important part of the fairness, accountability, and transparency (FAT) group of machine learning approaches and is often talked about in connection with deep learning. The benefits of Explainable AI include reduced risks and liabilities, increased trust, and greater acceptance and improvement of decisions.

The architecture of Explainable AI is based on the methods and approaches used to provide transparency and interpretability in machine learning models. This is a combination of the three components: a machine learning model, an explanation algorithm, and a user interface, which is built upon a variety of technologies and platforms, such as web applications and mobile apps, visualizations, and explanations based on the approaches of Explainable AI, such as feature importance and attribution.

Limits of XAI include computational complexity, limited scope, domain specificity, and lack of standardization and interoperability. (Ramadhani et al., 2024) Use cases of explainable AI can help reveal potential opportunities and challenges of this framework. (Ali et al., 2024) One of them is Computer Vision: Explainable AI techniques can help computer vision elucidate some of the most important factors in the co-occurrence and classification of images. (Salvi et al., 2023) For instance, explainable AI can identify and delineate regions of images that are most important in classifying objects and explain most of the prediction for a particular class. (Patel et al., 2023)

3. Mechanisms of Emotional Vulnerability

Emotion-based cyberattacks leverage emotional cognitive biases. As per dual process theory, emotional stimuli like fear, urgency, and greed lead to rapid, sub-optimal System 1 decision-making over rational System 2. This mode exhibits extreme bias towards optimism and authority. Affective Computing uses Sentiment analysis, which is NLP-based emotion detection. It was designed for mental health triage and customer service, and is now extensively used for adversarial purposes like manipulating writing style, detecting and responding to users' emotional states. It allows for effective, large-scale generation of adversarial content and trust calibration at scale. Generative AI will help augment the believability of deepfakes by generating all the supporting trust signals that we are typically used to trusting with ease. For example, trusting a big name is often sufficient. Loss Aversion in Prospect Theory explains the effectiveness of use cases based on deal-making and auctions, like threats based on winning an auction.

Example: If you receive a coupon that you won through a lucky draw as a prize and then get a discount on a product, this is a reward-based threat, and at that time the subconscious mind will activate so fast that the user is unable to do cross-verification. This theory applies because in the human mind the fear of loss is always more than the fear of gain. (Lee & John, 2025) The four lenses summary: Dual-Process Theory encompasses Decision-making; Affective Computing includes technology (the AI - based approach); HCI Trust Calibration considers Human-Computer Interaction (Acting without thinking); Prospect Theory includes Decision-making involving the fear of loss. (Somoray & Miller, 2023) This is not a complete model that defines the entire "Emotional malware" theory, but it classifies types of malware using the behavioral science approach. (Waqas et al., 2023b)

4. Core Concepts and Emotional Malware Mechanisms

4.1 Social Engineering and Psychological Manipulation

Psychological Principle	Description	Purpose in Emotional Malware	Representative Attack Example
Trust	Exploits the victim's confidence in a familiar person, brand, or institution	Obtain sensitive information or credentials without raising suspicion	Fake bank email; trusted-brand impersonation; HR verification scam
Authority	Impersonates an individual or organization with perceived authority	Encourage immediate compliance with malicious requests	CEO fraud; fake government notice; IT-support impersonation
Urgency	Creates time pressure to force rapid decisions	Reduces critical thinking and bypasses security awareness	"Your account will be suspended within 24 hours"
Fear	Triggers anxiety by	Manipulate victims into	Fake security alert;

	presenting threats or consequences	immediate, unverified action	ransomware warning; tax-penalty notice
Curiosity	Exploits the desire to discover unknown information	Persuades users to open malicious links or attachments	"See who viewed your profile"; "confidential document attached"
Greed	Offers unrealistic financial or personal rewards	Attract victims into fraudulent schemes	Lottery scam; cryptocurrency investment scam; fake prize notice
Empathy	Exploits compassion and emotional concern for others	Encourage financial assistance or disclosure of information	Fake medical emergency; charity-donation scam; family-emergency request

Table 1 Psychological principles exploited in emotion-driven cyberattacks(Zaoui et al., 2024b)

4.2 Emotional Triggers and Attack Techniques

Manipulation involves getting a person to take actions that will likely go against that person's best interests. This can be done by obtaining information, gaining access, or getting the person to do a specific action. Manipulation skills deal with people skills and not logic. There are three primary attack vectors: Profiling, Affective Personalization, and Multimodal Realism. Knowing someone's profession, interests, and activities helps one understand how to leverage that person's emotional weaknesses. Affective Personalization employs NLP to perform sentiment analysis used in the construction of contextual victim advice. Multimodal Realism uses deepfake technology to create audio-video to achieve a high level of trust. The focus is on the human element of thinking and decision-making. Most of the frameworks were built to combat software-based attacks and, consequently, built-in security solutions struggle to address the void created by emotion-related attacks; hence the need to integrate technical and psychological classification.

4.3 Conceptual Lifecycle Framework

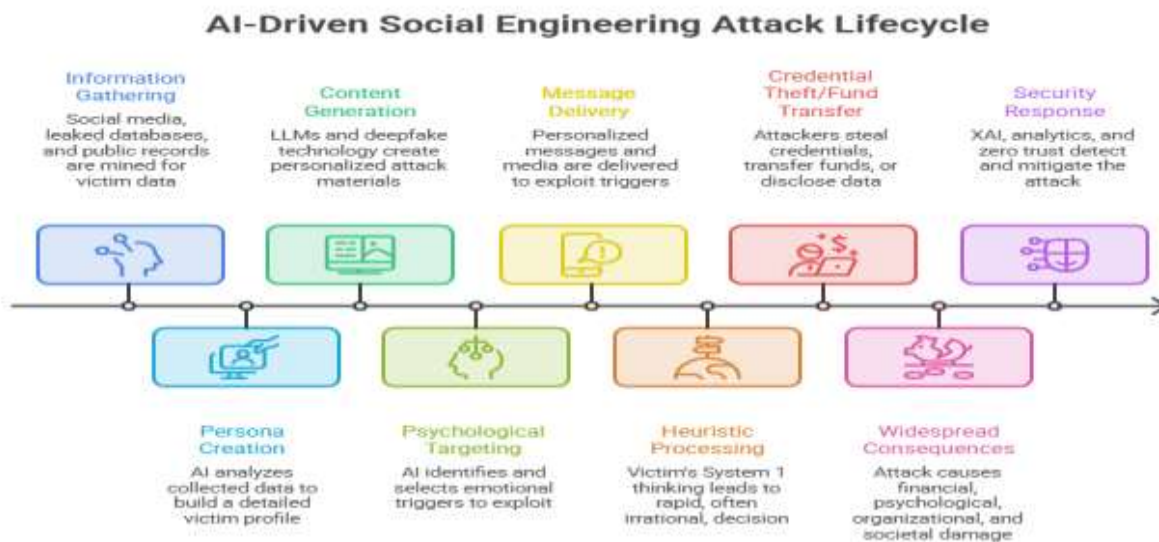


Figure 1 Conceptual lifecycle of an Emotional Malware attack

4.4 Positioning Relative to Adjacent Categories

Dimension	Traditional Malware	Classical Engineering	Social	AI-Enabled Deception (generic)	Emotional Malware (this framework)
Primary target	Software/system vulnerability	Human trust	trust	Perceptual/generative realism	Human emotion as the organizing variable across all of the above
Delivery mechanism	Code execution, exploits	Direct human contact (call, email)		Automated, AI-generated content	Channel-agnostic; classified by trigger,

				not medium (Section 4.4)
Psychological component	Largely absent	Central but under-theorized	Acknowledged but secondary to technical novelty	Central and theoretically grounded (Section 3)
AI involvement	Optional (e.g., AI-assisted malware generation)	Typically none	Central and definitional	Present but not required (cf. MGM case, Section 7.1)
Dominant detection paradigm	Signature/behavioural malware analysis	Awareness training	Content-level forensic detection	Layered: content, behavioural, provenance, and process controls (Section 5)
Academic treatment to date	Extensive, mature	Extensive but largely descriptive	Rapidly growing, technically focused	Proposed synthesis; not yet an established field

Table 2 Comparative positioning of the Emotional Malware framework against adjacent, more established categories. The table clarifies that Emotional Malware is not a replacement for these categories but a cross-cutting analytical lens applied to cases that fall within them.

This comparison describes that existing methods analyze which tool or vulnerability is used for an attack. Still, emotional malware analyzes which emotion is used for an attack and, therefore, it does not matter which tool the attacker uses. The MGM case example is the same as the classification of either classical social engineering or emotional malware.

5. Detection and Measurement Methods

Emotional malware is much harder to detect than software-based malware, and the reasons for this are that emotional malware does not have a signature, hash value, or string. Its success ratio is dependent on human emotion and behavior. Six categories define strengths, limitations, and open research questions as follows.

5.1 NLP-based sentiment and linguistic analysis

NLP-based sentiment analysis relies on chunks of text to interpret emotions such as fear, curiosity, urgency, trust, and greed. Sentiment Analysis is part of NLP and is used to determine emotions and attitudes behind opinions in linguistic data. Sentiment Analysis can be interpreted from text as positive, neutral, or negative on a given numeric scale (1-5). Emotion detection determines how text expresses joy, sadness, anger, fear, etc. Aspect-based sentiment analysis focuses on individual features of a product. Multilingual sentiment analysis is very complex due to differences in culturally expressed opinions and syntactical forms, especially in linguistic analysis. Intent-based sentiment analysis focuses on the context behind the sentiment in the text and is typically used in text summarization. Its primary advantage is Scalable, Cost-Effective Analysis and improved decision-making. There are, however, limitations: the domain of data is restricted, it can only handle neutral and mixed sentiments, and it is poor-quality data. (Hillman et al., 2023a)

5.2 Multimedia forensics and multimodal deepfake detection

Multimedia forensics assesses the context of multimedia signals, such as audio, video, and images. It analyzes and validates the integrity of content in its original form and identifies the device from which the content was captured. It also attempts to restore information from multimedia signals. It consists of facial inconsistency, audio authenticity verification, and cross-modal consistency.

It does this by matching both audio and video consistency and detecting deepfakes. Multimodal deepfakes use the realism of content. AI-based detection technology has progressed, but so too have deepfakes. Briefly, deepfakes utilize rapid evolution and generalization to create trust. False positives and negatives also constitute a challenge to detection.

5.3 Explainable AI (XAI).

Explainable AI is important for fairness and accountability. There are many advantages of explainable AI, including increased trust, improved decision-making, and a reduction of liability. Google is using XAI technologies in medical imaging, natural language processing, and computer vision. However, there are also some drawbacks. High computational complexity may limit its use. Use in different domains and applications may also be limited due to the different approaches to XAI and the lack of uniformity. Different XAI methods use diverse assessment criteria. (Mansoor & Iliev, 2025)

5.4 Behavioural analytics.

This looks at communication activity and transactions and sees abrupt shifts. The flexibility of it is that it can work with all domains, but in the absence of established workflows in a company or organization, this can produce false positives. (Zhuo et al., 2024) Behavioral analysis suggests that “something went wrong” and is not able to provide information about what went wrong and why the abnormal situation occurred. So the main question is how to minimize false positives while maintaining high detection accuracy. (Gallo et al., 2024)

5.5. Content provenance and digital watermarking.

There is cryptographically verifiable metadata that describes the origin of content, timestamps, and device information. The detection of this standard is based on the point in time the content was created. This technology has challenges, such as watermarks being removed by content owners or other watermarks. As of now, it is not an industry-wide standard. Content that is produced on external platforms and is not adjacent to participating platforms cannot be controlled. (Trad et al., 2024)

5.6. Zero-trust architectures.

Zero trust architecture operates on the principle of “never trust, always verify”. This extends beyond just the security of the network. This involves the safety of communication, transactions, and decisions. For example, if it receives any High-risk instruction, it takes that as a request for verification. A great strength of this architecture is that it operates independently, and for that reason, it would not verify if high authorities or any deepfake video instructed it to. Sometimes, it does work quite slowly and introduces labour productivity lag. The biggest problem is that a lot of organizations still do not adopt this, especially in high-impact crises. (Mushtaq et al., 2025)

Impact Dimension	Description
Individual Level	Beyond direct financial loss and identity theft, victims frequently report anxiety, depression, and diminished trust in digital platforms following exploitation (Shukla et al., 2023). Psychological recovery is an under-addressed dimension of cyber-incident response, which remains predominantly oriented toward technical remediation.
Organizational Level	Business Email Compromise (BEC) and executive-impersonation attacks target decision-makers directly, producing reputational damage, credential compromise, and direct financial loss—the Hong Kong deepfake CFO case (Section 7) being a paradigmatic example, alongside operational disruption illustrated by the MGM Resorts incident.
Societal Level	At scale, emotion-driven deception undermines public trust in digital communication generally—a phenomenon with implications well beyond individual fraud cases.
Misinformation and Digital Democracy	Deepfake technology capable of fabricating statements by public figures poses risks to electoral integrity and public discourse, since even debunked synthetic content can leave a residual "liar's dividend" effect, where genuine content becomes easier to dismiss as fake.
Public Confidence in Institutions	Attacks that impersonate government agencies or emergency-communication systems risk eroding legitimate institutional trust, with second-order effects on public compliance during genuine emergencies.
Privacy	Victim-profiling techniques (Section 4.2) depend on aggregating personal data from social

	media and breached databases, raising privacy concerns independent of any resulting fraud.
Ethical and Legal Dimensions	The same generative-AI capabilities that enable emotional malware (voice cloning, synthetic video) exist in a regulatory grey zone in most jurisdictions; content-provenance standards such as C2PA represent an emerging but not yet universally adopted governance response, and evidentiary frameworks for admitting or challenging synthetic media in legal proceedings are still being developed.
Governance and Regulatory Response	Emerging frameworks (e.g., deepfake-detection challenges convened by national bodies such as the UK Home Office, and forensic-standardization efforts such as NIST's open media-forensics evaluation) indicate growing institutional recognition of the problem, but current governance remains fragmented across jurisdictions and technology providers rather than harmonized.
Overall Significance	Collectively, these impacts justify treating Emotional Malware as an interdisciplinary governance challenge rather than a purely technical or purely individual-harm issue.

Table 2: Multidimensional Impact of Emotional Malware Across Individual, Organisational, Societal, and Governance Domains(Stalans et al., 2025)

7. Real-World Cases of Emotional Malware

Each case is analyzed using the same structure: background, attack methodology, emotional trigger, psychological mechanism, impact, lessons learned, and relevance to the framework to enable systematic comparison.

7.1 MGM Resorts Cyberattack (USA, 2023)

MGM Resorts International, one of the world's largest hospitality and entertainment companies, suffered a defining moment in the history of business continuity from cyberattacks in September 2023. Social engineering initiated by a single phone call became a \$100 million disaster affecting 29 company properties and resulting in 10 days of complete business disruption. Some major impacts were a \$100 million loss, 1a 10-dayduration, affecting 29 properties, hotels, and casinos, a loss of Personal data, and affecting millions of customers. The attack vector was social engineering via a phone call. The attack was executed in a series of steps. First, the attacker gathered the employee's personal data posted on LinkedIn. Next, the attacker called, posing as an employee and attempted to contact the help desk and convinced them to change the password. The help desk staff complied and provided the requested password. The attack was successful in exploiting trust and authority. This caused a disruption to operations, a financial loss to the company, and damaged the company's reputation. There were three main impacts. Operational impact: the company's hotels and casinos stopped operating completely. The company experienced a high financial loss. Technical impact: the attacker successfully bypassed the company's technical and physical controls and bypassed the firewalls, indicating that the attacker was bypassing the human security control. From the MGM Resorts Cyberattacks, there are four key takeaways regarding how to secure systems and business operations. The limitation of technical security: First, technical security mechanisms fail to eliminate social engineering attack vectors.

Organizational Trust Boundaries-Help desk and customer support services are also vulnerable to fraud attacks. Verification Hardening-the old protocols used for detecting and verifying voice will require an upgrade. Independence-The security team does not have to only focus on the IT infrastructure to be hardened, but also on the human level. (Schmitt & Flechais, 2024)

7.2 Hong Kong Deepfake CFO Fraud (2024)

A cyber incident occurred in January 2024 where deepfake technology (AI) was able to analyze pictures and videos of a person as well as their voice recordings and synthesize sophisticated forgeries of that person's likeness. They also set up a multi-participant deepfake video conference call, and for the rest of the participants, the AI also generated video conference participants. The tech employees on the call were persuaded using trust, authority, and urgency, and the audio and video appeared so authentic that it was hard to cross-verify. As a result, US\$25 million was transferred by that employee. From Hong Kong Deepfake CFO Fraud, we have the following Key Lessons: Visual/Auditory Verification- Sometimes what we see and hear is not the truth. It can also be a part of deepfake technology. Zero-Trust

Principle- Independent verification is important, and use Least-privilege access. High-Value Instructions- Be cautious when high-value instructions are given as a powerful authority is standing in front; then scrutinize those instructions too. (Pedersen et al., 2025b)

7.3 AI-Generated Celebrity Deepfake Investment Scams (2025)

AI-based deepfake scams allow celebrities to promote fictitious investment schemes by appearing in deepfake videos. Social media has the means to circulate videos on a global scale. The duo of trust and greed entices people to invest money and gets them a good return by using celebrity endorsement and advertisement videos. This leads to immense financial loss for several countries. This attack reveals key information: user awareness does not fully resolve the issue, while moderation at the platform level and verification of provenance are equally important. This attack does not stop at the corporate level, but has extended to the societal level and has created a situation known as the “Liar's Dividend,” where the actual content or evidence appears to be fake. (Alanazi et al., 2025)

7.4 Patna Deepfake Extortion Case (India, 2026)

An AI-based deepfake was used to blackmail a hotel owner in Patna, India, by demanding a payment of ₹97 lakh. Police have filed an FIR against three people, including a woman. The blackmailer threatened to circulate the fabricated videos and allegedly filed false criminal cases. The hotel owner paid ₹97 lakh in four separate payments at different locations in Muzaffarpur, Sitamarhi and Patna. The blackmailer has used deep emotions of fear and shame. Humans have social reputation fear and social standing anxiety; by targeting these emotions, the attacker creates fear in the victim's mind and makes them so isolated that they lose the power of rational thinking. Whether the content is correct or not is irrelevant.

To avoid social disruption, the victim moves the money and learns crucial lessons. As attackers employ new tricks (including new emotions and new technologies and media) to manipulate victims, perpetrators use the new procedures to exploit their targets. Detection Gap: Most of the older detection mechanisms do not allow assessing if evidence is original. Legal Gap: There has to be a robust victim support mechanism that aids the victim. Need for fast takedown mechanisms that erase the fake content posted on the internet (Pedersen et al., 2025c).

Year	Incident	Primary Emotional Trigger(s)	Attack Technique	Key Impact
2023	MGM Resorts Cyberattack (USA)	Trust, Authority	Help-desk engineering, credential compromise	Major operational and financial loss
2024	Hong Kong Deepfake CFO Fraud	Trust, Authority, Urgency	AI-generated deepfake video-conference impersonation	~US\$25 million fraudulently transferred
2025	AI-Generated Celebrity Investment Scams	Trust, Greed	Deepfake endorsement videos of public figures	Large-scale, cross-jurisdictional financial victimization
2026	Patna Deepfake Extortion Case (India)	Fear, Shame	AI-generated non-consensual explicit deepfake content	Reported extortion payment of ~₹97 lakh

Table 3. Comparative analysis of real-world Emotional Malware incidents (2023–2026), classified using the four-dimensional taxonomy in Section 4.4.

9. Prevention

Most security awareness training (SAT) is focused on behavioral prevention. Behavioral prevention is the teaching and awareness of spelling and grammatical errors, tone, and poor threats. While some may use SAT to teach, this method will never fix issues related to power or decision-making in high-risk scenarios. SAT methods are used to generate rules and do not address cognitive biases. Authority, positivity, and a lack of questioning are just a few of the cognitive biases

that SAT methods ignore. One of the main issues is AI-generated content, as this is often used to train employees on something that is used to recognize cues. The main focus should be on cognitive-bias inoculation. The aim is not to allow cognitive biases to dictate decision-making, but rather to recognize emotional and cognitive vulnerabilities, not just attacker tactics. “Never Trust, Always Verify” is the essence of Zero Trust and Verification, based on the principle of “Least Privilege Access” and “Continuous Monitoring and Microsegmentation.” In traditional security models, everything in the trusted network is accepted as secure. However, the Zero Trust model requires user identification and verification even outside the internal network. Lastly, Technical Prevention includes Content Provenance and Watermarking. This includes, on a large scale, the verification of content and media authenticity. Forensic detection or even AI-based biometric analysis is used to perform anomaly detection on audio, video, or even images to discover what artifacts, micro-expressions, unnatural patterns of blinking, or voice frequency anomalies are hidden in the data. It includes all generated content from harmful tools or from outside provenance-enabled platforms. When combined with provenance and zero-trust, it creates a multilayered defence system (Hillman et al., 2023b).

9. Conclusion

This paper provides an initial conceptualization of emotional malware. It combines psychological manipulation, personalization aided by artificial intelligence (AI), and social engineering. It targets the trust of individuals and subsequently their confidence in their decision-making, so that the attack surface is the human population. The level of success in an emotional malware attack is dependent on the strength of the emotional manipulation and the use of emotional triggers. Examples of present-day attack scenarios are the recent breaches of MGM Resorts and deepfake financial frauds. In today’s world, human-centered cybersecurity is of utmost importance. Generative AI is showing exponential growth, making existing detection methods obsolete. Emotional malware is designed to employ new attack surfaces and techniques. A layered approach to security is needed to identify and counter real-world threats and attacks with the help of AI and machine learning. Content provenance and zero-trust verification should be implemented in every organization or individual. There are five impact dimensions: Organizational, Psychological, Legal, Ethical, and Societal. It encourages the use of an interdisciplinary approach which integrates Cybersecurity, Psychology, Behavioral Science, HCI, and Design. It is critical to understand the future research directions and their associated challenges. Empirical Validation: Formalize the Emotional Malware framework through mathematical and practical trials. The evaluation of emotional trigger analysis and detection assesses the success of content generation based on AI and examines explainable AI (XAI) and multimodal detection. Victim Recovery helps people heal after emotional abuse and scams. Conduct a comparative analysis and restructure Security Awareness Training from cue-based to cognitive-bias inoculation.

"While it is vital to improve the detection of emotion-driven cyber threats, we must consider the cognitive vulnerabilities of humans, which will improve our resilience to those attacks."

References:

- Alanazi, S., Asif, S., Caird-Daley, A., & Moulitsas, I. (2025). Unmasking deepfakes: a multidisciplinary examination of social impacts and regulatory responses. *Human-Intelligent Systems Integration*, 2025 7:1, 7(1), 131–153. <https://doi.org/10.1007/S42454-025-00060-4>
- Ali, G., Rashid, J., Ul Hussnain, M. R., Tariq, M. U., Ghani, A., & Kwak, D. (2024). Beyond the Illusion: Ensemble Learning for Effective Voice Deepfake Detection. *IEEE Access*, 12, 149940–149959. <https://doi.org/10.1109/ACCESS.2024.3457866>
- Gallo, L., Gentile, D., Ruggiero, S., Botta, A., & Ventre, G. (2024). The human factor in phishing: Collecting and analyzing user behavior when reading emails. *Computers & Security*, 139, 103671. <https://doi.org/10.1016/J.COSE.2023.103671>
- Gonzaga, K., Serra, S., Gomes, M., & Malta, S. (2026). AI-Powered Social Engineering: Emerging Attack Vectors, Vulnerabilities, and Multi-Layered Defense Strategies. *Computers*, 2026, Vol. 15, Page 128, 15(2), 128. <https://doi.org/10.3390/COMPUTERS15020128>
- Hillman, D., Harel, Y., & Toch, E. (2023a). Evaluating organizational phishing awareness training on an enterprise

- scale. *Computers & Security*, 132, 103364. <https://doi.org/10.1016/J.COSE.2023.103364>
- Hillman, D., Harel, Y., & Toch, E. (2023b). Evaluating organizational phishing awareness training on an enterprise scale. *Computers & Security*, 132, 103364. <https://doi.org/10.1016/J.COSE.2023.103364>
- Lee, J., & John, K. K. (2025). Detecting Deepfakes Through Emotion?: Facial Expression and Emotional Contagion as Dual Indicators of Deepfake Credibility. *Applied Cognitive Psychology*, 39(6), e70141. <https://doi.org/10.1002/ACP.70141>
- Mansoor, N., & Iliev, A. I. (2025). Explainable AI for DeepFake Detection. *Applied Sciences*, 2025, Vol. 15, Page 725, 15(2), 725. <https://doi.org/10.3390/APP15020725>
- Mushtaq, S., Mohsin, M., & Mushtaq, M. M. (2025). A Systematic Literature Review on the Implementation and Challenges of Zero Trust Architecture Across Domains. *Sensors*, 2025, Vol. 25, Page 6118, 25(19), 6118. <https://doi.org/10.3390/S25196118>
- Owen, M., Flowerday, S. V., & van der Schyff, K. (2024). Optimism bias in susceptibility to phishing attacks: an empirical study. *Information and Computer Security*, 32(5), 656–675. <https://doi.org/10.1108/ICS-02-2023-0023>
- Patel, Y., Tanwar, S., Gupta, R., Bhattacharya, P., Davidson, I. E., Nyameko, R., Aluvala, S., & Vimal, V. (2023). Deepfake Generation and Detection: Case Study and Challenges. *IEEE Access*, 11, 143296–143323. <https://doi.org/10.1109/ACCESS.2023.3342107>
- Pedersen, K. T., Pepke, L., Stærmose, T., Papaioannou, M., Choudhary, G., & Dragoni, N. (2025a). Deepfake-Driven Social Engineering: Threats, Detection Techniques, and Defensive Strategies in Corporate Environments. *Journal of Cybersecurity and Privacy*, 2025, Vol. 5, Page 18, 5(2), 18. <https://doi.org/10.3390/JCP5020018>
- Pedersen, K. T., Pepke, L., Stærmose, T., Papaioannou, M., Choudhary, G., & Dragoni, N. (2025b). Deepfake-Driven Social Engineering: Threats, Detection Techniques, and Defensive Strategies in Corporate Environments. *Journal of Cybersecurity and Privacy*, 2025, Vol. 5, Page 18, 5(2), 18. <https://doi.org/10.3390/JCP5020018>
- Pedersen, K. T., Pepke, L., Stærmose, T., Papaioannou, M., Choudhary, G., & Dragoni, N. (2025c). Deepfake-Driven Social Engineering: Threats, Detection Techniques, and Defensive Strategies in Corporate Environments. *Journal of Cybersecurity and Privacy*, 2025, Vol. 5, Page 18, 5(2), 18. <https://doi.org/10.3390/JCP5020018>
- Ramadhani, K. N., Munir, R., & Utama, N. P. (2024). Improving Video Vision Transformer for Deepfake Video Detection Using Facial Landmark, Depthwise Separable Convolution and Self-Attention. *IEEE Access*, 12, 8932–8939. <https://doi.org/10.1109/ACCESS.2024.3352890>
- Salvi, D., Liu, H., Mandelli, S., Bestagini, P., Zhou, W., Zhang, W., & Tubaro, S. (2023). A Robust Approach to Multimodal Deepfake Detection. *Journal of Imaging*, 2023, Vol. 9, Page 122, 9(6), 122. <https://doi.org/10.3390/JIMAGING9060122>
- Schmitt, M., & Flechais, I. (2024). Digital deception: generative artificial intelligence in social engineering and phishing. *Artificial Intelligence Review*, 2024 57:12, 57(12), 324-. <https://doi.org/10.1007/S10462-024-10973-2>
- Somoray, K., & Miller, D. J. (2023). Providing detection strategies to improve human detection of deepfakes: An experimental study. *Computers in Human Behavior*, 149, 107917. <https://doi.org/10.1016/J.CHB.2023.107917>
- Stalans, L. J., Chan-Tin, E., Moran, M., Hart, A., & Pardonek, J. (2025). Being phished and reporting phishing emails: emotional and knowledge antecedents and training gaps. *Journal of Crime and Justice*, 48(4), 533–551. <https://doi.org/10.1080/0735648X.2024.2415321>
- Trad, F., Semaan-Nasr, E., & Chehab, A. (2024). MLPhishChain: A machine learning-based blockchain framework for reducing phishing threats. *Frontiers in Blockchain*, 7, 1484894. <https://doi.org/10.3389/FBLOC.2024.1484894/TEXT>
- Waqas, M., Hania, A., Yahya, F., & Malik, I. (2023a). Enhancing Cybersecurity: The Crucial Role of Self-Regulation, Information Processing, and Financial Knowledge in Combating Phishing Attacks. *SAGE Open*, 13(4). <https://doi.org/10.1177/21582440231217720>
- Waqas, M., Hania, A., Yahya, F., & Malik, I. (2023b). Enhancing Cybersecurity: The Crucial Role of Self-Regulation, Information Processing, and Financial Knowledge in Combating Phishing Attacks. *SAGE Open*, 13(4). <https://doi.org/10.1177/21582440231217720>
- Zaoui, M., Yousra, B., Yassine, S., Yassine, M., & Karim, O. (2024a). A Comprehensive Taxonomy of Social Engineering Attacks and Defense Mechanisms: Toward Effective Mitigation Strategies. *IEEE Access*, 12, 72224–72241. <https://doi.org/10.1109/ACCESS.2024.3403197>
- Zaoui, M., Yousra, B., Yassine, S., Yassine, M., & Karim, O. (2024b). A Comprehensive Taxonomy of Social

Engineering Attacks and Defense Mechanisms: Toward Effective Mitigation Strategies. *IEEE Access*, 12, 72224–72241. <https://doi.org/10.1109/ACCESS.2024.3403197>

Zhuo, S., Biddle, R., Recomendable, J. D., Russello, G., & Lottridge, D. (2024). Eyes on the Phish(er): Towards Understanding Users' Email Processing Pattern and Mental Models in Phishing Detection. *ACM International Conference Proceedings Series*, 1, 15–29.

<https://doi.org/10.1145/3688459.3688465>;GROUPTOPIC:TOPIC:ACM-PUBTYPE