Social engineering in Ukraine: threats and intelligent detection approaches[*]

Victoria Vysotska[1,†], Kirill Smelyakov[2,†], Anastasiya Chupryna[2,†], Dmytro Darahan[2,*,†], Oleh Torubara[2,†] and Oleh Shyshymenko[3,†]

[1] *Kharkiv National University of Internal Affairs, L. Landau Avenue 27 61080 Kharkiv, Ukraine*
[2] *Kharkiv National University of Radio Electronics, 14 Nauky Ave., Kharkiv, 61166, Ukraine*
[3] *National University "Odesa Law Academy", Fontanska doroga, 23, Odesa , Ukraine*

Abstract

This research provides a technical and legal analysis of social engineering threats in Ukraine's security landscape, where digital attacks have become weaponized instruments of hybrid warfare. The study examines how state-sponsored actors exploit technical vulnerabilities through AI-enhanced phishing and deepfakes, while operating within legal gray zones of encrypted platforms. Was analyzed the convergence of criminal infrastructures—including illicit gambling and drugs distribution—with information operations, creating complex jurisdictional challenges. A number of modern intelligent detection method was considered. On the basis of literature research was proposed Hybrid AI Detection Framework addresses these dual challenges through hybridization of transformer architecture and fuzzy deontic inference. Such an approach enables simultaneous technical detection and legal assessment of sophisticated social engineering campaigns, providing a comprehensive solution for modern cybersecurity and law enforcement applications in conflict-affected jurisdictions.

Keywords

software engineering, cybercrime, law enforcement, deep learning, artificial immune systems, fuzzy deontic logic, natural language processing, machine learning.

1. Introduction

The digital threat environment in Ukraine presents a critical and rapidly evolving security challenge, where social engineering has transcended conventional cybercrime to become a weaponized instrument of hybrid warfare. This multifaceted threat landscape is characterized by the deliberate convergence of hostile state-sponsored subversion with entrenched criminal enterprises, including illicit gambling, narcotics distribution, and large-scale fraud [1]. The ongoing conflict has accelerated this fusion, creating a complex ecosystem where attacks are designed to exploit human psychology not merely for profit, but to destabilize social structures, erode institutional trust, and undermine national resilience. This strategic dimension elevates social engineering beyond a technical nuisance to a paramount national security concern, establishing Ukraine as a pivotal case study in contemporary information warfare and necessitating the urgent development of sophisticated, intelligent countermeasures [2]. The most significant threat evolution lies in the orchestration of social engineering campaigns by Russian state-aligned actors. These operations are psychologically sophisticated and strategically targeted, aiming to demoralize the civilian population and compromise military personnel.

Tactics include the systematic manipulation of families of prisoners of war through coordinated disinformation and the sophisticated recruitment of vulnerable civilians for acts of sabotage. These campaigns are executed through encrypted messaging platforms and social networks, leveraging

[*] *CIAW-2025: Computational Intelligence Application Workshop, September 26-27, 2025, Lviv, Ukraine*

[*] Corresponding author.

[†] These authors contributed equally.

victoria.a.vysotska@lpnu.ua (V. Vysotska); kyrylo.smelyakov@nure.ua (K. Smelyakov); anastasiya.chupryna@nure.ua (A. Chupryna); dmytro.darahan@nure.ua (D. Darahan); oleh.torubara@nure.ua (O. Torubara); shishimenko228@gmail.com (O. Shyshymenko)

0000-0001-6417-3689 (V. Vysotska); 0000-0001-9938-5489 (K. Smelyakov); 0000-0003-0394-9900 (A. Chupryna); 0000-0003-4918-4579 (D. Darahan); 0009-0002-0933-0754 (O. Torubara); 0009-0004-5087-124X(O. Shyshymenko).

personalized approaches that build false trust over time. The objective is to create a pervasive sense of uncertainty and internal division, making these operations a core component of modern hybrid conflict that directly targets the human element as the most vulnerable component of any defense system.

Parallel to these state-sponsored activities, and often intersecting with them, is the aggressive expansion of criminal ecosystems that thrive on social engineering. The poorly regulated digital advertising of online gambling and the proliferation of illicit drug markets on platforms like Telegram have created profound social vulnerabilities [4]. These criminal infrastructures are not isolated; they are increasingly exploited as vectors for broader subversive activities. Organized groups use the promise of financial gain to recruit individuals from vulnerable demographics, gradually luring them from minor infractions into serious criminal acts, including espionage and sabotage. This creates a dangerous synergy where criminal profit and state-sponsored destabilization activities become inextricably linked, each amplifying the impact of the other.

The imperative to pioneer and implement advanced intelligent detection approaches has never been more pressing. The complex, adaptive, and multi-vector nature of the social engineering threat landscape in Ukraine demonstrates that reactive and siloed solutions are insufficient. The aim of our study is to research modern state of social engineering in Ukraine, study its threats and characteristic features. Our main goal is to analyze existed artificial intelligence methods and aproaches of detection of such malicious activities. Performed analysis should help identify a framework for research and development of hybrid adaptive software system, that is capable to support proactive approach in cybercrime detection and law enforcement [3].

2. Related works

2.1. Legal landscape

Information technology has become a common tool for committing criminal offences. Compared to traditional types of crime in Ukraine, cybercrime is a relatively new phenomenon and at the same time one of the greatest threats of the 21st century. Analysing criminal activity, there is a noticeable trend towards shifting criminal activity from the real world to the digital environment.

According to statistics from the Department of the Cyberpolice of the National Police of Ukraine, 2,500 cybercrimes were registered in Ukraine in 2024 that caused losses of UAH 396.7 million.

The most common threats in 2024 were: phishing, illegal content (glorification of extremism, terrorism, drug addiction, creation and distribution of pornography, including child pornography, violence against children (online bullying, grooming, "death communities"), the cult of cruelty and violence, copyright infringement), online fraud, database theft, interference with websites (cyberattacks). Negative factors affecting the crime situation in Ukraine's cyberspace include the russian federation's armed aggression against Ukraine, the spread of global cyber threats such as supply chain attacks, data breaches, the creation of new ways of committing criminal offences that are not yet covered by legal regulation, in particular with usage of artificial intelligence and the dissemination of personal data, which facilitates the commission of cybercrimes[5]. The main method of commitment aforementioned crimes is social engineering, thus its detection could help to prevent and reduce damage damage caused by the described illegal activities.

It should be noted that the development of cybercrime is facilitated by the lack of regulation of certain deeds that contain elements of a criminal offence at the legislative level. In particular, advertising of narcotic drugs and psychotropic substances is formally prohibited by Article 35 of the Law of Ukraine "On Narcotic Drugs, Psychotropic Substances and Precursors", but administrative or criminal liability for this act is not provided for by the current Criminal Code. The relevant draft law "On Amendments to the Criminal Code of Ukraine Regarding the Criminalisation of Advertising or Propaganda of Narcotic Drugs, Psychotropic Substances, Their

Analogues or Precursors" No. 5496 was sent to the Verkhovna Rada of Ukraine, but the draft was never considered by parliament.

The introduction of artificial intelligence technologies that can transform the activities of law enforcement agencies in the field of combating cybercrime, from criminal analytics, which will allow identifying trends in huge data sets, to biometrics, which allows quickly identifying criminals, can strengthen the capacity to counter cybercrime. AI tools can optimise decision-making processes at both the operational and strategic levels, enabling law enforcement agencies to better detect and eliminate criminal threats at the preventive stage[6].

For example, with the help of AI-based analysis, investigators can analyse large amounts of information to identify anomalous patterns in banking, telecommunications, online activity, and illegal online marketplaces, and monitor extremist content on social platforms.

In particular, the UK police already use AI in areas such as back office/business support functions, risk management of warrants, facial recognition, redaction, forensic analysis of data, intelligence and demand forecasting, resource allocation (officer and vehicle), intelligence (facial recognition, uncovering hidden links, mapping), performance optimisation (optimising investigative timelines), risk reduction, and data bias[7].

In Ukraine, the use of AI tools in law enforcement is not yet widespread. In 2021, the Minister of Justice of Ukraine announced that Ukraine would begin using Cassandra, AI software that analyses the likelihood of repeat offences by criminals. The purpose of the programme was to assist probation officers in preparing a "pre-trial report" describing the personality of the accused and assessing the likelihood of them reoffending.

Using foreign experience, it would be appropriate to introduce an auxiliary system with AI elements in Ukraine, which would make it possible to analyse cyberspace and identify possible cybercrimes for the effective prevention of criminal offences using information technology.

2.2. Features of cyberthreats

The landscape of cyberthreats is continuously evolving, with social engineering representing a particularly insidious challenge because it exploits human psychology rather than technical vulnerabilities. These attacks manipulate human emotions like fear, urgency, curiosity, and trust to deceive individuals into divulging sensitive information, granting system access, or transferring funds [8]. With the advent of artificial intelligence (AI), the nature of these attacks has transformed; AI is now weaponized to create more personalized and convincing campaigns. Conversely, AI-powered defensive tools are becoming essential for identifying and mitigating these advanced threats by analyzing vast datasets to spot subtle, machine-generated anomalies that would be imperceptible to the human eye.

At the core of every social engineering attack is a calculated appeal to human emotion. Attackers craft scenarios designed to provoke a swift, unthinking reaction, bypassing logical scrutiny. Common emotional triggers include a fear of losing money or account access, the urgency of a superior's request, the allure of a too-good-to-be-true offer, or curiosity about a mysterious link or file. These psychological levers are employed across a spectrum of attack types, each with its own methodology but sharing the same fundamental principles of deception.

Phishing, one of the most prevalent forms, involves attackers impersonating trusted entities via email, SMS, or voice calls to steal credentials or deliver malware. A more targeted variant, spear phishing, uses personalized information gathered from sources like social media to craft highly convincing messages tailored to a specific individual or organization [9]. This approach is further refined in whaling attacks, which focus exclusively on high-level executives to authorize fraudulent financial transactions [9]. Another sophisticated technique, Business Email Compromise (BEC), sees attackers impersonate executives to manipulate employees into initiating wire transfers or changing banking details, often relying on a forged but convincing narrative and causing massive financial losses.

Other common techniques include baiting, which lures victims with promises of free goods or services that instead lead to malware download or information theft, and scareware, which bombards users with false alarms about non-existent infections to frighten them into installing malicious software or paying ransoms [10]. Pretexting involves building a fabricated scenario where the attacker, posing as an authority figure like an IT support technician or law enforcement officer, establishes a false sense of trust to extract sensitive data [10]. Quid pro quo attacks offer a service, such as fake IT support, in exchange for login credentials or other critical information.

In addition to traditional malicious techniques, artificial intelligence has dramatically amplified the scale, efficiency, and persuasiveness of social engineering attacks. Generative AI and large language models (LLMs) can now produce phishing emails with impeccable grammar and style, eliminating the spelling errors and awkward phrasing that were once key red flags. These AI tools can scrape and analyze public data from social media, company websites, and data breaches to build detailed profiles of targets, enabling a level of personalization previously only possible through labor-intensive manual research. This allows attackers to reference a target's colleagues, recent projects, or personal interests, making fraudulent communications appear highly credible.

Furthermore, AI facilitates multi-vector attacks that can be coordinated across email, voice, and text simultaneously, increasing the chances of success. The most alarming advancement is the use of deepfakes and real-time voice cloning. Attackers can now create highly realistic video and audio impersonations of company executives, which can be used in video calls or voice messages to lend an undeniable layer of authenticity to their fraudulent requests. This technology significantly shortens the time required to execute a sophisticated campaign and makes verification through traditional means exponentially more difficult.

Despite their growing sophistication, AI-powered social engineering attacks exhibit distinct characteristics that can be identified by advanced AI defense systems. These systems leverage behavioral analysis, anomaly detection, and natural language processing to differentiate between legitimate and malicious communications.

Robust detection algorithms depends on linguistic and behavioral features of analyzed text. AI security tools can be trained to analyze the writing style and metadata of incoming emails. While generative AI can mimic human prose, it may still produce text with a consistent, "manufactured" tone or fail to perfectly replicate the unique and varied writing patterns of a specific individual over time. Defensive AI can flag messages that deviate from an established user's baseline communication style. Furthermore, AI can detect the emotional manipulation tactics embedded in the text, such as heightened urgency, fear, or offers that seem too good to be true, even when the grammar is flawless [11].

Behavioral Content Features focus on the manner in which users and systems interact with content, building a dynamic baseline of normal activity to spot anomalies. This is often operationalized through User and Entity Behavior Analytics (UEBA) [12]. UEBA systems use machine learning to analyze thousands of data points around user activity, such as typical login times, access patterns, the sensitivity of data normally interacted with, and communication habits. By establishing this behavioral baseline, AI can identify deviations that suggest compromise, such as a user suddenly accessing large volumes of sensitive data they never use, logging in from unusual locations, or sending emails at odd hours.

Another critical area is network features. AI systems monitor for anomalies in communication patterns, such as emails originating from spoofed domains that are visually similar to legitimate ones but have subtle differences in the URL. They can also analyze the digital fingerprints of messages and websites, identifying indicators that suggest automated generation or association with known phishing infrastructures [13]. For deepfakes and cloned audio, specialized AI detection tools are being developed to identify subtle digital artifacts, inconsistencies in lighting, blinking patterns, or vocal cadences that are not present in genuine human recordings.

A third component of defense is authentication and protocol features analysis. AI can help enforce and monitor multi-factor authentication (MFA), which remains a critical barrier even if credentials are stolen. More advanced systems are exploring the use of behavioral analytics to differentiate between legitimate users and fraudsters based on patterns of use, such as login times, locations, and typical actions. AI can also monitor for behavioral anomalies in how protocols are used, such as login attempts from unusual geographical locations or at strange times, which could indicate compromised credentials.

Multimedia Content Analysis represents a critical evolution in threat detection, moving beyond text to analyze images, video, and audio. Modern detection systems employ AI-powered featurization services to automatically extract and analyze attributes from visual and auditory content. This involves breaking down multimedia elements into quantifiable data points, such as color composition, emotional tone of images, scene background, and aesthetic qualities. For deepfakes and synthetic media, specialized models conduct a granular analysis of video frames to identify subtle digital artifacts, inconsistencies in lighting, unnatural facial micro-movements, or blinking patterns that are not present in genuine human recordings. Similarly, audio analysis in the time, frequency, and cepstral domains can detect synthetic voice clones by identifying unnatural vocal cadences and emotional inconsistencies. This multi-modal analysis is essential as adversaries increasingly use AI-generated images and videos for sophisticated phishing and disinformation campaigns.

The following Table 1 summarizes key social engineering threats and the specific characteristics that AI systems can be trained to detect.

Table 1 Types of Social Engineering attacks and their characteristics

Threat Type	Analysis Focus	AI-Detectable Characteristics
Spear Phishing & Whaling	Behavioral & Contextual	Highly personalized content from social media scraping, impersonation of executives,
Phishing	Linguistic & Network	Spoofed sender addresses, urgency/fear language, malicious links
Smishing & Vishing	Protocol & Content	SMS messages with fraudulent links, calls from spoofed numbers, requests for credentials
Baiting	Content & Network	"Too good to be true" offers, ads for free downloads leading to malicious sites
Pretexting	Linguistic & Behavioral	Fabricated scenarios to build false trust, impersonation of authority figures (IT, police, bank)
Business Email Compromise (BEC)	Behavioral & Contextual	Email account impersonation of high-level executives, requests for urgent wire transfers to new accounts,
Deepfakes	Multimedia Analysis	Visual artifacts, unnatural facial movements or blinking, synthetic voice patterns

2.3. Threat detection aproaches

Traditional security methods relying on predefined rules struggle to maintain efficiency against adaptive, continuously evolving cyber threats. ML-powered systems address this limitation by leveraging adaptive learning and advanced pattern recognition to analyze massive datasets for subtle anomalies that human analysts or rule-based filters often miss. This analysis involves two primary analytical tracks: semantic (or linguistic) and structural.

Modern detection systems using advanced DL models, such as Transformers (e.g., BERT or RoBERTa) [14], often integrate semantic and structure feature extraction implicitly. These architectures leverage internal mechanisms, such as attention, to generate highly contextualized word embeddings, meaning they perform sophisticated feature engineering during the forward pass. This context awareness allows the model to differentiate between subtle meanings.

Semantic analysis focuses on the meaning, intent, and psychological frame of the language, essential for detecting social engineering and hate speech. Structural analysis, conversely, focuses on non-textual attributes, such as the composition of a URL, network traffic logs, or account metadata. The most effective detection frameworks must incorporate parallel feature processing streams—for instance, utilizing a 1D Convolutional Neural Network (CNN) for efficient structural analysis of a URL string, while concurrently employing a Transformer model for deep linguistic context analysis of the accompanying text payload.

Classical models remain relevant for specific tasks, especially where computational efficiency or reliable binary classification is prioritized.

The Support Vector Machine (SVM) is highly effective for binary text classification, demonstrating strong performance in generalized crime detection [15]. Its primary limitation arises in complex, imbalanced multi-class problems, although modifications like Cost-Sensitive SVM can address this by weighting the cost of misclassification.

Ensemble Methods are frequently utilized due to their robustness and ability to manage noisy data. Random Forest (RF) is consistently identified as a frequently employed and reliably effective classification algorithm in cyber threat analysis [16]. RF is particularly suitable for multi-class classification involving imbalanced data, achieving strong accuracy by aggregating results from multiple decision trees. Similarly, AdaBoost has demonstrated high effectiveness and precision when applied to cyber-threat detection using real-time datasets. Other models, such as Naïve Bayesian and K-Nearest Neighbor (KNN), typically serve as computational baselines in comparative studies.

Deep learning models are crucial for capturing the non-linear relationships and intricate sequential dependencies inherent in modern threats, exceeding the capability of classical ML approaches.

Convolutional Neural Networks (CNNs) excel at extracting local patterns and sequential motifs from data. The 1D CNN architecture is especially efficient for processing fixed-format sequential data like URL strings or network logs [17], allowing for fast, low-latency analysis.

Recurrent Neural Networks (RNNs) and their variants, such as Long Short-Term Memory (LSTM) and Gated Recurrent Unit (GRU) networks, are designed to handle long-range dependencies in textual data, making them crucial for semantic understanding in tasks like hate speech classification [18]. Often, superior performance is achieved through Hybrid Architectures[19], such as the Multichannel RCNN (combining CNN and RNN components), which synthesizes the feature extraction power of CNNs with the sequential modeling strengths of RNNs [20]. Furthermore, a hybrid DL approach, the AutoEncoder + XGBoost (AEXB) Model, uses an AutoEncoder for robust feature extraction and dimensionality reduction, followed by XGBoost for rapid classification.

Evolutionary algorithms provide a bio-inspired paradigm for developing adaptive detection systems that evolve to recognize novel threats. A prominent approach is the Artificial Immune System (AIS) [21], which emulates the natural immune system's ability to distinguish between self and non-self. Early AIS models, however, faced challenges with premature convergence. The integration of Danger Theory [22] significantly enhances these models by shifting the detection paradigm from a rigid self/non-self distinction to a more contextual response to "danger signals" indicative of system harm. This led to the development of advanced algorithms like dt-aiNet, which incorporates a danger zone to modulate antibody concentrations.

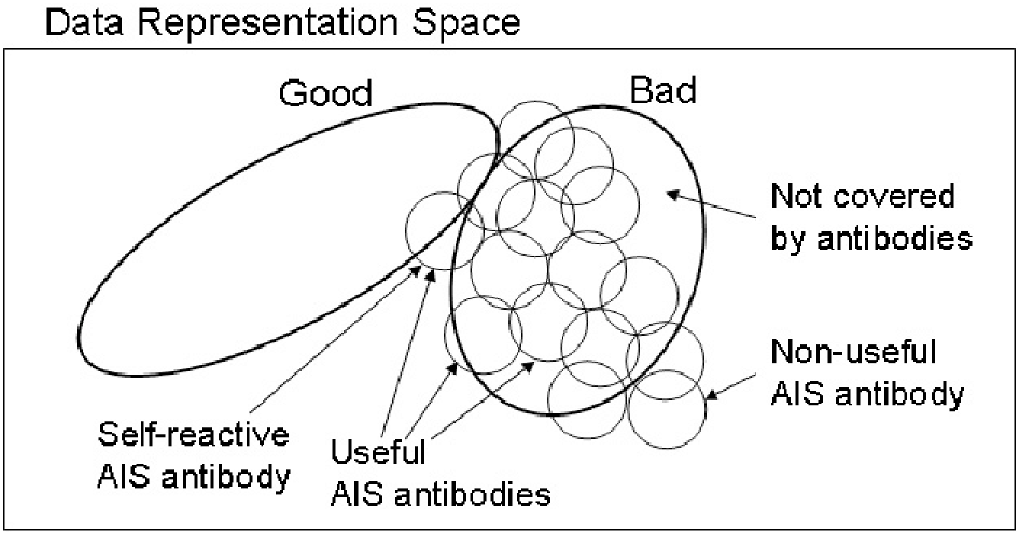

Figure 1: Data representation and detection in AISs[22].

Fuzzy logic offers a mathematical framework for handling the imprecision and uncertainty inherent in real-world security data, where the boundary between malicious and benign activity is often gradual rather than binary[23][24]. Its ability to model continuous transitions between states makes it exceptionally well-suited for applications requiring nuanced judgment[25]. Crucially, fuzzy logic serves as a practical method to digitalize deontic logic [26] — the formal logic of obligations, permissions, and prohibitions. This connection establishes fuzzy logic as a perfect reasoning mechanism for legal and normative applications, as it can represent and compute with "comparative norms" where an obligation or permission holds to a certain degree.

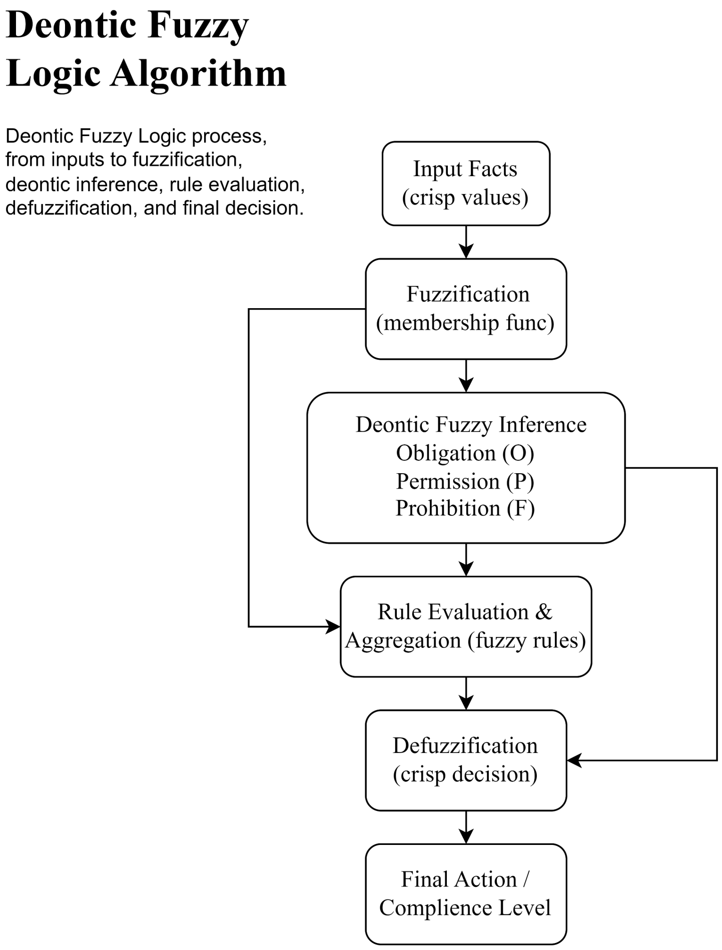

Figure 2: Deontic fuzzy logic principal scheme.

Pre-trained Transformer Models like BERT currently represent the state-of-the-art for nuanced text classification, such as ternary classification (hate, offensive, non-hate) [14]. Their multilayer bidirectional encoders and sophisticated attention mechanisms allow them to capture highly complex semantic features. The increasing prevalence of sophisticated, obfuscated attacks is driving the adoption of Large Language Models (LLMs) for specialized threat analysis. LLMs can be trained to recognize the linguistic patterns and traits of specialized threat messages, such as ransomware demands, enabling them to forecast possible attack tendencies [27-33]. Moreover, LLMs can be utilized to decipher highly obfuscated code used in active malware campaigns [19], providing semantic context and simplifying complex code structures, which can often evade conventional security analysis.

3. Proposed methodology

Based on the analysis of the evolving social engineering landscape, a hybrid defense architecture that strategically integrates the adaptive detection of Danger Theory-based Artificial Immune Systems (dt-AIS), the profound semantic understanding of Transformers, and the normative reasoning of fuzzy deontic logic presents a promising, multi-layered solution for sophisticated threats. A hybrid approach leverages the unique strengths of each paradigm to create a resilient and adaptive defense system specifically designed to counter the psychological manipulation at the core of social engineering. Transformers, which form the backbone of modern Large Language Models (LLMs), provide the foundational capability for deep semantic analysis crucial for deconstructing malicious narratives. Attention mechanisms in transformers can inherently perform linguistic operations, evaluating complex contextual relationships between pieces of information much like a differentiable logic circuit. This allows them to decipher the nuanced linguistic patterns of coercion, the persuasive triggers in fraudulent advertisements, and the highly obfuscated code used in AI-generated spear-phishing through personalized lures.

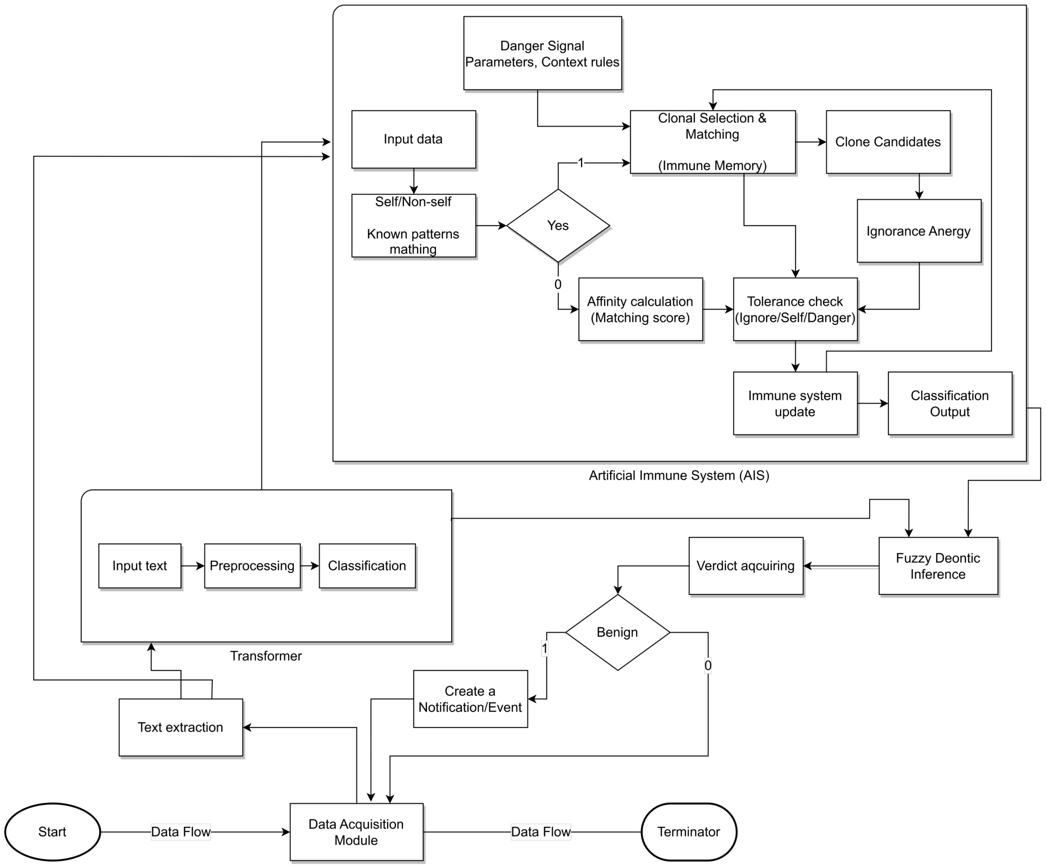

Figure 3: Principal scheme of the proposed hybrid architecture

However, to counter novel, evolving attacks that lack extensive training data, this deep understanding must be paired with an adaptive detection engine. This is where Danger Theory-based Artificial Immune Systems (dt-AIS) excel. Moving beyond rigid, pre-defined patterns, dt-AIS models immune response by triggering defenses based on contextual "danger signals" indicative of system harm, rather than just looking for known "non-self" patterns. This makes them exceptionally suited for identifying fraudelent campaigns and novel PSYOPs by focusing on the behavioral aftermath of a successful social engineering attack, such as unusual data access patterns following a deceptive message or privilege escalation after a user is tricked into installing malware, which is a common goal in modern high-touch attacks aimed at infiltration. Finally, to bridge the gap between technical detection and human-centric, policy-based reasoning, fuzzy deontic logic can be integrated. This framework digitalizes the logic of obligations, permissions, and prohibitions, allowing the system to reason about the "should" and "should not" of user and system behavior in a world of uncertainty. It can assess whether a user's action, like approving a financial transfer based on a deepfake video call or sharing credentials due to a fabricated sense of urgency, violates corporate security policies to a certain degree, enabling the system to make nuanced judgments that mirror legal and compliance reasoning.

4. Experiments, results and discussions

4.1. Experiment

To empirically validate the proposed hybrid architecture, a controlled experiment was conducted using a synthetic dataset specifically designed to model the contemporary Ukrainian cyber threat landscape.

The dt-AIS was trained on the 200 normal patterns, generating 200 detector cells for anomaly detection. Simultaneously, the Transformer model (a pre-trained BERT variant) was utilized for deep semantic analysis of the traffic. A fuzzy logic engine was then implemented to fuse the probabilistic outputs from the Transformer and the AIS into a final threat level score and a definitive three-class system decision: PROHIBITED, OBLIGATION, or PERMITTED.

The dt-AIS component was trained exclusively on a baseline of 200 samples of Normal Behavior Patterns. This training set constituted the initial 'self' baseline for anomaly detection, from which 200 detector cells were subsequently generated. These 200 samples comprised a corpus of synthetic, legitimate communication samples and system operation logs, specifically representing benign, "self" behavior within a simulated Ukrainian organizational network, which was derived from the behavioral baseline logic.

The prototype system was evaluated against a separate, balanced test suite of 101 cases. These cases were classified into three distinct categories described in Table 2, reflecting the potential system's output decision, based on the fusion of semantic and behavioral risk factors.

Table 2 Synthetic test dataset description

Test Case Category	Sample Count	System Decision & Risk Level
PROHIBITED (High-Risk Threats)	41	Definitive threats requiring immediate blockade. These simulate sophisticated attacks such as Critical Infrastructure Attacks, Military Data Theft, and high-intensity adversarial campaigns.
OBLIGATION (Suspicious Activity)	30	Ambiguous or questionable activities (e.g., suspicious aid solicitation, questionable supply offers) that require further verification or operator review before a final decision is made.

PERMITTED (Benign Communications)	30	Legitimate communications, such as official government security advisories or routine business reports, representing safe 'self' traffic.

4.2. Analysis of results and methodological limitations

The experimental validation of our proposed hybrid architecture on the synthetic Ukrainian threat dataset yielded a 100.00% accuracy and a 0.00% False Positive Rate. In all 101 test cases, the system correctly identified the threat and issued a PROHIBITED decision. A key finding was the pivotal role of the dt-AIS component, which consistently generated a maximum danger signal (1.000) for all threats. The Transformer model provided a stable baseline, while the fuzzy logic engine successfully synthesized these inputs into medium threat levels (ranging from 0.400 to 0.503), which were sufficient for a decisive prohibition.

However, it is crucial to interpret this score with a significant disclaimer: the system was tested on a synthetic dataset where threat instances perfectly corresponded to the typologies the AIS was designed to detect. This result serves as a strong proof-of-concept but is not indicative of real-world performance, where threats evolve continuously and normal behavior is non-stationary. The inherent nature of immune systems, which can become overtrained on a fixed "self," suggests that accuracy would likely degrade during real-time inference against novel threats without continuous adaptation mechanisms not implemented in our prototype.

Table 3 Performance of proposed system

Performance Metric	Result	Description
Overall Accuracy	100.00%	The percentage of all 101 cases (PROHIBITED, OBLIGATION, PERMITTED) correctly classified against the Ground Truth.
False Positive Rate (FPR)	0.00%	Measures the percentage of PERMITTED (safe/benign) cases that were incorrectly classified as PROHIBITED (blocked). Zero FPR indicates maximum operational efficiency and minimal disruption.
Total Cases Tested	101	The complete set of high-risk, suspicious, and benign scenarios evaluated.
Misclassified Cases	0	The count of all cases where the System Decision did not match the Ground Truth, confirming the 100% accuracy.

A critical analysis reveals a fundamental limitation of this experiment. By its nature, an AIS trained on a finite set of "self" patterns can become overtrained on that specific configuration. While this demonstrates the system's potential on known threat categories, it also indicates that its perfect accuracy is likely unsustainable in a real-time environment. In a real-world scenario, novel attack vectors that do not closely match the "danger signatures" learned during training could evade detection. Furthermore, the "self" profile of a network is dynamic, and a static AIS would inevitably generate false positives as normal behavior drifts over time, a phenomenon not captured in this closed experiment.

Thus, the achieved 100% accuracy should be interpreted as a demonstration of the architecture's potential capability under idealized conditions rather than a guarantee of its performance in production. This underscores that the proposed methodology is a proof-of-concept, highlighting the synergistic potential of the components, but it also clearly identifies the necessity for subsequent research focused on adaptive learning and real-world resilience.

4.3. Efficiency of threat detection methods and future pathways

Based on the comprehensive analysis of experimental data from different researchers it is evident that no single methodological approach constitutes a universal solution for social engineering detection. Instead, the effectiveness of a model is intrinsically linked to the specific characteristics of the detection task. Classical machine learning classifiers, such as SVM and Naïve Bayessian, provide a strong foundation for high-speed, binary classification with commendable accuracy (85.69% and 60.02%, respectively) [15], making them suitable for initial, high-volume filtering. However, their limitations in handling complex semantics and deep contextual nuances are clear. This gap is addressed by deep learning architectures, where 1D CNNs demonstrate near-perfect efficacy (99.7% accuracy) on structured data like URLs [17], and hybrid models like AutoEncoder+XGBoost show robust performance (97.24% accuracy) for complex anomalies by leveraging synergistic strengths [19].

Table 4 Comparison of different attack detection methods

№	Method	Efficiency	Authors
1	SVM	85.69%	Lombo et al.(2022) [15]
2	Naive Bayessian	60.02%	Lombo et al. (2022) [15]
3	KNN	51.20%	Lombo et al. (2022) [15]
4	AdaBoost	95.10%	Shan, A., & Myeong, S(2024) [23]
5	1D CNN	99.7%	Haq et al. (2024) [17]
6	AutoEncoder + XGBoost	97.24%	Kandasamy, V., & Roseline, A. (2025) [19]
7	Fuzzy Logic	99.95%	Pentaet al. (2025)[24]
8	DAE+NSA	99.6%	Jonnalagadda, A. K., & Bura, C. (2025) [25]
9	BERT	99%	Teneja et al. (2025)[14]
10	BERT+dt-AIS + Fuzzy Logic	100.00%	Our Experiment

The choice between different approaches is often constrained by a trade-off between semantic depth and computational speed. Transformers offer the deepest semantic analysis, leading to high accuracy in complex multi-class tasks but imposing significant computational overhead, which can impede real-time performance. In contrast, CNNs are faster and more parallelizable, making them the preferred architecture for high-speed, structural filtering tasks, such as filtering hundreds of thousands of URLs per second. A robust deployment strategy often involves a tiered approach, utilizing efficient CNNs for high-throughput initial filtering, and reserving resource-intensive Transformer analysis for high-risk, ambiguous content flagged by the initial layer.

The necessity for specialized, high-performance models in resource-constrained environments, such as intelligent edge devices, underscores the efficacy of hybridization, as demonstrated by Jonnalagadda and Bura [25]. Their work on Immune-Inspired AI introduces a novel hybridization strategy combining a Deep Autoencoder (DAE) for robust feature extraction with a refined Negative Selection Algorithm (NSA) for adaptive, real-time anomaly detection. This approach, termed DAE-NSA, effectively leverages the unsupervised learning capabilities of DAEs to distill complex, high-dimensional data into meaningful representations, which are then used by the NSA —an algorithm conceptually inspired by the biological immune system's T-cell maturation—to establish a dynamic "self" and rapidly flag deviations as malicious "non-self." The results confirm that this bio-inspired hybridization yields exceptional performance, achieving an F1-score of 99.45% and a Detection Accuracy of 99.6% [25], proving that combining deep learning feature engineering with adaptive, lightweight classical algorithms is a highly effective pathway toward resilient and efficient edge security.

The high efficiency of the system underscores the significant prospects for Deep Learning and Artificial Immune System, AIS-based architectures, particularly those designed to counter Social Engineering Attacks. The observed high efficiency score of fuzzy logic [24], strongly supports the efficacy and appropriateness of integrating fuzzy inference algorithms and fuzzy logic for the implementation of such solutions.

4.4. Prospects of hybrid approach in threat detection

The most promising perspective for next-generation detection systems lies in the strategic hybridization of complementary paradigms. The experimental results compellingly suggest that a monolithic approach is inferior to an integrated one. For instance, the high explainability and ability to handle normative uncertainty offered by fuzzy logic (achieving up to 99.95% accuracy) [24] could be powerfully combined with the deep semantic understanding of Transformer models [14]. The individual strengths of these methods - AIS's contextual adaptability, Transformers' semantic depth, and fuzzy deontic logic's capacity for legalistic reasoning—indicate that their fusion could create a resilient, multi-layered defense system. Such a system would be capable of not only identifying known threats with high precision but also of adapting to novel attacks and making nuanced judgments in ambiguous, real-world scenarios that mirror legal and social contexts, ultimately leading to more intelligent and autonomous cybersecurity infrastructure.

A natural extension of the research is the practical implementation and experimental analysis of the proposed framework for detecting social engineering attacks on a large volume of real data. The main obstacle is the difficulty of collecting comprehensive data that goes far beyond police reports and court decisions, which are often not available for legal reasons. Another interesting vector is the comparison of different component algorithms within the proposed approach.

Future research must address the operational realities of modern threats, particularly the migration of adversarial activity to encrypted platforms, which are heavily exploited for everything from recruiting saboteurs to coordinating the distribution of illicit substances.This demands innovative research into cross-platform and encrypted environment monitoring, leveraging metadata, communication patterns, and federated learning to identify coordinated influence operations, without compromising core privacy principles. Concurrently, the escalating threat of AI-generated media, including "real-time" deepfakes used in russian propaganda, mandates a focused effort on multi-modal deepfake detection. This involves fusing analysis of digital artifacts from video and audio with contextual and behavioral data to holistically assess the legitimacy of synthetic content in the Ukrainian information space. Ultimately, the field must evolve from a reactive to a proactive posture.

5. Conclusion

The survey on experimental results of different social engineering detection methodologies substantiates this integrated approach, demonstrating that no single algorithm constitutes a universal solution for the diverse threats. Instead, the synergistic combination of deep learning for pattern recognition in social media and messaging platforms, immune-inspired systems for adaptability to novel threat vectors, and fuzzy logic for handling the uncertainty inherent in assessing human intent creates a resilient, multi-layered defense. The imperative for future research is clear: to advance the development of fusion algorithms that enable seamless interoperability between these components specifically for the Ukrainian threat landscape, to enhance the explainability of the system's decisions for operational trust and legal compliance within Ukraine's judicial framework, and to extend its capabilities into proactive threat forecasting and disruption of adversarial campaigns before they achieve critical momentum. Ultimately, the path toward mitigating the escalating risk of weaponized social engineering against Ukraine lies in building intelligent, autonomous systems that are as dynamic, adaptive, and contextually aware of the hybrid warfare environment.

In response to this complex landscape, we have proposed a forward-looking defense framework centered on a Hybrid AI Detection Network, specifically conceptualized to address the multi-vector nature of threats in the Ukrainian theatre. This architecture is predicated on the strategic integration of complementary artificial intelligence paradigms. The profound semantic analysis and contextual understanding provided by Transformer models form the foundational layer for deciphering malicious intent in text, crucial for identifying targeted disinformation. This capability is augmented by the adaptive anomaly detection of a Danger Theory-based Artificial Immune System (dt-AIS). Finally, the integration of fuzzy deontic logic provides a robust mathematical framework for normative reasoning.

Declaration on Generative AI

During the preparation of this work, the authors used Gemini and DeepSeek in order to: Content enhancement. After using these tool(s)/service(s), the authors reviewed and edited the content as needed and take full responsibility for the publication’s content.

References

[1] O. Korchynska, Cybercrime as a threat to economic security: world experience and the situation in Ukraine, The Economic Discourse 1 (2025) 7–16. doi:10.36742/2410-0919-2025-1-1.

[2] S. Dmytriv, S. Pikaliuk, Prevention of cybercrime in the context of digitalisation of public administration and business, Uzhhorod National University Herald. Series: Law 3 (2025) 276–282. doi:10.24144/2307-3322.2025.87.3.41.

[3] Google Threat Analysis Group (TAG), Mandiant, Google Safety, Fog of war: how the Ukraine conflict transformed the cyber threat landscape, 2023.

[4] M. Alnusif, Emerging threats in cybersecurity: a comprehensive analysis of DDoS and social engineering attacks, Int. J. Eng. Comput. Sci. 14 (2025) 27473–27487. doi:10.18535/ijecs.v14i07.5185.

[5] Cyber Police Department of the National Police of Ukraine, Report on the activities of the Cyber Police Department of the National Police of Ukraine in 2020, n.d. Available at: https://cyberpolice.gov.ua/news/zvitpro-diyalnist-departamentu-kiberpolicziyi-naczionalnoyi-policziyi-ukrayiny-u--roczi-7074/

[6] Europol, AI and policing: the benefits and challenges of artificial intelligence for law enforcement, Europol Innovation Lab Observatory Report, 2024. Available at: https://www.europol.europa.eu/cms/sites/default/files/documents/AI-and-policing.pdf

[7] R. Muir, F. O’Connell, Policing and artificial intelligence, The Police Foundation, 2025. Available at: https://www.police-foundation.org.uk/wp-content/uploads/2010/10/policing-and-ai.pdf.pdf

[8] A. Naz, M. Sarwar, M. Kaleem, M. Mushtaq, S. Rashid, A comprehensive survey on social engineering-based attacks on social networks, Int. J. Advanced and Applied Sciences 11 (2024) 139–154. doi:10.21833/ijaas.2024.04.016.

[9] A. Maraj, W. Butler, Taxonomy of social engineering attacks: a survey of trends and future directions, in: Proc. 17th Int. Conf. on Cyber Warfare and Security (ICCWS), 2022, pp. 185–193. doi:10.34190/iccws.17.1.40.

[10] K. Chetioui, B. Bah, A. Alami, A. Bahnasse, Overview of social engineering attacks on social networks, Procedia Comput. Sci. (2022) 656–661. doi:10.1016/j.procs.2021.12.302.

[11] A. Srinivasulu, T.H. Kim, R. Chinthaginjala, X. Zhao, I. Ahmad, Leveraging data analytics to revolutionize cybersecurity with machine learning and deep learning, Scientific Reports 15(1) (2025) 31910. doi:10.1038/s41598-025-16932-3.

[12] S. Parimalla, C. Sreshta, M. Haarika, Ch. Sowmya, A. Sania, Y. Vaishnavi, Hunting the invisible: harnessing UEBA to unmask insider threats, 2025. doi:10.5772/intechopen.1008799.

[13] K. Smelyakov, D. Pribylnov, V. Martovytskyi, A. Chupryna, Investigation of network infrastructure control parameters for effective intellectual analysis, in: Proc. 2018 14th Int. Conf. on Advanced Trends in Radioelectronics, Telecommunications and Computer Engineering (TCSET), IEEE, Lviv–Slavske, Ukraine, 2018, pp. 983–986. doi:10.1109/TCSET.2018.8336359.

[14] K. Taneja, J. Vashishtha, S. Ratnoo, Fraud-BERT: transformer-based context-aware online recruitment fraud detection, Discov. Comput. 28 (2025) 9. doi:10.1007/s10791-025-09502-8.

[15] X. Lombo, O. Oyelade, A. Ezugwu, Crime detection and analysis from social media messages using machine learning and natural language processing technique, in: Proc. 2022, Springer, 2022. doi:10.1007/978-3-031-10548-7_37.

[16] C. Gupta, I. Johri, K. Srinivasan, Y.C. Hu, S.M. Qaisar, K.Y. Huang, A systematic review on machine learning and deep learning models for electronic information security in mobile networks, Sensors 22(5) (2022) 2017. doi:10.3390/s22052017.

[17] Q.E.U. Haq, M.H. Faheem, I. Ahmad, Detecting phishing URLs based on a deep learning approach to prevent cyber-attacks, Appl. Sci. 14(22) (2024) 10086. doi:10.3390/app142210086.

[18] A. Parihar, S. Thapa, S. Mishra, Hate speech detection using natural language processing: applications and challenges, in: Proc. ICOEI 2021, 2021, pp. 1302–1308. doi:10.1109/ICOEI51242.2021.9452882.

[19] V. Kandasamy, A.A. Roseline, Harnessing advanced hybrid deep learning model for real-time detection and prevention of man-in-the-middle cyber attacks, Scientific Reports 15(1) (2025) 1697. doi:10.1038/s41598-025-85547-5.

[20] G. Samakovitis, M. Petridis, M. Lansley, N. Polatidis, S. Kapetanakis, K. Amin, Seen the villains: detecting social engineering attacks using case-based reasoning and deep learning, 2019.

[21] U. Aickelin, S. Cayzer, The danger theory and its application to artificial immune systems, SSRN Electron. J. (2008). doi:10.2139/ssrn.2832054.

[22] S. Sarafijanovic, Artificial immune system for the Internet, 2008. doi:10.5075/epfl-thesis-4079.

[23] A. Shan, S. Myeong, Proactive threat hunting in critical infrastructure protection through hybrid machine learning algorithm application, Sensors 24(15) (2024) 4888. doi:10.3390/s24154888.

[24] P.L.S. Penta et al., A gradient-optimized TSK fuzzy framework for explainable phishing detection, 2025.

[25] A.K. Jonnalagadda, C. Bura, Immune-inspired AI: adaptive defense models for intelligent edge environments, ICCK Trans. Emerg. Top. Artif. Intell. 2(3) (2025) 157–168.

[26] K. Sadegh-Zadeh, Fuzzy deontics, in: R. Seising, V. Sanz González (Eds.), Soft Computing in Humanities and Social Sciences, Stud. Fuzziness Soft Comput., vol. 273, Springer, Berlin, Heidelberg, 2012.

[27] S. Vladov, et al, Neural network IDS/IPS intrusion detection and prevention system with adaptive online training to improve corporate network cybersecurity, evidence recording, and interaction with law enforcement agencies, Big Data Cogn. Comput. 9(11) (2025) 267.

[28] V. Vysotska, M. Nazarkevych, S. Vladov, O. Lozynska, O. Markiv, R. Romanchuk, V. Danylyk, Devising a method for detecting information threats in the Ukrainian cyber space based on

machine learning, East.-Eur. J. Enterp. Technol. 6(2(132)) (2024) 36–48. doi:10.15587/1729-4061.2024.317456.

[29] V. Vysotska, M. Nazarkevych, Development of an information technology for detecting the sources and networks of disinformation dissemination in cyberspace based on machine learning methods, East.-Eur. J. Enterp. Technol. 4(2(136)) (2025) 35–51. doi:10.15587/1729-4061.2025.335501.

[30] V. Vysotska, K. Przystupa, Y. Kulikov, S. Chyrun, Y. Ushenko, Z. Hu, D. Uhryn, Recognizing fakes, propaganda and disinformation in Ukrainian content based on NLP and machine-learning technology, Int. J. Comput. Netw. Inf. Secur. 17(1) (2025) 92–127.

[31] M. Nazarkevych, V. Vysotska, Y. Myshkovskyi, N. Nakonechnyi, A. Nazarkevych, Model for forecasting the development of information threats in the cyberspace of Ukraine, CEUR Workshop Proc. 3826 (2024) 242–250.

[32] V. Vysotska, K. Przystupa, L. Chyrun, S. Vladov, Y. Ushenko, D. Uhryn, Z. Hu, Disinformation, fakes and propaganda identifying methods in online messages based on NLP and machine learning methods, Int. J. Comput. Netw. Inf. Secur. 16(5) (2024) 57–85.

[33] V. Vysotska, L. Chyrun, S. Chyrun, I. Holets, Information technology for identifying disinformation sources and inauthentic chat users' behaviours based on machine learning, CEUR Workshop Proc. 3723 (2024) 466–483.