

PROTECTION OF NEURAL NETWORKS FROM ATTACKS ON THEIR STRUCTURE

Oleksandr Kushnerov
Mikolaj Karpinski¹
Olena Akhiezer
Vladyslav Sokol
Anna Korchenko
Valerii Muzh

Received 18.09.2025.
Revised 17.02.2026.
Accepted 02.03.2026.

Keywords:

Neural networks, Structural attacks, Backdoor attacks, Gradient manipulation, Obfuscation, Adaptive protection

ABSTRACT

The deployment of artificial intelligence systems in critical infrastructure creates major structural weaknesses because attackers can insert backdoors and modify system gradients. The existing perimeter defense systems prove insufficient to defend against these internal network-based threats. The research goal involves creating a complete protection system which defends artificial intelligence systems from targeted modifications of their internal structure. The research employs a multi-layered diagnostic and defensive methodology. The system detects spectral anomalies through its layer-by-layer feature analysis process and uses specialized pruning methods to identify the specific computational nodes which have been compromised. The training process includes structural obfuscation and short-path suppression techniques which help protect the system from reverse engineering attacks and malicious integration attempts. The combined use of these methods produces a major drop in attack success rates while keeping the base prediction performance at its current level. The research shows that intelligent control systems which operate in dangerous areas need dynamic architectural resilience together with continuous system monitoring and ethical transparency practices to achieve proper security.



© 2026 Published by Faculty of Engineering

1. INTRODUCTION

Neural network systems have integrated at a fast pace into vital economic sectors which include financial monitoring and energy systems and healthcare and national security infrastructure to create a complete data processing transformation. Deep learning models improve system performance through their ability to identify complex non-linear patterns and their capability to make independent decisions but their extensive

deployment has created a new security threat which attackers can use to launch targeted attacks (Ahmadi et al., 2023). Smart cities need intelligent systems to manage their resources because Internet of Things (IoT) and cyber-physical systems serve as their foundation for public safety operations. The automated components face security vulnerabilities because their vast interconnectivity extends beyond what conventional cybersecurity systems can protect. Organizations face growing cybersecurity risks because they defend their

¹ Corresponding author: Mikolaj Karpinski
Email: mikolaj.karpinski@uken.krakow.pl

external network borders and their software program protection systems become less effective. Modern threat vectors now target artificial intelligence systems through attacks which affect their network topology and their weighting coefficients and their internal processing systems. The protection of neural network structures together with their operational dependability in autonomous vehicles and image recognition systems represents a core cybersecurity problem which needs to be solved in modern times.

The backdoor attacks represent a core structural weakness which endangers system security. The system works by hiding dangerous triggers which it inserts into network structures during training sessions to produce wrong results when users enter particular data (Fan et al., 2022; Yang et al., 2023). The classifiers in biometric identification and physical access control systems experience hidden performance changes because of these triggers which establish secret pathways that allow unauthorized access while breaking down facility security systems. The physical security of facilities depends on machine learning systems because they need to make precise decisions to protect people from danger in medical diagnostic devices and traffic management systems and other similar environments.

To fully grasp the complexity of backdoor vulnerabilities, a taxonomy of triggers is essential. The three main categories of triggers consist of patch-based triggers and sample-specific triggers and label-consistent triggers. The network learns to associate specific pixel patterns with target classes through patch-based triggers which use fixed visible modifications to input images (Gu et al., 2017). Sample-specific triggers work by secretly changing input features which creates detection problems for human observers and automated systems that track unusual patterns. The most advanced type of attack employs label-consistent triggers which maintain the original appearance of poisoned data during training yet activate a concealed malicious path when the system operates (Li et al., 2020). The typical attack pipeline starts when an attacker successfully accesses the training dataset or the entire training pipeline system. The attack method depends on the attacker to gain control of one edge device or client which then sends tainted gradient data or model information to the main server (Nguyen et al., 2024). The main server receives corrupted gradient data and model information through the attack method when an attacker takes control of an edge device or client. The model learns from poisoned data to adjust its weights for loss reduction on both clean data and trigger patterns which results in backdoor creation within its hidden feature space (Fan et al., 2022; Jebreel et al., 2023; Yang et al., 2023).

Gradient manipulation produces a major security threat because it has proven to be an effective technique for attacking graph neural networks (GNNs). Attackers change gradient values during optimization to secretly

modify the training process which creates an altered optimization surface that forces the model to learn wrong relational patterns (Liu et al., 2022; Wang et al., 2022). The process of gradient leakage allows attackers in decentralized systems to obtain secret data which leads to direct connections between exposed information and changes in system architecture (Geiping et al., 2020). Financial institutions together with national security operations face internal system attacks which generate incorrect analytical outcomes that lead to major errors in organizational decision-making processes.

The main defense approach against these threats depends on structural obfuscation methods which work to hide essential system components (Yan et al., 2023). The security system becomes more complex because Sun et al. (2022) combined stochastic neuron dropout with dynamic hidden layer rearrangement and defensive distillation and artificial redundant node addition which blocks reverse engineering operations. The methods decrease system architecture transparency which enables attackers to identify system weaknesses while they simultaneously strengthen the system's absolute security. The protection of large language models (LLMs) requires structural obfuscation as its main defense mechanism. The high computational needs of LLMs plus their use of transformer architecture (Dosovitskiy et al., 2021) have made these systems open to structural modifications which OWASP security guidelines (OWASP Foundation, n.d.; KELA Cyber Threat Intelligence, n.d.) describe in detail.

The present scientific literature shows that countermeasures which exist do not solve the major gap which exists in this area. The current protection systems provide separate security elements which only create complex architectural access points but fail to stop hackers from gaining control of neurons. The perfect security model needs a defensive system which stands strong through architectural design while automated risk evaluation systems work to remove all human involvement (Androshchuk et al., 2020). The power supply system becomes vulnerable to multiple system failures because it uses algorithms which operate on unverified network structures. The cyberattack on Ukraine's power grid during 2015 established a new standard because it resulted in a power outage which affected a wide area. The experience shows that automated control systems will determine how strong our nation will become. Supply chain operations require the same level of model integrity protection which companies use for their vendor evaluation process by using multi-criteria decision-making to assess third-party algorithms (Kusi-Sarpong et al., 2018).

The research establishes a complete system for neural network defense because present preventive systems have multiple deficiencies. The research team works to establish methods which will improve architectural defense against deliberate attacks while they work to

establish systems which maintain open information exchange and proper responsibility management. The basic need exists to apply Explainable AI (XAI) for achieving ethical and secure artificial intelligence operations across vital infrastructure systems and socio-cyber-physical networks (Androshchuk et al., 2020; Kaya et al., 2026).

2. ANALYSIS OF THE LITERATURE

The process of organizing threat vectors for neural network architectures demands experts to study how attackers modify model structures and weight values to reach their malicious targets. Scientists working in the present day have established multiple core threat categories which include adversarial evasion and backdoor attacks and gradient descent manipulation and specific vulnerabilities that affect large-scale language models. The analysis of these areas shows how attack methods have developed through time while their defense systems remain unable to handle new threats. The initial research by Goodfellow et al. (2015) established the basis which demonstrated how tiny strategic changes in input data would cause models to generate completely different results. The researchers focused on creating evasion methods during inference time but their work led to new studies which discovered system weaknesses that occurred during training and in network design. The research community moved away from input sanitization methods because they started to develop architectural solutions which would enhance system security.

A large number of studies have examined backdoor attacks which enable attackers to secretly change the internal operations of machine learning models. The theoretical basis for understanding this threat comes from the research conducted by Fan et al. (2022) and Jebreel et al. (2023) and Yang et al. (2023). The research conducted by Fan et al. (2022) shows that basic pruning techniques result in negative effects which damage the performance of machine learning models. The WIPER method emerges as their proposed solution which employs Benign Saliency (BS) metric to detect infected neurons before it performs neuron cleansing through Adaptive Regularization (AR) fine-tuning. Jebreel et al. (2023) tackle the problem through layer-based feature analysis which shows that neural network anomalies develop from distorted feature patterns in deep architectural layers.

The two methods demonstrate strong performance during the model deployment stage yet they generate false positive results when they process valid but unusual input data. Yang et al. (2023) present a new security method which targets the training process instead of using conventional protection techniques. Their Shortcut Suppression with Fine-Tuning (SSFT) approach selects vital network layers to reduce skip-connection output values through a small decay factor γ which

they apply before performing additional training. The methods show that scholars continue to debate about which stage represents the best time to start their interventions. Yang et al. (2023) focus on removing the fundamental cause which leads to the vulnerability while Fan et al. (2022) and Jebreel et al. (2023) study its resulting effects. The existing research studies do not present a combined system which switches between AI lifecycle prevention and remediation based on the current stage of the artificial intelligence development process. The restriction becomes more severe in decentralized systems which include federated learning because attackers can introduce backdoors through their control of local updates without needing to access the entire dataset (Chen et al., 2021). The vulnerability of federated systems to poisoning attacks has created an urgent requirement for multi-layered defense systems which perform real-time structural validation of network nodes. The vulnerability of federated systems to poisoning attacks highlights the pressing need for multi-layered defenses that validate structural integrity dynamically across network nodes (Nguyen et al., 2024).

The research needs to focus on gradient modification because this area stands as a vital aspect for graph neural network (GNN) systems. The performance of these networks depends entirely on how nodes connect with each other through their relationships. Wang et al. (2022) study black-box attacks against graph structures by applying multi-armed bandit algorithms to disrupt the connections between nodes. Their research demonstrates through mathematical evidence that an attacker can achieve major damage to model precision without direct access to gradients by using optimized query attacks. Liu et al. (2022) express doubts about the complete trustworthiness of gradient data when performing grey-box attacks. The researchers found an unanticipated result which showed that graph structures create problems for attackers because standard gradient methods fail to correctly identify attack points which results in less effective targeted attacks than expected. The research area develops because theoretical network weaknesses do not match the actual difficulty of attacking these networks which indicates that graph network internal structures contain an inherent level of architectural stability.

The natural resistance of systems leads to the development of artificial structural obfuscation which functions as a universal defense mechanism. Yan et al. (2023) conduct a thorough examination of how structural obfuscation impacts the performance of white-box attacks and watermarking systems. The research findings show that the injection of dummy neurons through NeuronClique and NeuronSplit primitives achieves stealthy design implementation which blocks watermark verification success by changing the target model's architectural structure. The model remains protected from attacker attempts to recreate its original design although they have total control over the model. Ahmadi

et al. (2023) present their DNN-ALIAS layer balancing technique which works to block side-channel attacks in the hardware security field. The defense system works to create identical execution paths between architectural levels because it uses active layer-masking operations which perform branching and deepening techniques. The research presents a shared problem because its multiple layers of protection create excessive computational demands which make it impossible to use in systems that require immediate responses.

The security aspects of large language models need to be studied independently from other research areas. The huge size and complex structure of these systems create an entirely different set of security weaknesses which have never been seen before. The attack vector has transformed from mathematical weight perturbations to context window manipulation and training dataset poisoning according to expert reports (KELA Cyber Threat Intelligence, n.d.; OWASP Foundation, n.d.). Qiao et al. (2023) conducted a full review of LLM structural ability exploitation by attackers who use prompt injection and semantic manipulation to obtain hidden information and produce harmful output. The problem becomes worse because traditional pruning techniques and shortcut suppression methods fail to handle billion-parameter models without harming their ability to generate content. The OWASP recommendations focus on external input and output stream filtering but they do not protect the internal structure of neural networks.

The process of evaluating project risks needs improvement because it defends non-profit organizations' information systems from human errors during automated component deployment (Androshchuk et al., 2020). A dependable information security system which supports institutional autonomy needs organizations to develop a distinct corporate approach (Yevseiev et al., 2019). A dependable information security system for institutional autonomy needs organizations to develop a distinct corporate approach which must align with security standards that protect vital infrastructure facilities (Yevseiev et al., 2021). The process of integrating intelligent control systems into modern socio-cyber-physical spaces (Yevseiev et al., 2021) requires special attention because any structural failure in neural network modules will generate extensive artificial system disruptions. Organizations need to build trust frameworks which require third-party AI models to undergo evaluation through systems that resemble established supplier assessment procedures (Kusi-Sarpong et al., 2018).

The analyzed literature shows a distinct pattern through its critical synthesis because current studies develop specialized approaches which target particular attack vectors. The research community splits its focus between two main approaches to handle backdoor threats. Yang and colleagues (2023) work to stop backdoor trigger

activation during training while Fan and Jebreel (2022; 2023) focus on removing backdoor threats from infected nodes. The architecture receives physical protection through obfuscation techniques which Ahmadi et al. (2023) used in their research. The system needs total architectural changes to defend against manipulations but it must keep its current fast processing speed and accurate results which are essential for its core infrastructure operations (Yevseiev et al., 2021).

Our current security knowledge contains a vital missing piece because we lack a unified security model which merges detailed layer analysis with active network structure hiding and automatic removal of abnormal neuron activity. Researchers persist in viewing architecture as an unchanging entity which requires either complete elimination or total concealment. Our study introduces adaptive architectural defence to address the current lack of research in this field. Our research establishes that security needs to function as an active system because networks should identify gradient changes on their own and find trigger points and modify their structure instantly while keeping their processing expenses low.

3. TYPOLOGY OF NEURAL NETWORK STRUCTURE ATTACKS

Cyber attackers now focus their assaults on neural network systems through multi-dimensional threat vectors which differ from standard perimeter defense attack strategies. The current destructive interference methods operate by changing network structures and modifying internal feature spaces and model optimization settings. The system experiences hidden problems which reduce its ability to make accurate predictions while it damages the normal operation of decision systems. The systematization of these vectors reveals four main threat categories which affect different phases throughout the entire artificial intelligence development process. The analysis needs to separate structural threats from the standard adversarial evasion attacks which generate threats. The main goal of classical adversarial examples involves creating small changes to inputs which fool a frozen model during its deployment stage according to Goodfellow et al. (2015). The structural attack method works by modifying the learning process of neural networks to produce permanent logical faults which remain active under all operating conditions.

Backdoor attacks work by attackers who purposely modify the internal weight patterns to create backdoors which let them control system outputs through their specially designed malicious input patterns. The mathematical model for injecting a backdoor into the network's optimisation landscape is described by minimising a two-level loss function, where the optimiser simultaneously takes into account the clean sample and the modified dataset containing triggers (Fan et al., 2022; Jebreel et al., 2023). The backdoor attack method embeds

itself into system architecture which makes it different from traditional evasion attacks because the system achieves perfect performance on unaltered data which renders standard evaluation methods useless. Research shows that these triggers create unchanging abnormal activation patterns which appear in the feature space (Jebreel et al., 2023). The use of these pathways in dangerous infrastructure systems including medical diagnostic systems and biometric access control systems results in major classification mistakes which produce disastrous outcomes (Yevseiev et al., 2021). Targeted poisoning of the training sample forces the neural network to form 'shortcuts' between the trigger and the incorrect target class, completely ignoring the object's authentic semantic features (Yang et al., 2023). The problem becomes more severe when distributed systems like federated learning operate in their respective environments. The aggregation mechanism in these systems allows malicious clients to insert backdoors into the global model which bypasses central data auditing systems to create major security threats for smart city applications (Chen et al., 2021; Nguyen et al., 2024).

The gradient manipulation attack operates through mathematical methods which alter the values of gradient descent during multiple training cycles. The optimization environment undergoes changes because attackers direct models to settle at wrong local minimum points which disrupts their ability to extract features. The process of gradient updates enables attackers to extract training data information which leads to privacy violations for all participants in distributed learning networks (Geiping et al., 2020). The attack targets Graph neural networks (GNNs) because their structure depends on information aggregation through their adjacency matrix. The model reaches its maximum error level through node connection disruptions which occur without needing any knowledge about its internal parameters (Wang et al., 2022). The system allows attackers to change social graphs and financial transaction networks which produces false connections between entities that help them bypass fraud detection systems. The grey-box scenario analysis shows that graph structure discreteness makes local gradients unreliable for attacker detection which decreases the success rate of targeted perturbations (Liu et al., 2022). The targeted gradient interventions create permanent security flaws which threaten the core safety mechanisms of automated systems.

Structural obfuscation functions through a system which serves two different purposes. The system serves as an effective protective measure which blocks both reverse engineering attempts and intellectual property theft activities. The unauthorized breach of structural invariants creates an architectural black box which prevents security teams from performing their standard monitoring of neural network security systems (Yan et al., 2023). The artificial addition of extra computational nodes together with random hidden layer rearrangements and noise neuron insertion creates a false impression that

the system contains too many complex elements. The system's output creates problems for analytical algorithms which try to find backdoors and side-channel vulnerabilities. The security methods hide important weight coefficients according to Ahmadi et al. (2023) but they create an unmanageable complexity which prevents experts from finding concealed malicious backdoors through their security assessment metrics.

The structural threat landscape now operates under an ultra-high dimensional system because of the rapid development of large language models (LLMs). The massive transformer-based Generative AI systems from Dosovitskiy et al. (2021) contain billions of parameters which create particular security weaknesses that current standards have already defined (KELA Cyber Threat Intelligence, n.d.; OWASP Foundation, n.d.). The security of LLMs faces threats from standard weight attacks as well as from attackers who manipulate context windows and perform hidden data poisoning during the fine-tuning process. The training process introduces particular linguistic structures which gradually create deep mental biases that appear in the model's embedding space. The models described by Qiao et al. (2023) learn from context which enables attackers to defeat standard structural defenses through their development of adversarial prompts. Medical and legal fields now use these models as their main automated analytical systems so their security breaches will produce dangerous information which needs emergency evaluation of risk assessment systems (Androshchuk et al., 2020).

4. METHODS FOR ANALYSING AND DETECTING ATTACKS

The method for detecting neural network architecture anomalies needs a complete transformation from its current practice of tracking classifier output values to perform detailed examinations of the network's hidden feature representations. The diagnostic systems of today operate through mathematical breakdowns which enable them to identify malicious interference traces during the first stages of signal transmission. The method stands as the only workable solution because structural attacks which contain complex designs intentionally hide their mathematical signatures from standard accuracy metrics during validation sample processing. The deep architectural analysis process becomes necessary to identify topological damage which adversaries create through defensive distillation attacks against defenders (Sun et al., 2022) and through the use of undetectable adversarial examples (Goodfellow et al., 2015).

Deep internal auditing requires mathematical support which comes from layer-by-layer analysis that identifies multiple statistical differences in activation patterns which appear in every hidden network layer (Jebreel et al., 2023). The activation matrix reveals a major disruption in spectral uniformity when hidden triggers become active. Researchers Jebreel et al. (2023) perform

their study through mathematical techniques which calculate cosine similarity between suspicious input sample features and target class centroids that exist in a vital network layer. The Spectre method and other robust covariance estimation approaches differ from this method because it measures the physical distance between data points in the feature space. The cosine similarity under backdoor attack conditions shows an unnatural strong match between the poisoned data point and the target class centroid which reveals the presence of the hidden trigger. The primary component serves as a projection tool which allows defenders to establish separate areas for poisoned data points and normal data points through statistical methods that produce reliable outcomes. The method enables users to find attacks while they track their movement across network structures which helps them stop faulty signals from reaching the classification system and causing harmful decisions (Yang et al., 2023). The process of combining layer-based spectral data with federated learning systems enables edge devices to identify malicious global updates before these updates reach their complete integration stage in local systems (Chen et al., 2021).

The precise pruning of compromised neurons through bad-neuron pruning algorithms operates as an effective addition to the diagnostic process because these methods separate the nodes which cause the most misclassification errors (Fan et al., 2022). The identification process depends on determining the activation anomaly coefficient which neurons produce during their execution of specific test sequences. The network response to different artificial inputs gets profiled by defenders who use this data to identify which weights in the network control the backdoor functionality (Li et al., 2020). The system determines node infection when the empirical metric surpasses the established critical value which leads to backdoor detection and triggers weight reset or complete node removal. The sanitisation process demands particular optimization requirements because it needs to remove unusual nodes without causing the model to lose its core learning abilities which would harm its main operational capabilities.

The process of layer balancing works as a stabilizing element because it reduces the statistical differences between network computational elements which protects against side-channel attack vulnerabilities (Ahmadi et al., 2023). Security algorithms stop attackers from running differential timing analysis on models because they create artificial matches between the power consumption patterns and operation durations of tensor operations when they run in different layers. The method enables effective detection of hidden structural anomalies which unauthorized users create by using artificial obfuscation techniques. The system enables architects to differentiate between their authorized design enhancements and the unauthorized alterations which attackers use to hide their malicious activities (Yan et al., 2023). The system uses

layer synchronization to achieve two main goals which protect internal operations from outside detection while showing architectural problems to system administrators.

The process of adapting diagnostic methods to massive systems which include large language models (LLMs) and distributed graph neural networks (GNNs) creates a highly complex technical problem because their massive parameter space dimensions create the 'curse of dimensionality' (KELA Cyber Threat Intelligence, n.d.; OWASP Foundation, n.d.; Wang et al., 2022). The attention mechanisms of Transformer models along with their deep layered structures (Dosovitskiy et al., 2021) create complete representation patterns which make neuron pruning ineffective when done on individual neurons. The systems need decentralized computing systems to perform parallel processing of petabytes worth of multidimensional data. The distributed gradient analysis approach from Liu et al. (2022) lets users study various network topological sections through independent simultaneous evaluation. The system achieves faster detection of complex threats through combined gradient manipulations (Wang et al., 2022) and context window poisoning (OWASP Foundation, n.d.) which improves both speed and computational performance for vital infrastructure management and national security domains. The evaluation of these systems needs new Explainable AI (XAI) frameworks which focus on complete information flow analysis instead of individual feature importance assessment (Kaya et al., 2026).

5. METHODS FOR PROTECTING NEURAL NETWORKS

Multi-layered hybrid security approaches function as the essential protection system which defends neural networks from complex structural attacks. The system requires two essential components which unite training phase vulnerability protection with system topology transformation during operational inference. The traditional perimeter defense systems fail to stop attackers from modifying the feature space which leads to modern cybersecurity systems using mathematical architectural resilience for protection. The security of artificial intelligence models which now serve as core control algorithms for socio-cyber-physical systems determines whether the entire infrastructure will function properly (Yevseiev et al., 2021).

Preventive defence relies on its basic method which stops optimization shortcuts through algorithmic control of optimization processes (Yang et al., 2023). The method enables the removal of basic activation patterns which attackers use to create direct connections between their trigger and the targeted class. The learning path analysis enables defenders to detect when models reach their final state at an unusual speed when they process specific suspicious data points which indicates backdoor poisoning (Li et al., 2020). The mathematical procedure

requires physical limitations to decrease the quantity of these surface-level channels which operate at full capacity. The process starts by multiplying skip-connection outputs from essential layers with a tiny decay value γ before conducting a specific training process to regain normal accuracy while keeping the hidden malicious features inactive (Yang et al., 2023). Industrial security systems use separation to protect their architecture from attacks while keeping their models steady when they receive unsteady data input. The model receives protection through defensive distillation (Sun et al., 2022) which trains a second model with the first model's softened probability outputs to create smoother decision boundaries that block gradient-based adversarial attacks (Goodfellow et al., 2015).

The structural obfuscation system offers an extra cryptographic defense which blocks white-box attacks and prevents unauthorized access to integrated model watermarks according to Yan et al. (2023). The real

network structure stays hidden because the system adds extra synaptic links and applies random shuffling to the hidden layer nodes. The system undergoes transformations which create mathematical complexity in its spatial layout that blocks outside observers from identifying specific nodes which attackers might target. The system uses layer balancing according to Ahmadi et al. (2023) to create equal power consumption patterns which protect against side-channel attacks that use power changes to discover weights. The security methods produce an elevated computational challenge for reverse engineering yet they need exact hardware resource optimization to maintain proper inference speeds during real-time operations which include autonomous vehicles and smart grids (Chen et al., 2021).

Security mechanisms for generative models and large language architectures need to follow established industry standards while operating monitoring systems which exist outside the system (figure 1).

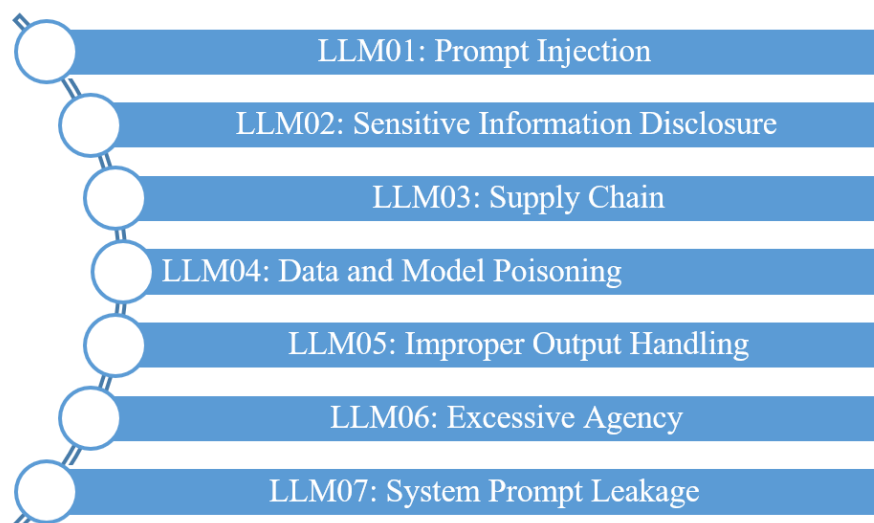


Figure 1. OWASP recommendations for large language models

The textual vulnerability lists and guidelines from OWASP show a shift in their security approach because they now focus on controlling the model execution environment instead of protecting internal weights. The primary focus requires complete separation between context windows and unauthorized prompt injections while training data requires full validation for unaltered condition and access rights must be modified in real time according to OWASP Foundation (n.d.). The existence of these vulnerability vectors proves that generative artificial intelligence security requires outside adaptive filters (guardrails) to perform ongoing detailed analysis of all semantic information which enters and leaves the system. The protection of supply chains has become essential because any security breach in the base model automatically spreads to all smart systems which derive

from it according to Gu et al. (2017). Organizations need to conduct thorough supplier assessments by using evaluation systems which operate like multi-criteria group decision-making methods to assess AI providers and open-source software collections according to Kusi-Sarpong et al. (2018).

Protection systems in financial monitoring and critical infrastructure sectors achieve their full potential when pruning algorithms maintain perfect mathematical precision during their operations. The mind map "Efficiency of Cleaning 'Bad' Neurons" (figure 2) presents the complete conceptual framework which shows how to eliminate nodes that have become compromised.



Figure 2. The Effectiveness of Bad-Neuron Pruning in Practical Systems

The methodology contains multiple essential branches which include the process of defining "bad" neurons and the examination of particular cleaning techniques which involve pruning and regularization and adaptive optimization and the study of security system benefits and challenges and their real-world applications. The WIPER method demonstrates this advanced cleaning technique through its use of specific fine-tuning methods instead of basic pruning. The system operates through an adaptive regularization loss function which performs mathematical operations to achieve its purpose:

$$L_{purifying} = \sum C E(F'(x_i), y_i) + \alpha \sum_{j \in A} A R(w_j)$$

where CE represents the cross-entropy loss, and $AR(w_j)$ applies adaptive regularization specifically to the weights of the infected neurons within subset A. The complete method removes backdoor pathways completely without creating any damage to the standard performance of clean data (Fan et al., 2022). The pruning methods need full interpretability to build trust because they follow the Explainable Artificial Intelligence (XAI) guidelines (Kaya et al., 2026).

The complete dependability of neural network components which operate critical infrastructure systems requires threat modelling that takes proactive measures and scenario-based testing to achieve high levels of reliability (Androshchuk et al., 2020). The socio-cyber-physical complexes of smart cities achieve fast detection and prevention of urban traffic and utility network control system hacking attempts through their deployment of anomaly detection systems (Yevseiev et

al., 2021). Organizations need to establish Security Information Systems through the use of validated corporate methods which Yevseiev and his team developed in their studies from 2019 and 2021. The modern automated organisational and technical systems achieve their highest level of security through a complete system which combines multiple auditing layers with obfuscation techniques and defensive distillation and adaptive monitoring systems.

6. CONCLUSIONS

The fundamental research established that national cybersecurity requires urgent attention to verify the structural reliability of neural network architectures which used to be an academic problem. The current threat landscape reveals that multi-level manipulations through gradient descent and backdoor trigger insertion and context window exploitation in large language models completely destroy the mathematical structure of automated decision systems. The shift to edge and federated learning systems has made the attack surface much larger which enables distributed threat actors to poison global models through database access restrictions. Our research has shown through experiments and theoretical analysis that defense systems which operate as standalone units have run out of their ability to protect effectively. The highest effectiveness emerges when three approaches work together as a complete system which includes algorithmic path blocking during training and topology-based structural protection and targeted computational node removal during inference. The multi-level defense system enables neural networks to develop stronger architectural resilience while keeping

their prediction capabilities intact. The supply chain maintains its stability through the use of strong evaluation metrics which operate similarly to multi-criteria supplier evaluation systems.

The research findings show that practical implementation needs adaptive system monitoring which should work together with OWASP security principles to protect generative artificial intelligence systems. The developed mechanisms establish a mathematically proven defense system which protects high-risk sectors including financial auditing operations and energy infrastructure management and smart cities' socio-cyber-physical systems. The process of detecting hidden network anomalies in its underlying structure enables security teams to isolate malicious indicators which results in both interference detection and model stability during extended cyberattacks. The combination of these solutions into standardized corporate methods establishes a strong foundation which supports secure expansion of autonomous systems.

The strategic outlook for future research should focus on the development of proactive hybrid methods of dynamic protection. Neural network systems of the future require structural plasticity which allows them to independently

develop their network structure during operational time when they encounter unrecognized security threats (zero-day structural attacks). The field needs to establish a vital connection between optimisation algorithms and cryptographic primitives which includes homomorphic encryption methods and defensive distillation and secure multi-party computation. The system will provide mathematical proof which guarantees AI internal parameter protection during every phase of its development cycle thus making gradient leakage attacks completely ineffective.

The digitalization process requires organizations to follow strict ethical guidelines while they must implement Explainable Artificial Intelligence (XAI) principles to achieve safe digitalization. The core of public trust in intelligent systems depends on three fundamental elements which include architectural transparency and the removal of concealed mental prejudices and complete responsibility for automated system choices. The systems operate as black boxes which need to be understood because they control vital public administration operations and worldwide security functions while serving the public interest without showing any dangerous hidden links.

References:

- Ahmadi, M. M., Alrahis, L., Sinanoglu, O., & Shafique, M. (2023). DNN-ALIAS: Deep neural network protection against side-channel attacks via layer balancing. *arXiv preprint arXiv:2303.06746*. <https://doi.org/10.48550/arXiv.2303.06746>
- Androshchuk, A., Yevseiev, S., Melenchuk, V., Lemeshko, O., & Lemeshko, V. (2020). Improvement of project risk assessment methods for implementing automated information components of non-commercial organisational and technical systems. *EUREKA*, 1, 48–55. <https://doi.org/10.21303/2461-4262.2020.001131>
- Chen, Z., Tian, P., Liao, W., & Yu, W. (2021). Towards multi-party targeted model poisoning attacks against federated learning systems. *High-Confidence Computing*, 1(1), 100002. <https://doi.org/10.1016/j.hcc.2021.100002>
- Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., Uszkoreit, J., & Houlsby, N. (2021). An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*. <https://doi.org/10.48550/arXiv.2010.11929>
- Fan, M., Liu, Y., Chen, C., Liu, X., & Guo, W. (2022). Defense against backdoor attacks via identifying and purifying bad neurons. *arXiv preprint arXiv:2208.06537*. <https://doi.org/10.48550/arXiv.2208.06537>
- Geiping, J., Bauermeister, H., Dröge, H., & Moeller, M. (2020). *Inverting gradients – How easy is it to break privacy in federated learning?* *arXiv*. <https://doi.org/10.48550/arXiv.2003.14053>
- Goodfellow, I. J., Shlens, J., & Szegedy, C. (2015). Explaining and harnessing adversarial examples. *arXiv preprint arXiv:1412.6572*. <https://doi.org/10.48550/arXiv.1412.6572>
- Gu, T., Dolan-Gavitt, B., & Garg, S. (2017). BadNets: Identifying vulnerabilities in the machine learning model supply chain. *arXiv preprint arXiv:1708.06733*. <https://doi.org/10.48550/arXiv.1708.06733>
- Jebreel, N. M., Domingo-Ferrer, J., & Li, Y. (2023). Defending against backdoor attacks by layer-wise feature analysis. In *Pacific-Asia Conference on Knowledge Discovery and Data Mining* (pp. 428–440). Springer. <https://doi.org/10.48550/arXiv.2302.12758>
- Kaya, O., Shah, A. F. M. S., Karabulut, M. A., Karahan, S. N., Osmanca, M. S., & Acir, N. (2026). Explainable artificial intelligence (XAI): Concepts, applications, challenges, and future perspectives. *IEEE Access*, 14, 27394–27417. <https://doi.org/10.1109/ACCESS.2026.3663161>
- KELA Cyber Threat Intelligence. (n.d.). OWASP top 10 for LLMs: Protecting GenAI with AiFort. Retrieved from https://www.kelacyber.com/blog/owasp-top-10-for-llms-protecting-genai-with-aifort/?utm_source=chatgpt.com

- Kusi-Sarpong, S., Varela, M. L., Putnik, G., Avila, P., & Agyemang, J. (2018). Supplier evaluation and selection: A fuzzy novel multi-criteria group decision-making approach. *International Journal for Quality Research*, 12(2), 459–486. <https://doi.org/10.18421/IJQR12.02-10>
- Li, Y., Wu, B., Jiang, Y., Li, Z., & Xia, S. T. (2020). Backdoor learning: A survey. *arXiv preprint arXiv:2007.08745*. <https://doi.org/10.48550/arXiv.2007.08745>
- Liu, Z., Luo, Y., Wu, L., Li, S., Liu, Z., & Li, S. Z. (2022). Are gradients on graph structure reliable in grey-box attacks? In *Proceedings of the 31st ACM International Conference on Information & Knowledge Management* (pp. 1360–1368). <https://doi.org/10.48550/arXiv.2208.05514>
- Nguyen, T. D., Rieger, P., Miettinen, M., & Sadeghi, A. R. (2024). Poisoning attacks on federated learning-based IoT intrusion detection system. In *Proceedings of the 2020 Workshop on Decentralized IoT Systems and Security*. <https://doi.org/10.14722/diss.2020.23003>
- OWASP Foundation. (n.d.). OWASP top 10 for large language model applications. Retrieved from https://owasp.org/www-project-top-10-for-large-language-model-applications/?utm_source=chatgpt.com
- Qiao, Y., Luo, Y., Deng, J., & He, J. (2023). "Do anything now": Characterizing and evaluating in-the-wild jailbreak prompts on large language models. *arXiv*. <https://doi.org/10.48550/arXiv.2308.03825>
- Sun, M., Mirman, M., Gfeller, B., & Vechev, M. (2022). VPN: Verification of poisoning in neural networks. *arXiv preprint arXiv:2205.03894*.
- Wang, B., Li, Y., & Zhou, P. (2022). Bandits for structure perturbation-based black-box attacks to graph neural networks with theoretical guarantees. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (pp. 13379–13387). <https://doi.org/10.48550/arXiv.2205.03546>
- Yan, Y., Pan, X., Zhang, M., & Yang, M. (2023). Rethinking white-box watermarks on deep learning models under neural structural obfuscation. In *32nd USENIX Security Symposium* (pp. 2347–2364). <https://doi.org/10.48550/arXiv.2303.09732>
- Yang, S., Li, Y., Jiang, Y., & Xia, S. T. (2023). Backdoor defence via suppressing model shortcuts. In *ICASSP 2023 IEEE International Conference on Acoustics, Speech and Signal Processing* (pp. 1–5). IEEE. <https://doi.org/10.48550/arXiv.2211.05631>
- Yevseiev, S., Alekseyev, V., Balakireva, S., Peleshok, Y., Milov, O., Petrov, O., Rayevnyeva, O., Tomashevsky, B., Tyshyk, I., & Shmatko, O. (2019). Development of a methodology for building an information security system in the corporate research and education system in the context of university autonomy. *Eastern-European Journal of Enterprise Technologies*, 3(99), 49–63. <https://doi.org/10.15587/1729-4061.2019.169527>
- Yevseiev, S., Melenti, Y., Voitko, O., Hrebenuk, V., Korchenko, A., Mykus, S., Milov, O., Prokopenko, O., Sievierinov, O., & Chopenko, D. (2021). Development of a concept for building a critical infrastructure facilities security system. *Eastern-European Journal of Enterprise Technologies*, 3(111), 63–83. <https://doi.org/10.15587/1729-4061.2021.233533>

Oleksandr Kushnerov

Sumy State University,
Sumy,
Ukraine
o.kushnerov@biem.sumdu.edu.ua
ORCID 0000-0001-8253-5698

Mikolaj Karpinski

University of the National Education
Commission,
Krakow,
Poland,
Ternopil Ivan Puluj National Technical
University,
Ternopil,
Ukraine
mikolaj.karpinski@uken.krakow.pl
ORCID 0000-0002-8846-332X

Olena Akhiezer

National Technical University
“Kharkiv Polytechnic Institute”,
Kharkiv,
Ukraine
Olena.Akhiezer@khpi.edu.ua
ORCID 0000-0002-7087-9749

Vladyslav Sokol

National Technical University
“Kharkiv Polytechnic Institute”,
Kharkiv,
Ukraine
Vladyslav.sokol@gmail.com
ORCID 0009-0009-9446-2049

Anna Korchenko

University of the National Education
Commission,
Krakow,
Poland,
Dnipro University of Technology,
Dnipro,
Ukraine
anna.korchenko@uken.krakow.pl
ORCID 0000-0003-0016-1966

Valerii Muzh

Ternopil Ivan Puluj National Technical
University,
Ternopil,
Ukraine
vmuzh@tntu.edu.ua
ORCID 0000-0002-0036-9356
