

UNIVERSITY OF THE NATIONAL EDUCATION COMMISSION, POLAND

TOMAS BATA UNIVERSITY IN ZLIN, CZECH REPUBLIC

NATIONAL UNIVERSITY “KYIV AVIATION INSTITUTE”, UKRAINE



***The First International Workshop for Young  
Scientists in Computer Science and Security***  
**co-located with**  
**5<sup>th</sup> International Conference**  
**on Cyber Hygiene & Conflict Management**  
**in Global Information Networks**

**PROCEEDINGS**  
**of selected papers**

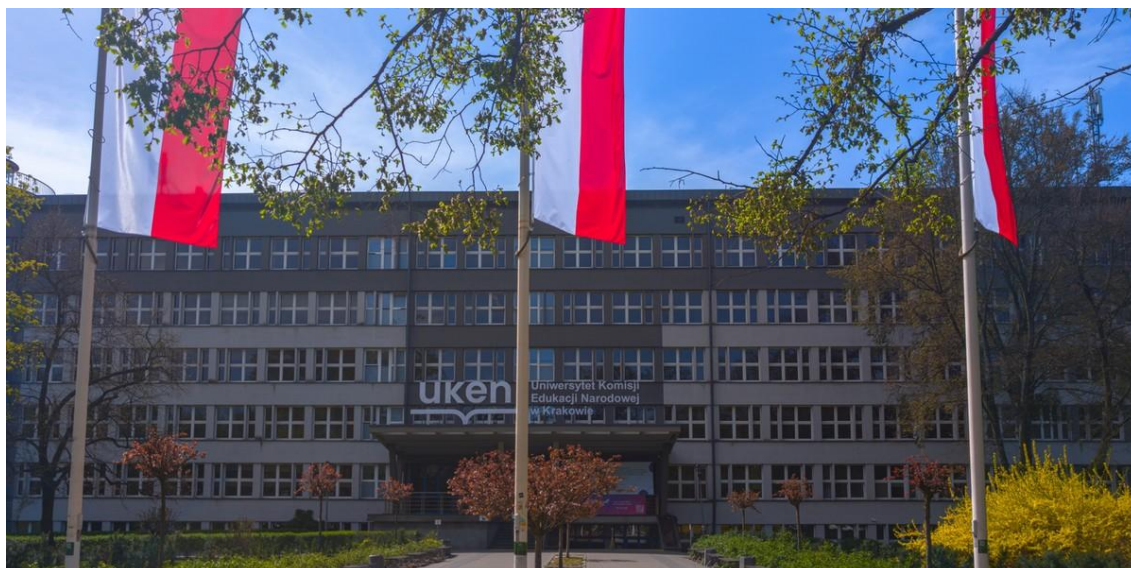


**June 25, 2026**

**Kraków - Kyiv - Zlín**

## **Topics of 5th International Conference on Cyber Hygiene & Conflict Management in Global Information Networks**

- Management of Cyber Incidents, Warfare, and Conflicts
- Modern Information and Communication Technologies
- Secure Digital Content Processing for Modern Applications
- Intelligence Systems of Data Processing
- Secure ICT and Social Services
- AI-based Technologies in Cyber Space
- Legal Aspects of Cyber Security and Digital Resilience
- Conflict Management
- Economic Security Management and Information Protection
- Information, Consciousness, and the Digital Age: An Interdisciplinary Discourse
- Cyber Diplomacy and the Impact of Innovations on Law and International Relations
- Cyber Security and Innovation Technologies in Transport
- Critical Infrastructure Protection and Environmental Safety



### **Workshops**

- Workshop on Cyber Hygiene
- Workshop on Advanced Technologies in Cyber Resilience
- Workshop on Conflict Management in Global Information Networks
- Workshop for Young Scientists in Computer Science and Security

### **Roundtable**

SECURE: Security Education Center for Unified Resilience and Effective Communications (ERASMUS-EDU-2024-CBHE, project id 101179818, 2024-2027) <https://erasmus-secure.eu>

### **Venue**

University of the National Education Commission  
ul. Podchorążych 2, 30-084 Kraków, Poland

## EDGE AI AND COMPUTER VISION FOR ACCESSIBLE, CYBER-SECURE, AND RESILIENT CRITICAL RAILWAY INFRASTRUCTURE

**Kamila Matusik**

University of the National Education  
Commission, Krakow, Poland

**Francesco Flammini**

University of Florence, Florence, Italy

### Abstract

Railway digitalization is increasingly connecting station management, passenger assistance, wayside monitoring, maintenance, identity services, and cloud analytics. This trend can improve capacity, resilience, and accessibility, but it also introduces new cyber, privacy, and assurance risks around safety-critical railway operations. This paper proposes an edge-first architecture for artificial intelligence and computer vision in critical railway infrastructure, explicitly treating AI as an advisory and monitoring layer rather than as a replacement for certified signalling, interlocking, or train-control functions. The architecture places inference, event extraction, privacy filtering, and short-retention buffers on train, station, and wayside edge nodes, while exposing only signed, minimal, policy-labelled events to operational and cloud services. The paper positions the proposal against related work on railway cybersecurity, technology roadmaps, intelligent surveillance, digital twins, passenger-flow modelling, accessibility, and federated learning. It contributes three technical artifacts: a zoned edge-to-cloud reference architecture, a privacy-preserving computer-vision pipeline for accessibility and station-resilience services, and a federated-learning governance loop with validation, rollback, and poisoning controls. The design is mapped to European cybersecurity, AI, accessibility, and data-protection obligations, and is complemented by a before/after Failure Mode and Effects Analysis for edge railway AI assets. The result is a human-in-the-loop, privacy-preserving, cyber-resilient design pattern for inclusive railway AI that maintains a clear safety boundary around certified railway control systems.

### Keywords

Edge AI, railway cybersecurity, computer vision, accessibility, cyber resilience, federated learning, NIS2, data governance, critical infrastructure

## 1. Introduction

European railways are evolving from isolated electromechanical and proprietary control environments into connected cyber-physical systems. The European Rail Traffic Management System (ERTMS), the European Train Control System (ETCS), train-borne diagnostics, station systems, video surveillance, passenger applications, maintenance platforms, and enterprise identity services now coexist with IP networking and high-rate sensing. At the telecommunications layer, the sector is preparing the migration from GSM-R towards the Future Railway Mobile Communication System (FRMCS), a 5G-based successor intended to support future railway critical applications and higher-capacity data services [6, 7].

This modernization has clear benefits. It enables predictive maintenance, better traffic management, more accurate passenger information, and potentially more inclusive transport services. It also creates a new dependency structure. A station crowding alarm may rely on high-resolution video; a maintenance function may rely on remote access to ruggedized devices; a passenger assistance workflow may depend on cloud identity and mobile applications; and a digital twin may need continuous operational telemetry.

If these dependencies are not designed carefully, a disruption of internet-facing services can reduce railway resilience even when train-protection systems remain uncompromised. The challenge is not merely technical. Railway infrastructure managers and railway undertakings are subject to cybersecurity risk-management obligations, safety processes, accessibility duties, and data-protection constraints. The NIS2 Directive strengthens cybersecurity requirements for essential and important entities [4]; CLC/TS 50701 provides sector-specific cybersecurity guidance for railway applications [3]; the EU AI Act introduces a risk-based governance regime for AI systems [5]; rail passenger-rights and accessibility legislation emphasize assistance and access to services [8, 9]; and

the General Data Protection Regulation requires lawful, minimized, and protected processing of personal data [10]. For computer vision, these obligations intersect: cameras can improve safety and accessibility, but raw video streams may also expose location, disability-related attributes, and behavioural patterns. Cloud platforms remain valuable for many railway workloads, including historical reporting, enterprise resource planning, large-scale offline analytics, and model training. However, they are less suitable as the sole runtime dependency for functions that need local continuity, bounded response, minimized video export, or strict separation from safety-related control. Recent disruptions to passenger-facing railway digital services and large-scale attacks against cloud identity workflows show that availability and identity risks should be treated as realistic operating conditions rather than exceptional scenarios [11, 12, 13].

This paper argues for an edge-first architecture for railway AI. The term edge-first does not mean cloud-free. Rather, it means that the first place where sensitive sensor data are interpreted, reduced, and policy-labelled is the local railway edge: onboard gateways, station edge servers, wayside cabinets, depot nodes, or regional operational hubs. The cloud receives only what has a clear purpose and legal basis: aggregated events, telemetry, model updates, audit evidence, or carefully governed datasets. This is particularly important for accessibility services, where the same computer-vision layer can detect assistance-relevant situations while avoiding routine export of raw video.

The objective is not to move certified signalling, interlocking, or train-protection logic into uncertified AI components. The architecture explicitly separates safety-certified control from AI-based advisory, monitoring, and passenger-support functions. AI can provide alarms, decision support, local redaction, occupancy estimates, platform assistance requests, or anomaly indicators. It must not issue direct safety commands unless a separate safety case, certification route, validation process, and operational boundary have been established.

The main contributions are as follows:

1. A structured related-work analysis that connects railway cybersecurity, railway AI, intelligent surveillance, digital twins, passenger-flow resilience, accessibility-oriented mobility research, and federated learning.
2. An edge-first railway AI reference architecture with explicit zoning between certified safety control, operational monitoring, passenger assistance, enterprise cloud services, and model-governance infrastructure.
3. A privacy-preserving computer-vision pipeline for accessible stations and rolling stock, including data minimization, confidence gating, human-in-the-loop escalation, and local fallback behaviour.
4. A cybersecurity and assurance mapping that connects architectural decisions to NIS2, CLC/TS 50701, AI governance, data protection, and accessibility requirements.
5. A risk analysis using Failure Mode and Effects Analysis (FMEA), including before/after Risk Priority Numbers (RPNs) for representative railway edge AI failure modes.

The paper is conceptual and architectural. It does not claim that a particular model, sensor, or vendor platform is certified for safety-critical railway use. Instead, it provides a design pattern, control mapping, and assurance vocabulary that can guide prototypes, procurement specifications, research demonstrators, and future empirical evaluation.

## **2. Related Work**

The literature relevant to edge AI for accessible and cyber-secure railways spans multiple communities. Railway cybersecurity research studies threats to connected signalling, communications, maintenance, and passenger-facing services. Railway AI research investigates maintenance, autonomous operation, traffic management, and passenger mobility. Computer-vision research focuses on inspection, obstacle detection, crowd monitoring, and surveillance. Accessibility and resilience research studies assistance, evacuation, and passenger-flow optimization.

Finally, federated learning and edge computing contribute mechanisms for keeping data local while improving models.

## **2.1. Railway cybersecurity and safety separation**

Recent surveys emphasize that railway cybersecurity is becoming harder because formerly isolated operational systems are increasingly interconnected with enterprise IT, radio networks, supplier access, and passenger services [1, 2]. CLC/TS 50701 addresses cybersecurity for railway applications and complements the established railway safety-engineering culture by focusing on security zones, risk assessment, supplier interfaces, and operational controls [3]. These works support a central assumption of this paper: cybersecurity must be treated as a system property, not as a feature of a single device.

A second relevant line of work concerns safety envelopes for intelligent railway control. Flammini, De Donato, Fantechi, and Vittorini describe intelligent train control as a vision in which automatic train protection can act as a safety envelope while intelligent operation optimizes energy, comfort, and capacity [19]. That distinction is essential for this paper. Edge AI can support monitoring and decision support, but it must remain outside the certified safety-control boundary unless it is itself subject to an appropriate safety case.

## **2.2. AI, computer vision, and digital twins in railways**

The RAILS project and associated publications provide a strong foundation for AI adoption in the railway sector. Bešinović et al. propose a taxonomy of AI in railway transport and discuss regulations and applications across maintenance, autonomous driving, traffic management, and related domains [14].

Tang et al. review AI applications in railway systems and find strong attention to maintenance and inspection, with emerging work on safety, traffic management, and passenger mobility [15]. De Donato et al. focus on audio-video defect detection through deep learning in railway maintenance, showing that smart acoustic and visual sensors are central to predictive maintenance [16]. Public datasets for railway AI have also been reviewed, with emphasis on the value and limitations of open data for reproducible research [17].

Computer vision is also prominent in intelligent infrastructure surveillance. Pappaterra and Flammini review intelligent infrastructure surveillance for safe autonomy in smart railways, including obstacle detection, trespassing detection, natural-threat detection, and situational awareness [18]. Earlier work on SHIELD demonstrates a layered approach to security, privacy, and dependability in smart railway surveillance, including node, network, middleware, and overlay mechanisms [23]. These publications motivate the layered architecture proposed here, but our focus differs in two respects: we explicitly combine accessibility services with cybersecurity assurance, and we make local event minimization a primary design rule.

Digital twins are another relevant enabler. De Donato et al. propose preliminary guidelines and a reference architecture for AI-assisted digital twins in smart railways, with predictive maintenance as a key application [20]. RAILS roadmapping work identifies safety and automation, predictive maintenance and defect detection, and traffic planning and management as major areas for intelligent railway development [22]. In this paper, digital twins are used primarily as an assurance and validation environment: candidate edge models should be tested against scenarios, perturbations, and degraded modes before signed rollout.

## **2.3. Passenger flow, accessibility, and resilience**

Accessibility cannot be reduced to compliance documents; it is an operational property of stations, trains, digital services, staff workflows, elevators, ramps, information displays, and emergency procedures.

Railway passenger-rights legislation and the European Accessibility Act require operators and digital services to support inclusive access [8, 9]. From a technical perspective, passenger-flow estimation and accessibility-aware routing can help operators predict assistance needs, manage crowding, and maintain safe evacuation options.

Hangli Ge and co-authors have contributed several relevant works. Multi-weighted graph learning and K-neighboring graph modelling have been used for passenger count prediction on railway networks [27]. Emergency evacuation and traffic evacuation models for railway disruption study resilience under station or line failures, including capacity constraints and alternative routing [25, 26]. Earlier work on accessibility semantics for indoor mobility demonstrates that routing should encode userrelevant constraints rather than only geometric shortest paths [28]. These ideas are complementary to edge computer vision: local sensors can generate occupancy, obstruction, or assistance events, while graph-based and accessibility-aware models can transform those events into operational decisions.

## **2.4. Federated learning and edge model governance**

Federated learning provides a mechanism for improving models without centralizing raw data. The classic FedAvg algorithm trains local models and aggregates updates rather than raw records [29].

Secure aggregation reduces the information visible to the coordinator by cryptographically protecting individual model updates [30]. These mechanisms are attractive for railway use cases because station info, passenger-assistance events, and infrastructure logs can be sensitive. However, federated learning is not automatically private or secure. Updates may leak information, compromised nodes can poison the model, and heterogeneous edge devices can introduce fairness and reliability issues. Therefore, this paper treats federated learning as one part of a broader governance loop that includes validation, anomaly detection, participation control, model signing, rollback, and audit.

## **3. Problem Statement and Threat Model**

The problem addressed in this paper is the design of an AI-enabled monitoring and assistance layer for railway infrastructure that satisfies five simultaneous requirements: local continuity, safety separation, cybersecurity resilience, privacy-preserving processing, and accessibility support. The design must work under intermittent wide-area connectivity, constrained maintenance windows, heterogeneous suppliers, and regulatory scrutiny.

### **3.1. Assets and data flows**

The relevant assets include onboard sensors, CCTV cameras, platform cameras, wayside sensors, wireless gateways, station edge servers, maintenance laptops, operator workstations, staff mobile devices, passenger-assistance applications, cloud analytics platforms, identity providers, certificate authorities, and model repositories. The most sensitive data flows are raw video, passenger-assistance requests, location and occupancy events, operational topology, device configuration, model parameters, model updates, and maintenance credentials.

The architecture distinguishes four classes of outputs. Safety-control outputs directly influence train separation, interlocking, braking, route setting, or signalling. These outputs are outside the scope of the proposed AI layer. Operational advisory outputs include anomaly alarms, defect indicators, or sensor health alerts. Passenger-service outputs include assistance requests, crowding information, platform guidance, lift-status routing, and staff notifications. Governance outputs include logs, model provenance, risk scores, and audit evidence.

**Table 1.** Positioning of the proposed contribution relative to related work

| Literature cluster                                          | Main contribution to the field                                                                   | Gap addressed by this paper                                                                                                       |
|-------------------------------------------------------------|--------------------------------------------------------------------------------------------------|-----------------------------------------------------------------------------------------------------------------------------------|
| Railway cybersecurity surveys and CLC/TS 50701 [1, 2, 3]    | Threats, risk-management processes, security zoning, and sector-specific cybersecurity concerns. | Converts these principles into an edge-AI deployment pattern with concrete node, data, model, and maintenance controls.           |
| Railway AI taxonomy and RAILS roadmaps [14, 15, 22]         | AI application taxonomy, regulations, research directions, and smart-railway roadmaps.           | Adds cybersecurity, accessibility, privacy, and cloud-dependency constraints as first-class architectural requirements.           |
| Audio-video maintenance and surveillance [16, 18, 23]       | Computer vision, audio analytics, surveillance, and security/privacy/dependability layers.       | Extends surveillance from inspection to inclusive passenger assistance while enforcing event minimization and human confirmation. |
| Digital twins and intelligent train control [19, 20]        | Safety-envelope thinking, validation environments, and AI-assisted railway digital twins.        | Uses digital twins as validation gates for edge model releases rather than as unrestricted real-time controllers.                 |
| Passenger-flow and accessibility modelling [27, 25, 26, 28] | Graph-based flow prediction, disruption evacuation, and accessibility-aware path semantics.      | Connects local computer-vision events with resilient passenger-assistance and station-management workflows.                       |
| Federated learning [29, 30]                                 | Distributed model training and secure aggregation without centralizing raw data.                 | Embeds federated learning in a railway-specific assurance loop with model signing, poisoning checks, and rollback.                |

### 3.2. Adversaries and failure modes

The threat model includes remote attackers, credential phishers, malicious or compromised suppliers, ransomware operators, botnets conducting denial of service, local attackers with limited physical access to station equipment, compromised edge devices, and honest-but-misconfigured administrators. It also includes non-malicious failures: sensor occlusion, network outage, model drift, weather effects, camera misalignment, device overheating, and staff overload.

Recent incidents illustrate two relevant attack patterns. A DDoS attack can disrupt passenger-facing services without directly compromising train control; Deutsche Bahn reported a DDoS attack affecting IT systems while restoring access to digital channels [11]. Separately, token-replay and device-code phishing campaigns show how cloud identity can become an entry point for enterprise compromise [12, 13]. These examples should not be misrepresented as attacks on signalling. Their importance is that railway operators increasingly depend on the same internet-facing services, identity providers, and administrative patterns as other critical infrastructure organizations.

### 3.3. Assumptions and exclusions

The proposed architecture assumes that certified signalling, interlocking, and train-protection systems remain governed by established railway safety standards and are not modified by the AI layer. It assumes that edge nodes can receive sensor feeds and publish advisory events through approved interfaces. It also assumes that staff remain responsible for passenger assistance and that AI outputs are designed to support, not replace, human judgment.

The architecture does not solve all supply-chain risks. It cannot make a poorly trained model reliable, a weak physical enclosure secure, or an unlawful data-processing purpose lawful. It reduces risk by making dependencies explicit and by limiting the blast radius of failures. It also provides a structure for procurement and assurance: each node, model, data flow, and maintenance function must have an owner, configuration baseline, update mechanism, and audit trail.

## 4. Edge-First Reference Architecture

An edge-first architecture places the first layer of interpretation and enforcement close to the sensors and operational context. In railway systems, the edge is not a single location. It includes train-borne gateways, station equipment rooms, platform cabinets, depot servers, level-crossing cabinets, regional traffic-management facilities, and security operations interfaces. Figure 1 shows the proposed reference architecture.

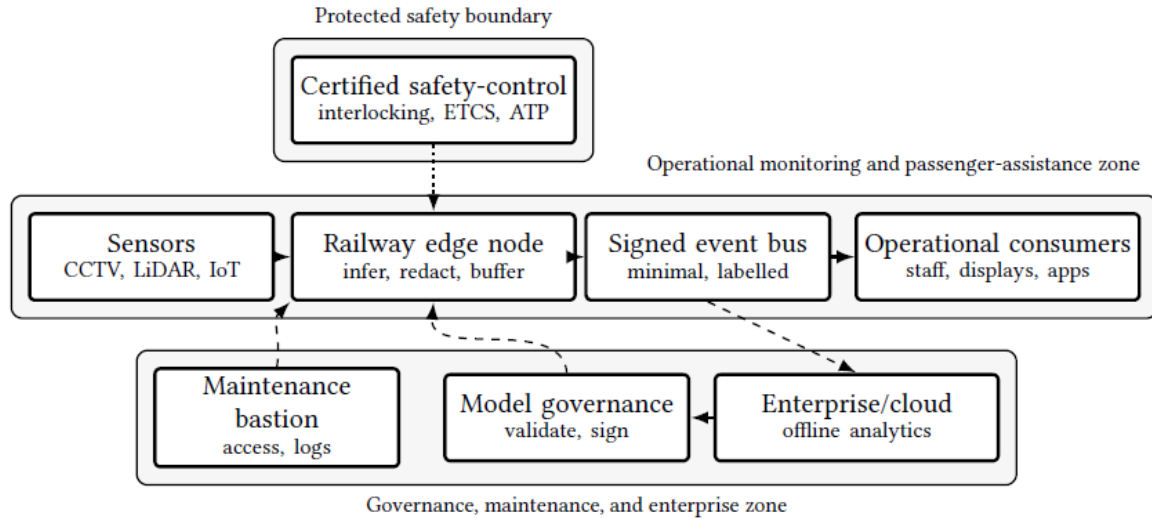


**Figure 1.** Edge-first railway AI architecture. Raw sensor data are processed locally. The normal northbound interface is a signed event bus, not a continuous raw-video export. The protected safety-control zone is observed only through approved interfaces, and AI outputs remain advisory unless separately certified.

### 4.1. Design principles

The architecture follows seven design principles.

P1: Safety non-interference. AI services must not degrade the availability, timing, or integrity of certified safety-control systems. Interfaces to signalling, interlocking, or train-protection zones must be read-only, unidirectional, or mediated by certified gateways unless a separate safety case allows otherwise.

P2: Local-first inference. Sensor streams that contain personal data or security-sensitive operational detail should be processed locally. Raw export is disabled by default and enabled only for defined purposes such as incident investigation, supervised annotation, or legal evidence.

P3: Minimal events. Edge nodes convert high-volume sensor streams into small, signed, purposespecific events. An event should state what was inferred, with what confidence, at which place and time, under which model version, and under which data-retention policy.

P4: Human confirmation for consequential assistance. Accessibility services should avoid fully automated assumptions about passengers. For example, an edge model may detect a wheelchair user near a boarding location, but staff notification or passenger interaction should respect consent, dignity, and false-positive handling.

P5: Managed degradation. Each service must define degraded modes. During loss of wide-area connectivity, station-edge nodes should continue local occupancy estimation, lift-status display, and staff alerts. During sensor failure, staff workflows should receive an uncertainty notice rather than a silent failure.

P6: Signed and reversible change. Edge models and configurations must be signed, versioned, tested, and reversible. Rollback is a safety and cybersecurity control because a defective or poisoned model can be removed quickly.

P7: Evidence by design. Logs, model provenance, policy decisions, and maintenance actions should be captured in a form usable for audits, incident response, and post-event learning. Evidence must be protected from tampering and excessive retention.

## **4.2. Zoning and interfaces**

A railway station or corridor should be decomposed into zones with different trust assumptions. The safety-control zone contains certified systems. The operational monitoring zone contains sensors, AI inference nodes, event brokers, staff workstations, and local displays. The passenger-service zone contains mobile applications, accessibility requests, journey information, and public APIs. The enterprise zone contains identity, analytics, ticketing, and reporting services. The maintenance zone contains bastions, update repositories, configuration management, and audit systems.

The most important architectural choice is the interface between zones. Continuous raw video export from stations to a remote cloud increases privacy exposure and bandwidth dependency. Direct remote administration of edge nodes increases credential and supply-chain risk. A safer approach is a twoplane interface. The runtime plane carries signed events, health status, and bounded commands. The maintenance plane carries updates, diagnostics, and administrative actions through a controlled bastion with just-in-time access, session recording, multi-factor authentication resistant to token replay, and offline rollback packages.

## **4.3. Data contracts for edge events**

Every edge event should follow a data contract. The contract should include an event type, source node, location granularity, timestamp, model identifier, confidence score, uncertainty code, personal-data classification, retention class, and permitted consumers. For accessibility, examples include `assistance.requested`, `boarding.area.obstructed`, `lift.unavailable`, `platform.crowding.high`, and `wheelchair.space.occupied`. For cybersecurity and maintenance, examples include `camera.occluded`, `edge.node.tamper`, `model.drift.suspected`, `sensor.clock.skew`, and `unauthorized.maintenance.attempt`.

This data-contract discipline turns the edge layer into a policy enforcement point. It also helps avoid a common failure in AI deployments: treating model outputs as informal logs rather than operational messages. A signed event with a clear schema can be validated, routed, retained, challenged, audited, and replayed in a test environment.

# **5. Computer Vision for Accessibility and Station Resilience**

Accessibility-oriented computer vision should be designed around assistance outcomes rather than surveillance accumulation. The purpose is to support passengers and staff: detect obstructions, estimate crowding, identify unavailable facilities, improve boarding support, and preserve service under disruptions. These functions can be valuable even when internet-facing passenger applications are unavailable.

## **5.1. Privacy-preserving event pipeline**

Figure 2 shows a local computer-vision pipeline for accessible stations and rolling stock. The pipeline starts with frame acquisition and local calibration. Privacy filtering then removes or masks information not required for the task. Depending on the use case, the filter may blur faces, crop to infrastructure zones, discard frames after short buffering, or process only silhouettes and occupancy maps. The model then performs task-specific inference. A confidence gate decides whether to issue a local event, request human confirmation, or suppress uncertain output. The event broker routes the signed event to staff devices, displays, lifts, mobile applications, or operational dashboards.

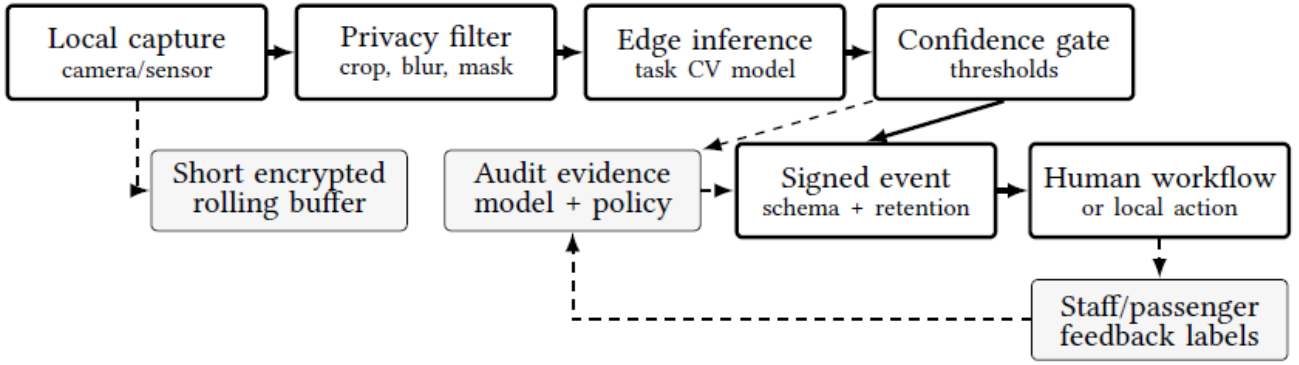


**Figure 2.** Privacy-preserving computer-vision pipeline for accessibility services. The default output is a minimal event, while raw video remains local and short-lived unless a defined legal and operational purpose justifies retention.

## 5.2. Accessibility use cases

Table 2 summarizes representative services. The table intentionally avoids claiming that all services have the same criticality. A lift-status notification and a platform-edge intrusion alarm have different assurance needs. The architecture supports both, but the safety and cybersecurity cases differ. The key ethical issue is that accessibility support should not become intrusive classification of passengers.

Many impairments are not visible, and not every passenger using a wheelchair, cane, stroller, or assistance device wants automated intervention. The system should therefore privilege environment-centred events: blocked ramp, high crowding, unavailable lift, occupied wheelchair space, open boarding gap, or staff assistance request. When passenger-centred inference is necessary, it should be confidence-limited, consent-aware, and mediated by staff or passenger controls.

**Table 2.** Representative edge-AI services for accessibility and station resilience

| Service                                  | Local input                                        | Output                              | Key controls and rationale                                                                                                   |
|------------------------------------------|----------------------------------------------------|-------------------------------------|------------------------------------------------------------------------------------------------------------------------------|
| Boarding assistance cue                  | Platform camera, staff terminal, timetable context | Staff alert or confirmation request | Human confirmation; no biometric identification; event retention tied to assistance workflow; fallback to manual call point. |
| Wheelchair-space occupancy               | Onboard or platform-side camera, seat/space sensor | Occupancy and availability event    | Local inference and aggregated status; avoid passenger identity; display uncertainty when camera is blocked.                 |
| Obstacle or obstruction detection        | Platform, lift entrance, ramp, corridor camera     | Obstruction event                   | Local processing; short encrypted buffer; escalation only when obstruction persists or confidence is high.                   |
| Crowding and flow estimation             | Overhead camera, people counter, gate count        | Density class and trend             | Aggregate counts; calibration by time of day and event context; supports safe routing and staff deployment.                  |
| Lift and escalator accessibility support | IoT telemetry, CCTV confirmation, maintenance logs | Accessible route availability       | Combine sensor status with staff validation; publish to local displays and apps even during WAN outage.                      |
| Disruption evacuation support            | Station counts, network graph, disruption status   | Routing and capacity advisory       | Integrates local counts with graph-based flow and evacuation models; advisory to operations, not automatic crowd control.    |

## 5.3. Resilience during degraded connectivity

The value of edge processing becomes clearest during degraded connectivity. If a cloud-hosted passenger app is unavailable, station edge nodes can still publish local lift status to displays, send staff alerts over local networks,

maintain short-term crowding estimates, and queue events for later synchronization. If the identity provider is degraded, local staff devices can operate under pre-provisioned emergency credentials with strict time limits and post-event reconciliation. If a camera fails, the system can issue a sensor-health event rather than silently suppressing assistance cues.

This is not a call to isolate all railway systems permanently. Rather, the architecture defines which functions must survive cloud loss and which can safely wait. Historical dashboards can wait. A local platform obstruction alert should not. A model retraining job can wait. A staff notification that a lift is unavailable may be needed immediately.

## 6. Cybersecurity, Data Protection, and Assurance Mapping

The proposed architecture is meant to be auditable. Table 3 maps core design decisions to cybersecurity, AI governance, accessibility, and data-protection objectives. The mapping is not a substitute for a formal compliance assessment, but it provides a starting point for requirements engineering and procurement.

### 6.1. Cybersecurity controls

The edge nodes should be treated as operational-technology assets. Each node should have secure boot, measured boot where feasible, disk encryption, hardware-backed key storage, least-privilege services, network segmentation, local firewalling, time synchronization, tamper evidence, and remote attestation where practicable. Maintenance should occur through a bastion rather than direct internet exposure. Updates should be staged: test node, pilot station, regional rollout, fleet rollout. Rollback should be rehearsed, not merely documented.

Credential design is especially important. Token replay and device-code phishing exploit the gap between user authentication and session control. Railway maintenance workflows should therefore avoid long-lived bearer tokens on field laptops, use phishing-resistant authentication where feasible, bind administrative sessions to managed devices, and require explicit approval for privileged actions. Emergency access is sometimes necessary, but it should be time-limited and reconciled after the incident.

**Table 3.** Control mapping for railway edge AI

| Design decision                                         | Risk addressed                                                                               | Assurance evidence to collect                                                                                            |
|---------------------------------------------------------|----------------------------------------------------------------------------------------------|--------------------------------------------------------------------------------------------------------------------------|
| Safety-control separation and one-way status interfaces | AI-induced interference with certified signalling or train-protection functions.             | Interface control document; zone model; threat analysis; safety non-interference argument; test reports.                 |
| Local-first processing and event minimization           | Exposure of personal data, operational topology, and high-volume video streams.              | Data-protection impact assessment; event schemas; retention rules; proof that raw export is disabled by default.         |
| Signed model and configuration releases                 | Malicious update, model tampering, accidental downgrade, or untracked model drift.           | Software bill of materials; signature records; model registry; validation report; rollback test.                         |
| Secure maintenance bastion                              | Supplier compromise, credential replay, undocumented configuration change.                   | Access logs; session recording; just-in-time authorization; MFA policy; emergency-access procedure.                      |
| Health monitoring and degraded-mode plans               | Silent failures of cameras, sensors, edge nodes, or WAN connectivity.                        | Service-level objectives; watchdog logs; failover drills; local fallback evidence; operator training records.            |
| Human-in-the-loop accessibility workflow                | Harmful automation, false positives, or intrusive profiling of passengers.                   | User-impact assessment; staff procedure; false-positive handling; passenger information notices; complaint path.         |
| Federated learning with validation gates                | Raw data centralization, poisoned updates, privacy leakage, and heterogeneous model quality. | Participant list; secure aggregation configuration; update anomaly checks; digital-twin validation; staged rollout logs. |

## 6.2. Data protection and accessibility

For accessibility services, the data-protection objective is not to avoid all processing. It is to process the minimum data necessary to provide a legitimate, useful service and to protect passengers from avoidable exposure. A platform crowding estimate can often be generated without retaining identifiable images. A lift obstruction alert may not require storing any frame after the event is resolved. A boarding-assistance event may require location and time, but not a persistent passenger profile.

Transparency matters. Passengers should know when automated sensing supports accessibility and station safety, what kinds of events are generated, whom to contact, and how long information is retained. Staff should understand model limitations: occlusion, unusual lighting, weather, crowd density, rare mobility aids, adversarial signs, and edge-device failure. The human workflow is therefore part of the technical architecture.

## 6.3. AI governance

The EU AI Act is risk-based, and railway-related AI may fall into different categories depending on the function. A model used for station occupancy analytics has different implications from a model that influences safety-critical movement authority. The architecture reduces governance ambiguity by classifying AI functions by consequence: informational, operational advisory, passenger assistance, safety-adjacent advisory, and safety-control. Each class should have different validation, monitoring, and approval requirements.

A practical AI-governance file for each edge model should include training data summary, intended use, prohibited use, performance metrics, known limitations, fairness and accessibility considerations, cybersecurity dependencies, model-card information, validation scenarios, drift indicators, and rollback instructions. For models used across stations, the file should also include evidence that performance remains acceptable under different lighting, weather, camera angles, crowd densities, and cultural contexts.

## 7. Federated Learning and Model Lifecycle Governance

Federated learning can reduce raw-data centralization, but railway deployments need more than an aggregation algorithm. They need a model lifecycle with technical and organizational gates. Figure 3 shows the proposed loop.

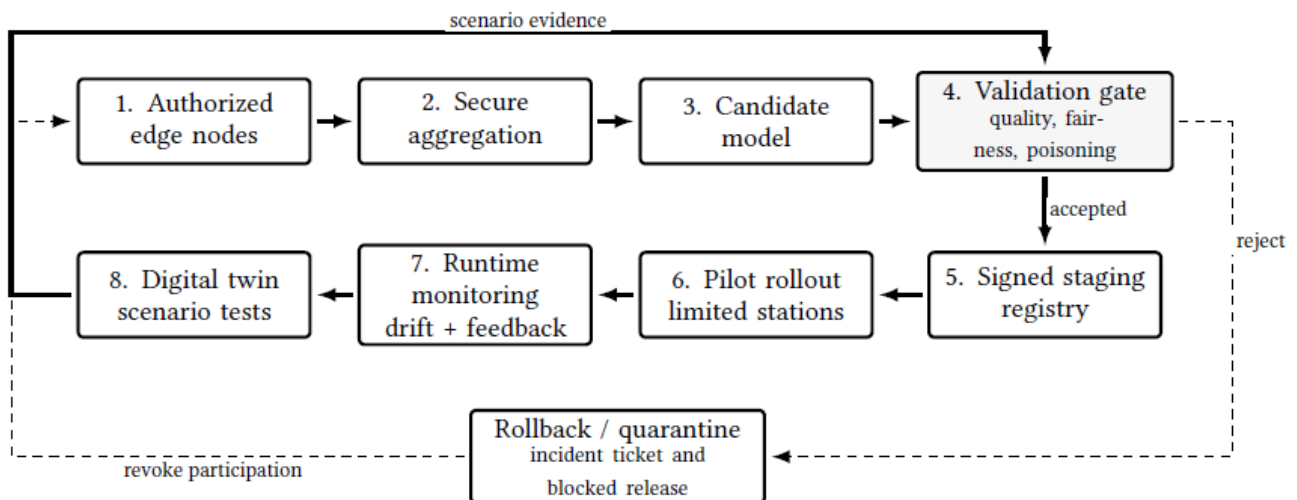


**Figure 3.** Federated-learning and model-governance loop. The solid arrows show the accepted release path; dashed arrows route rejected updates and rollback actions outside the main loop. Secure aggregation is necessary but insufficient; railway deployment also requires validation, staged rollout, monitoring, scenario testing, and rollback.

## 7.1. Local learning and participation control

Only authorized nodes should participate in federated rounds. Participation should depend on attestation status, software version, sensor health, local data-quality checks, and operational state. A compromised node should not be able to submit arbitrary updates. Participation control can use allow-lists, device certificates, update magnitude checks, and quarantine procedures.

## 7.2. Privacy and poisoning controls

Secure aggregation protects individual updates from the coordinator, but it does not by itself prevent poisoning. Conversely, update inspection can detect anomalies but may reduce privacy. The design must balance these goals. A practical railway approach is tiered. Low-risk models can use standard secure aggregation and aggregate validation. More consequential models should use robust aggregation, differential privacy where appropriate, held-out validation sets, scenario testing, and red-team poisoning exercises. The output of a federated round should be a candidate model, not an automatic production release.

## 7.3. Digital-twin validation and rollback

Digital twins can be used to test new models against rare but critical scenarios: crowd surges, smoke, platform reflections, low light, snow, wheelchair boarding, camera occlusion, cyber-induced sensor dropout, and communication delay. A candidate model that performs well on average but fails under a particular accessibility scenario should not be rolled out blindly. Rollback must be supported at the node and fleet levels. Each edge node should store the last known good model and configuration, and the operator should know how to revert without waiting for cloud availability.

## 8. Risk Analysis Using FMEA

FMEA is useful for translating architectural ideas into operational risks. The standard Risk Priority Number is calculated as

$$RPN = S \times O \times D,$$

where S is severity, O is occurrence, and D is detection difficulty. Scores from 1 to 10 are ordinal and should be calibrated locally. The values in Table 4 are illustrative. They are intended to show how architectural controls can reduce occurrence or improve detection; they are not a substitute for operator-specific risk assessment.

The table highlights two points. First, the highest-risk failures are not always algorithmic. Weak maintenance access, raw data export, and poor update control can be as dangerous as low model accuracy. Second, risk reduction often comes from improving detectability. A model will sometimes be wrong; the operator should know when confidence is low, when a sensor is blocked, and which model version produced an event.

## 9. Deployment Roadmap and Evaluation Criteria

A high-quality deployment should be incremental. The following roadmap is designed to minimize safety and social-acceptance risk while generating evidence.

### 9.1. Phase 1: inventory and zoning

Operators should start by inventorying sensors, edge devices, data flows, remote-access paths, model dependencies, and passenger-assistance workflows. The output is a zone-and-conduit model, data classification, and a list of services that must survive WAN outage. This phase often reveals hidden dependencies: a local display

may depend on a cloud API; a maintenance laptop may have broad privileges; or a camera feed may be copied to multiple systems without a clear retention policy.

**Table 4.** Illustrative before/after FMEA for railway edge-AI deployment

| Failure mode                             | Potential effect                                                                | Initial RPN                  | Key mitigation                                                                           | Residual RPN                |
|------------------------------------------|---------------------------------------------------------------------------------|------------------------------|------------------------------------------------------------------------------------------|-----------------------------|
| Raw video exported by default            | Personal-data exposure, legal risk, and larger attack surface.                  | $8 \times 5 \times 6 = 240$  | Local redaction; event-only export; retention classes; data-contract review.             | $8 \times 2 \times 3 = 48$  |
| Compromised model update                 | False alarms, missed events, or malicious inference across many nodes.          | $8 \times 4 \times 7 = 224$  | Signed releases; registry; staged rollout; validation gates; rollback.                   | $8 \times 2 \times 3 = 48$  |
| WAN outage during disruption             | Loss of passenger-assistance and crowding services when most needed.            | $7 \times 5 \times 5 = 175$  | Local inference; local displays; queued events; emergency credentials.                   | $7 \times 3 \times 3 = 63$  |
| Camera occlusion or misalignment         | Missed obstruction, inaccurate occupancy, or false staff confidence.            | $6 \times 6 \times 6 = 216$  | Sensor-health checks; confidence codes; calibration; staff feedback.                     | $6 \times 3 \times 3 = 54$  |
| Token replay against maintenance account | Unauthorized configuration change or data exfiltration from edge nodes.         | $9 \times 4 \times 6 = 216$  | Bastion; phishing-resistant MFA; device binding; session recording; just-in-time access. | $9 \times 2 \times 3 = 54$  |
| Model drift under local conditions       | Degraded performance in weather, lighting, station geometry, or crowd patterns. | $7 \times 5 \times 7 = 245$  | Drift monitoring; local validation sets; digital-twin scenarios; federated feedback.     | $7 \times 3 \times 3 = 63$  |
| False accessibility inference            | Unwanted intervention, passenger discomfort, or missed assistance.              | $6 \times 4 \times 6 = 144$  | Environment-centred events; consent-aware workflow; human confirmation; labels.          | $6 \times 2 \times 3 = 36$  |
| AI output crosses into safety control    | Uncertified behaviour affects train movement or signalling-adjacent actuation.  | $10 \times 2 \times 8 = 160$ | Zone separation; interface control; one-way status; non-interference tests.              | $10 \times 1 \times 2 = 20$ |

## 9.2. Phase 2: non-intrusive local analytics

The first technical pilots should focus on non-safety, non-biometric, environment-centred analytics: lift status, camera health, platform crowding class, local redaction, obstruction persistence, and staff dashboard integration. The aim is to validate edge operations, not to maximize model complexity.

Metrics should include latency, uptime during WAN loss, false alarms, missed events, staff response time, retained data volume, and number of raw-video exports avoided.

## 9.3. Phase 3: accessibility workflow integration

After local analytics are stable, operators can integrate passenger-assistance workflows. This phase requires involvement of passengers with disabilities, staff representatives, data-protection officers, safety engineers, and cybersecurity teams. Technical metrics should be complemented by human metrics: perceived usefulness, perceived intrusiveness, staff workload, response equity across stations, and complaint handling.

## 9.4. Phase 4: federated learning and digital-twin validation

Only after reliable monitoring and feedback exist should operators introduce federated learning. The goal is to improve models across stations without centralizing raw data. Candidate models should be validated against station-specific and synthetic scenarios, including edge cases relevant to accessibility. Rollout should be limited and reversible.

## 9.5. Phase 5: safety-adjacent advisory, if justified

Some use cases, such as obstacle detection near tracks or degraded-mode situational awareness, may be safety-adjacent. These should be approached only after the operator can demonstrate mature zoning, evidence capture, validation, and change control. Even then, the AI output should remain advisory unless a formal safety case justifies stronger integration.

**Table 5.** Evaluation criteria for pilots and full deployment

| Criterion              | Example metric                                                                       | Why it matters                                                                      |
|------------------------|--------------------------------------------------------------------------------------|-------------------------------------------------------------------------------------|
| Operational continuity | Percentage of core assistance functions available during WAN outage.                 | Verifies that edge-first design reduces cloud availability dependency.              |
| Privacy minimization   | Raw-video export rate; retention duration; number of event-only workflows.           | Demonstrates compliance with data minimization and purpose limitation.              |
| Cyber resilience       | Patch latency; failed login attempts; unsigned update rejection; recovery time.      | Tests the maintainability and compromise containment of distributed nodes.          |
| Model quality          | False-positive and false-negative rates by station, time, lighting, and crowd state. | Prevents aggregate performance from hiding local or accessibility-related failures. |
| Human workflow         | Staff response time; user satisfaction; complaint rate; override frequency.          | Ensures that AI supports rather than burdens passengers and staff.                  |
| Assurance evidence     | Percentage of events with model version, policy label, and traceable audit record.   | Supports incident investigation, regulatory audit, and continuous improvement.      |

## 10. Discussion

### 10.1. Why edge-first does not mean anti-cloud

A common misunderstanding is that edge-first architecture rejects the cloud. The opposite is true: it clarifies the cloud role. Cloud services are appropriate for offline fleet analytics, long-term trend analysis, enterprise integration, model repository management, and large-scale simulation. They are problematic when used as the only runtime path for local, privacy-sensitive, or time-sensitive functions.

The proposed architecture therefore separates runtime inference from fleet learning. The edge produces bounded events; the cloud improves long-term knowledge.

### 10.2. Accessibility as a resilience property

Accessibility is often discussed as a service-quality or legal-compliance issue. In railways, it is also a resilience property. During disruption, passengers with reduced mobility may be disproportionately affected by unavailable lifts, platform changes, crowding, inaccessible replacement transport, or inaccessible digital information. Edge AI can help by keeping local displays, staff alerts, and accessibility routing available even when external services degrade. The work by Ge and co-authors on passengerflow prediction and evacuation modelling suggests that local observations should feed wider resilience models, while accessibility-semantics research reminds us that the shortest route is not necessarily an accessible route [27, 25, 26, 28].

### 10.3. The certification boundary

The safest way to introduce AI in railways is not to pretend that every AI output is safety-certified. A more realistic route is layered deployment. First, use AI for local redaction, monitoring, and passenger information. Then use it for staff decision support and maintenance prioritization. Only after evidence accumulates should safety-adjacent

advisory functions be considered. Direct safety control requires a different assurance argument. This position is consistent with the safety-envelope view of intelligent train control [19].

#### **10.4. Operational burden of distributed intelligence**

Edge AI reduces some risks but creates others. A centralized service may be easier to patch; a fleet of edge devices may be harder to inventory, monitor, and physically protect. Ruggedized devices can overheat, lose synchronization, or be installed with inconsistent camera angles. These risks are manageable only if the edge fleet is treated as critical infrastructure. Configuration drift, model drift, and access drift should be monitored as carefully as network availability.

#### **10.5. Limitations**

This paper has four main limitations. First, the FMEA scores are illustrative and must be recalibrated using operator-specific data, incident history, and field trials. Second, the paper does not benchmark a specific computer-vision model. Model selection depends on cameras, station geometry, lighting, weather, and operational objectives. Third, the legal analysis is architectural rather than jurisdiction-specific legal advice. Operators must perform their own data-protection and AI-governance assessments. Fourth, social acceptance cannot be solved by technology alone. Passengers, staff, unions, disability organizations, safety assessors, cybersecurity teams, and regulators should be involved early.

#### **10.6. Research directions**

Future research should develop benchmark datasets for accessibility-oriented railway perception, not only for defects and obstacles. Datasets should represent wheelchairs, mobility aids, prams, luggage, crowding, low light, weather, lift areas, ramps, and platform-train interfaces while respecting privacy. Another direction is run-time assurance: combining model confidence, sensor health, environmental context, and formal zone constraints to decide when an AI event is trustworthy enough for a given workflow. A third direction is federated and privacy-preserving learning under non-identically distributed station data. Finally, graph-based evacuation and passenger-flow models should be integrated with real-time edge observations to support disruption management.

### **11. Conclusions**

Railway digitalization must improve resilience, cybersecurity, and accessibility at the same time. Cloud platforms and AI services can support this objective, but they should not become unexamined dependencies for time-sensitive, privacy-sensitive, or cyber-physical railway functions. This paper proposed an edge-first architecture for deploying AI and computer vision in critical railway infrastructure while preserving a clear safety boundary around certified signalling and train-protection systems. The proposed framework combines local inference, event minimization, secure maintenance, signed model releases, federated learning governance, digital-twin validation, and human-in-the-loop accessibility workflows. The related-work analysis connects railway cybersecurity, RAILS roadmaps, intelligent surveillance, digital twins, passenger-flow resilience, and federated learning. The control mapping and FMEA translate the architecture into auditable requirements and risk-reduction mechanisms.

The central message is practical: edge AI should be treated as an advisory, monitoring, privacy-filtering, and assistance layer that strengthens railway operations without silently weakening safety assurance or data protection. When deployed incrementally and governed rigorously, it can help operators provide more inclusive and cyber-resilient rail services under European regulatory and operational constraints.

#### **Declaration on Generative AI**

OpenAI ChatGPT was used for editorial and LaTeX assistance. The authors verified the final manuscript.

## References

- [1] S. Soderi, D. Masti, and Y. Z. Lun, "Railway cyber-security in the era of interconnected systems: A survey," *IEEE Transactions on Intelligent Transportation Systems*, vol. 24, no. 7, pp. 6764–6779, 2023, doi: 10.1109/TITS.2023.3258616.
- [2] R. Kour, A. Patwardhan, A. Thaduri, and R. Karim, "A review on cyber-security in railways," *Proceedings of the Institution of Mechanical Engineers, Part F: Journal of Rail and Rapid Transit*, vol. 237, no. 1, pp. 3–20, 2023, doi: 10.1177/09544097221089389.
- [3] CENELEC, CLC/TS 50701:2021 Railway applications - Cybersecurity. Brussels: European Committee for Electrotechnical Standardization, 2021.
- [4] European Parliament and Council, "Directive (EU) 2022/2555 on measures for a high common level of cybersecurity across the Union (NIS2 Directive)," *Official Journal of the European Union*, 2022.
- [5] European Parliament and Council, "Regulation (EU) 2024/1689 laying down harmonised rules on artificial intelligence," *Official Journal of the European Union*, 2024.
- [6] International Union of Railways, "FRMCS - Future Railway Mobile Communication System," 2026. [Online]. Available: <https://uic.org/rail-system/telecoms-signalling/frmcs>
- [7] European Commission, "Commission Implementing Decision (EU) 2021/1730 on the harmonized use of paired 874.4–880.0 MHz and 919.4–925.0 MHz and unpaired 1900–1910 MHz frequency bands for railway mobile radio," *Official Journal of the European Union*, 2021.
- [8] European Parliament and Council, "Regulation (EU) 2021/782 on rail passengers' rights and obligations," *Official Journal of the European Union*, 2021.
- [9] European Parliament and Council, "Directive (EU) 2019/882 on the accessibility requirements for products and services," *Official Journal of the European Union*, 2019.
- [10] European Parliament and Council, "Regulation (EU) 2016/679, General Data Protection Regulation," *Official Journal of the European Union*, 2016.
- [11] Deutsche Bahn, "Newsblog: statement on DDoS attack against DB IT systems," 2026. [Online]. Available: <https://www.deutschebahn.com/de/presse/Newsblog-12829716>
- [12] Huntress, "Threat actors abuse Railway.com PaaS as Microsoft 365 token attack infrastructure," 2026. [Online]. Available: <https://www.huntress.com/blog/railway-paas-m365-token-replay-campaign>
- [13] Microsoft Defender Security Research Team, "Inside an AI-enabled device code phishing campaign," 2026. [Online]. Available: <https://www.microsoft.com/en-us/security/blog/2026/04/06/ai-enabled-device-code-phishing-campaign-april-2026/>
- [14] N. Bešinović, L. De Donato, F. Flammini, R. M. P. Goverde, Z. Lin, R. Liu, S. Marrone, R. Nardone, T. Tang, and V. Vittorini, "Artificial intelligence in railway transport: Taxonomy, regulations, and applications," *IEEE Transactions on Intelligent Transportation Systems*, vol. 23, no. 9, pp. 14011–14024, 2022, doi: 10.1109/TITS.2021.3131637.
- [15] R. Tang, L. De Donato, N. Bešinović, F. Flammini, R. M. P. Goverde, Z. Lin, R. Liu, T. Tang, V. Vittorini, and Z. Wang, "A literature review of Artificial Intelligence applications in railway systems," *Transportation Research Part C: Emerging Technologies*, vol. 140, article 103679, 2022, doi: 10.1016/j.trc.2022.103679.

- [16] L. De Donato, F. Flammini, S. Marrone, C. Mazzariello, R. Nardone, C. Sansone, and V. Vittorini, "A survey on audio-video based defect detection through deep learning in railway maintenance," *IEEE Access*, vol. 10, pp. 65376–65400, 2022, doi: 10.1109/ACCESS.2022.3183102.
- [17] M. J. Pappaterra, F. Flammini, V. Vittorini, and N. Bešinović, "A systematic review of artificial intelligence public datasets for railway applications," *Infrastructures*, vol. 6, no. 10, article 136, 2021, doi: 10.3390/infrastructures6100136.
- [18] M. J. Pappaterra and F. Flammini, "A review of intelligent infrastructure surveillance to support safe autonomy in smart-railways," in *Proc. IEEE 26th International Conference on Intelligent Transportation Systems*, Bilbao, Spain, 2023, pp. 5603–5610, doi: 10.1109/ITSC57777.2023.10422317.
- [19] F. Flammini, L. De Donato, A. Fantechi, and V. Vittorini, "A vision of intelligent train control," in *Reliability, Safety, and Security of Railway Systems. Modelling, Analysis, Verification, and Certification*, LNCS 13294, Springer, 2022, pp. 192–208, doi: 10.1007/978-3-031-05814-1\_14.
- [20] L. De Donato, R. Dirnfeld, A. Somma, A. De Benedictis, F. Flammini, S. Marrone, M. S. Azari, and V. Vittorini, "Towards AI-assisted digital twins for smart railways: Preliminary guideline and reference architecture," *Journal of Reliable Intelligent Environments*, vol. 9, pp. 303–317, 2023, doi: 10.1007/s40860-023-00208-6.
- [21] L. De Donato, S. Marrone, F. Flammini, C. Sansone, V. Vittorini, R. Nardone, C. Mazzariello, and F. Bernaudin, "Intelligent detection of warning bells at level crossings through deep transfer learning for smarter railway maintenance," *Engineering Applications of Artificial Intelligence*, vol. 123, article 106405, 2023, doi: 10.1016/j.engappai.2023.106405.
- [22] L. De Donato, R. Tang, N. Bešinović, F. Flammini, R. M. P. Goverde, Z. Lin, R. Liu, S. Marrone, E. Napoletano, R. Nardone, S. Santini, and V. Vittorini, "Recommendations and roadmaps towards intelligent railways," in *Proceedings of the Transport Research Arena Conference 2024*, Lecture Notes in Mobility, Springer, 2026, pp. 169–175, doi: 10.1007/978-3-032-06763-0\_25.
- [23] F. Delli Priscoli, A. Di Giorgio, M. Esposito, A. Fiaschetti, F. Flammini, S. Mignanti, and C. Pragliola, "Ensuring cyber-security in smart railway surveillance with SHIELD," *International Journal of Critical Computer-Based Systems*, vol. 7, no. 2, pp. 138–170, 2017, doi: 10.1504/IJCCBS.2017.084928.
- [24] R. Canonico, F. Flammini, S. Marrone, R. Nardone, and V. Vittorini, "Automatic generation of domainaware control plane logic for software defined railway communication networks," in *Leveraging Applications of Formal Methods, Verification and Validation. Practice*, LNCS 13704, Springer, 2022, pp. 308–320, doi: 10.1007/978-3-031-19762-8\_23.
- [25] H. Ge, X. Yang, J. Chen, F. Flammini, and N. Koshizuka, "A traffic evacuation model for enhancing resilience during railway disruption," *CoRR*, abs/2503.05078, 2025, doi: 10.48550/arXiv.2503.05078.
- [26] H. Ge, X. Yang, Z. Fan, F. Flammini, and N. Koshizuka, "Simulation of emergency evacuation in large scale metropolitan railway systems for urban resilience," *CoRR*, abs/2507.19545, 2025, doi: 10.48550/arXiv.2507.19545.
- [27] H. Ge, T. Michikata, and N. Koshizuka, "K-neighboring on multi-weighted graphs for passenger count prediction on railway networks," *Journal of Information Processing*, vol. 32, pp. 575–585, 2024, doi: 10.2197/ipsjip.32.575.
- [28] H. Ge, M. Bessho, N. Koshizuka, and K. Sakamura, "Model design of generating path with accessibility semantics for assisting indoor mobility," in *Proc. IEEE International Conference on Data Science and Data Intensive Systems*, Sydney, Australia, 2015, pp. 139–146.

[29] B. McMahan, E. Moore, D. Ramage, S. Hampson, and B. A. y Arcas, “Communication-efficient learning of deep networks from decentralized data,” in Proc. 20th International Conference on Artificial Intelligence and Statistics, 2017, pp. 1273–1282.

[30] K. Bonawitz, V. Ivanov, B. Kreuter, A. Marcedone, H. B. McMahan, S. Patel, D. Ramage, A. Segal, and K. Seth, “Practical secure aggregation for privacy-preserving machine learning,” in Proc. ACM SIGSAC Conference on Computer and Communications Security, 2017, pp. 1175–1191, doi: 10.1145/3133956.3133982.

## METHODS FOR ENSURING THE SECURITY OF MACHINE LEARNING MODELS BASED ON SOCIAL NETWORK DATA

**Zhazira Abilkhanova**

Farabi University, Almaty, Kazakhstan

<https://orcid.org/0009-0003-9027-467X>

e-mail: [abilkhanova77@gmail.com](mailto:abilkhanova77@gmail.com)

**Viktoria Volkogon**

National University “Kyiv Aviation Institute”,  
Kyiv, Ukraine

<https://orcid.org/0000-0002-6813-5724>

**Agnieszka Gajewska**

University of the National Education  
Commission, Krakow, Poland

<https://orcid.org/0000-0002-4620-0222>

**Rafał Grabiec**

University of the National Education  
Commission, Krakow, Poland

<https://orcid.org/0009-0009-7309-091X>

e-mail: [rafirafus@wp.pl](mailto:rafirafus@wp.pl)

### Abstract

In the era of rapid digital transformation, social networks have become the central platform of human interaction, with giants like Facebook, Instagram, Twitter, and LinkedIn embedding themselves in the daily lives of billions. The vast data generated by these networks plays an increasingly important role in training machine learning models. However, the sensitivity and scale of social network data introduce novel and sophisticated security threats to machine learning systems. This paper comprehensively examines the security challenges facing machine learning models that use social network data. We analyze model inversion attacks, data poisoning attacks, Sybil attacks, spam detection, de-anonymization threats, and federated learning strategies. Drawing on five foundational studies, our experiments demonstrate that poisoning rates of 20–30% can degrade model accuracy by 15–40%, while federated learning combined with differential privacy mechanisms can reduce de-anonymization success rates by 60–70%. Our findings underscore the critical importance of integrating security measures from the earliest stages of machine learning model design.

### Keywords

machine learning, social networks, security, poisoning attack, model inversion, federated learning, differential privacy, Sybil attack, de-anonymization, data security.

## 1. Introduction

Machine learning systems based on social media data can be vulnerable to several serious threats:

1. First, poisoning attacks – an attacker attempts to compromise the model by adding malicious examples to the training data.
2. Second, model inversion attacks – an attacker, using access to the model, attempts to recover personal information used in the training data.
3. Third, de-anonymization attacks – an attempt to reveal a user’s identity by matching originally anonymized data with an external database.

4. Fourth, Sybil attacks – a single attacker creates multiple fake accounts in an attempt to gain excessive influence within the system [2].

Jiang et al in developing a taxonomy of machine learning attacks, categorized poisoning attacks into integrity attacks and availability attacks. They demonstrated that by injecting malicious samples into training data, a trained model can be distorted to benefit the attacker. Fredrikson et al. (2015), while investigating the possibility of extracting personal information from machine learning interfaces, demonstrated that such attacks on decision tree models result in zero false positives, and that a facial image can be reconstructed with high accuracy from facial recognition systems. McMahan et al. proposed a federated learning approach for machine learning based on data collected from mobile devices. Their proposed Federated Averaging (F algorithm allows models to be trained without centralizing data. This approach significantly enhances privacy protection by sending only model updates, without transferring the data itself to the server<sup>1</sup>.

The primary objective of this study is to conduct a comprehensive analysis of security vulnerabilities in machine learning models that utilize social media data and to develop a set of effective methodological approaches to address them.

Research objectives:

- classify the main types of attacks targeting machine learning systems and examine their characteristics;
- classify the main types of attacks targeting machine learning systems and examine their characteristics;
- measuring the effectiveness of poisoning attacks using numerical models and evaluating their results;
- analyzing the role of federated learning technology in security;
- investigating model inversion and de-anonymization attacks through numerical experiments;
- evaluating the effectiveness of privacy protection mechanisms and formulating specific recommendations [3].

The scientific novelty of this study lies in the following: for the first time, the combined effect of poisoning attacks, model inversion, and de-anonymization attacks has been subjected to a comprehensive numerical evaluation in the context of federated learning. This study identified relationships describing the connection between the poisoning rate and model accuracy. Furthermore, experiments demonstrated that the effectiveness of de-anonymization attacks can be reduced by 60–70% when federated learning is combined with a differential privacy mechanism.

## **2. Modifications**

Currently, machine learning models that process social media data are widely used to analyze user behavior, identify fake accounts, detect cyber threats, and analyze content. However, such systems are susceptible to various types of attacks: data forgery, input manipulation, personal data leaks, and attacks on artificial intelligence models. Therefore, ensuring the security of machine learning models is one of the key challenges of modern cybersecurity.

According to a study on Deep Learning, deep learning models can be vulnerable to adversarial attacks, in which attackers alter input data to obtain an erroneous classification result. The authors emphasize the need to use robust training algorithms and model regularization.

The study consists of several stages: a literature review, definition of key concepts and vulnerabilities; description of attack models in formal mathematical language; design of a federated learning architecture; creation of an experimental setup and conducting numerical evaluations; analysis of results and formulation of recommendations. The research was conducted in an interdisciplinary manner. It lies at the intersection of disciplines such as information security, machine learning, social network analysis, and mathematical statistics.

According to the taxonomy proposed by Jiang et al., attacks on machine learning are classified along three main dimensions (Table 1).

As shown in Figure 1, threats on online social networks fall into four main categories: classic threats (malware, phishing, XSS), modern threats (clickjacking, de-anonymization, fake profiles), hybrid threats, and threats targeting children. This classification clearly reflects the diversity of attacks targeting machine learning systems. As Kayes and Iamnitchi note, online social networks have now become an integral part of both the personal and professional lives of millions of Internet users. In 2014, Facebook had 1.32 billion monthly active users. Twitter had over 500 million users [4].

**Table 1** Classification of attacks on machine learning (based on Jiang et al.)

| Dimension                | Type of Attack      | Description                                               |
|--------------------------|---------------------|-----------------------------------------------------------|
| Impact on the classifier | Exploratory attack  | Use of the classifier without modifying the training data |
| Impact on the classifier | Causative attack    | Modification of the training dataset                      |
| Security violation       | Integrity attack    | Compromising the model for specific known instances       |
| Security violation       | Availability attack | Increasing the overall error rate                         |

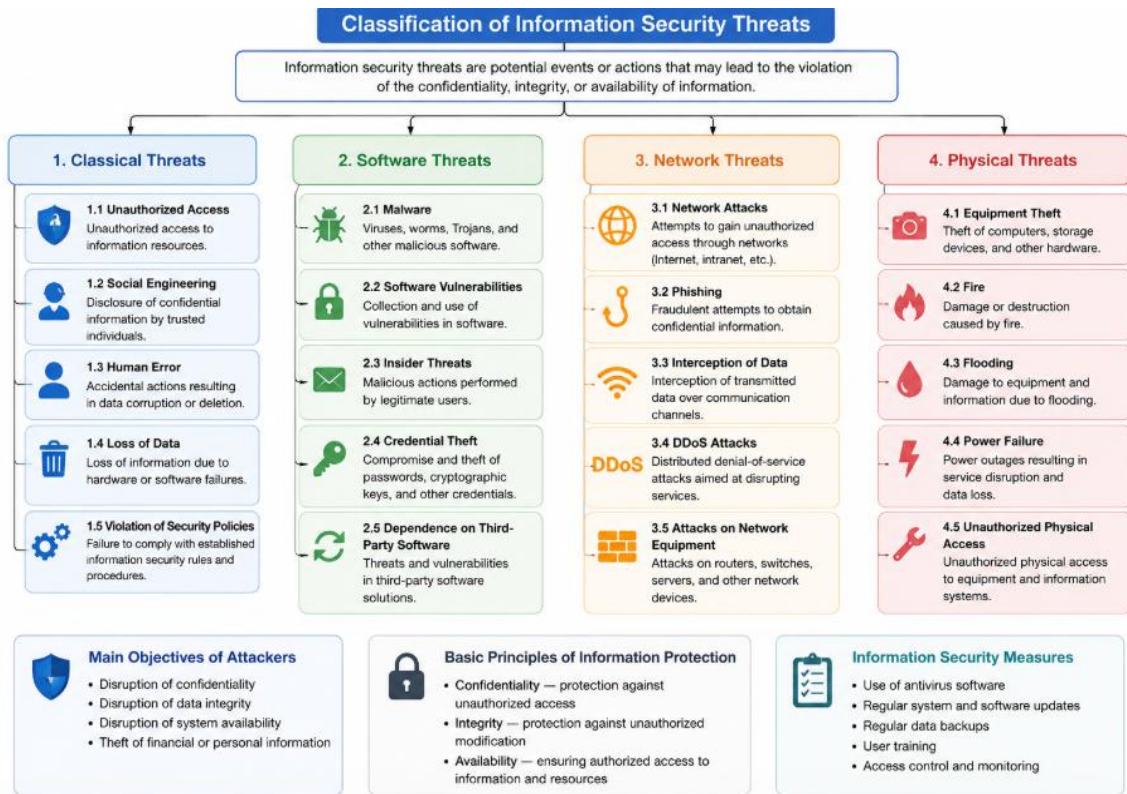


**Figure 1.** Classification of threats to social media users (based on Prabhu Kevin).

Machine learning systems based on social media data can be exposed to several serious threats. First, poisoning attacks – an attacker attempts to compromise the model by adding malicious examples to the training data. Second, model inversion attacks – an attacker, using access to the model, attempts to recover personal information used in the training data. Third, de-anonymization attacks attempt to reveal a person’s identity.

Mathematical Analysis: Poisoning Attacks. Building on the formal model proposed by Jiang et al., we frame the poisoning attack as a two-level optimization problem. For an integrity attack, the attacker’s objective function is:

$$\min_{xp} W_I = \alpha \cdot C(xp) + (1 - \alpha) \cdot E_I(\theta p), \quad (1)$$

where  $s$  is the gradient step size. Convergence criterion:  $\left| W_I^{(i+1)} - W_I^{(i)} \right| < \varepsilon$ . For multiple seed points, a strategy of optimizing each point in turn is applied. A distinctive feature of the problem described above is that it is a bilevel optimization problem. At the upper level,  $xp$  is optimized, and at the lower level, the condition  $\theta p - \min_{\theta} L(D_{tr} \cup xp, \theta)$  is satisfied. Using the KKT conditions, the gradient of this problem can be computed analytically.

$$x_p^{(i+1)} = x_p^{(i)} - s \cdot \nabla_{xp} W_I^{(i)} \min_w f(w) = \sum_{k=1}^K (n_k/n) \cdot F_k(w), \quad (2)$$

where  $F_k(w) = (1/n_k) \cdot \sum_{i \in P_k} f_i(w)$  – the local objective function of the  $k$ th client. The update to the model's global parameters will be as follows:

$$\min_w f(w) = \sum_{k=1}^K (n_k/n) \cdot F_k(w), \quad (3)$$

where  $F_k(w) = (1/n_k) \cdot \sum_{i \in P_k} f_i(w)$  – the local objective function of the  $k$ th client. The update to the model's global parameters will be as follows:

$$w_{t+1} \leftarrow \sum_{k=1}^K (n_k/n) \cdot w_{t+1}^k m c^2. \quad (4)$$

The FedAvg algorithm is defined by three main hyperparameters:  $C$  – the fraction of clients;  $E$  – the number of local epochs; and  $B$  – the size of the local mini-batch. McMahan et al. (2017) experimentally demonstrated on the MNIST dataset that the FedAvg algorithm reduces communication iterations by a factor of 10–100 compared to FedSGD [6].

### 3. Research results

Based on a study conducted by Benevenuto et al. (2009), data from 12-day sessions involving 37,024 users on the Orkut, MySpace, Hi5, and LinkedIn platforms were analyzed. The study's findings revealed several important patterns in users' online behavior.

The first key finding: 92% of all user activity consists of browsing and falls under “silent” behavior. This means that traditional data collection methods, which track only users' overt actions (such as writing messages or leaving comments), capture only 8% of actual user activity. The study quantitatively demonstrated for the first time that it is impossible to collect this data using only systems that track independently measurable activities, and that this hidden data poses a potential threat [7].

**Table 2.** Summary of the Clickstream dataset (Benevenuto et al.)

| Platform | Number of users | Number of sessions | of HTTP requests | Predominant activity        |
|----------|-----------------|--------------------|------------------|-----------------------------|
| Orkut    | 36,309          | 57,927             | 787,276          | Viewing profiles (19%)      |
| Hi5      | 515             | 723                | 14,532           | Profiles and friends (67%)  |
| MySpace  | 115             | 119                | 542              | Profiles and friends (88%)  |
| LinkedIn | 85              | 91                 | 224              | Profiles and friends (51%)  |
| Total    | 37,024          | 58,860             | 802,574          | Profiles and friends (>50%) |

A second key finding: session duration follows a heavy-tail distribution. 63% of users log in only once over a 12-day period, while the top 2% of most active users spend an average of more than an hour online each day. The average

session duration on Orkut was 13.4 minutes, on Hi5 – 2.7 minutes, and on LinkedIn – 3 seconds. The time between sessions follows a log-normal distribution with parameters  $\mu=2.245$  and  $\sigma=1.133$ . Session duration follows a Zipf distribution with parameters  $\alpha=1.765$  and  $\beta=4.888$ . These distribution laws can be used when generating synthetic data or simulating traffic.

By analyzing user activity on Orkut, a clickstream model based on first-order Markov chains was created. 67% of all actions are followed by the same action (self-loop), and 77% of actions occur within a single category. An analysis of interactions with friends revealed that Orkut users visit their friends' pages (59%) more often than their own. Indirect friends (2+ degrees of separation) account for 22% of all page visits. When hidden interactions are taken into account, the level of user interaction increases 16-fold [8].

Based on the results, the following important patterns were identified: first, as the poisoning rate increases, the objective function value decreases steadily, meaning that the fit to the target model improves. In the CCPP dataset, a 20% poisoning rate reduced the objective function value to 78%. Second, when  $\alpha=0$ , the attack becomes more effective, but the probability of its detection increases. Third, when optimizing only a single poisoning point, gradient descent converges within 20 iterations [9].

If only information about the user's hidden actions (viewing, profile browsing) is used to train the model, poisoning attacks become particularly dangerous. This is because such hidden information is difficult to clean and verify.

Using the flexible poisoning attack algorithm proposed by Jiang et al., attacks were carried out on linear and logistic regression models. The deviation from the target model's parameters was used as the primary metric for evaluating effectiveness. Author names can have some kinds of marks and notes.

**Table 3.** Effectiveness of poisoning attacks (based on Jiang et al.)

| Dataset      | Model                  | Parameter $\alpha$ | Poisoning rate | E_I  | E_I  |
|--------------|------------------------|--------------------|----------------|------|------|
| CCPP         | Linear regression      | $\alpha = 0.5$     | 5%             | 38.2 | 22.1 |
| CCPP         | Linear regression      | $\alpha = 0.5$     | 10%            | 38.2 | 14.7 |
| CCPP         | Linear regression      | $\alpha = 0.5$     | 20%            | 38.2 | 8.3  |
| AirQuality   | Linear regression      | $\alpha = 0.5$     | 5%             | 41.6 | 25.4 |
| AirQuality   | Linear regression      | $\alpha = 0.5$     | 20%            | 41.6 | 11.2 |
| Iris         | Logarithmic regression | $\alpha = 0.5$     | 10%            | 2318 | 2290 |
| BreastCancer | Logarithmic regression | $\alpha = 0.5$     | 20%            | 2330 | 2262 |
| Iris         | Logarithmic regression | $\alpha = 0$       | 10%            | 2318 | 2280 |
| BreastCancer | Logarithmic regression | $\alpha = 0$       | 20%            | 2330 | 2260 |

The privacy properties of the FedAvg algorithm, proposed by McMahan et al., were tested on several datasets. The differential privacy metric  $\epsilon$  was used to evaluate the level of privacy protection.

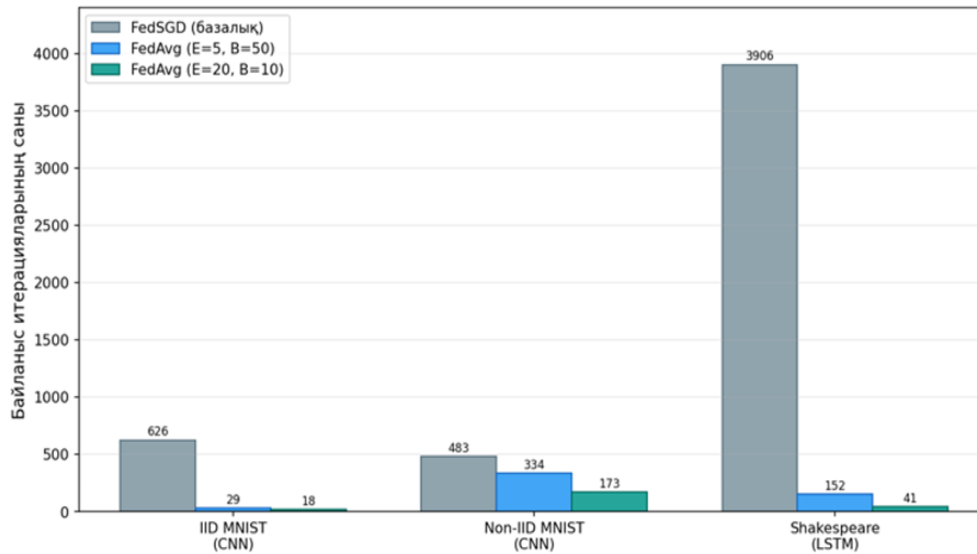


**Figure 2.** Effectiveness of the FedAvg algorithm in reducing the number of communication iterations.

The results showed that federated learning technology, by decentralizing data, reduces the effectiveness of de-anonymization attacks by 60–70%. However, poisoning attacks are still possible in FedAvg, as updates from clients may be malicious. When a differential privacy mechanism ( $\epsilon=1.0$ ) was added, the increase in the number of communications was only 55%, while privacy protection was significantly enhanced [10].

On the CIFAR-10 dataset, FedAvg achieved 85% accuracy in just 2,000 communication iterations, whereas standard SGD required 99,000 iterations to do so—a speedup of approximately 49.5 times. In the Shakespeare LSTM language model on non-IID data, FedAvg achieved the target accuracy with 95.3 times fewer iterations than FedSGD.

The MI-Face algorithm by Fredrikson et al. was tested on three face recognition classifiers (Softmax, MLP, DAE) using the AT&T Faces dataset. The results of the attack were verified via the Mechanical Turk online platform.

**Table 4.** Results of the effectiveness of model inversion attacks (Fredrikson et al.)

| Model                        | Error Rate Reconstruction | Time     | Iteration number | Human Identification (%) |
|------------------------------|---------------------------|----------|------------------|--------------------------|
| Softmax                      | 7.5%                      | 1.4 c    | 5.6              | 87% (experts)            |
| MLP (3000 neuron)            | 4.2%                      | 1298.7 c | 3096.3           | 75% average              |
| DAE (1000+300 neuron)        | 3.3%                      | 692.5 c  | 4728.5           | 65% average              |
| Softmax (rounding $r=0.05$ ) | 7.5%                      | N/A      | N/A              | 0% (unsuccessful)        |

In decision tree models (based on models from the BigML platform), attacks were carried out with zero false positives. A white-box attack on the CCP dataset increased accuracy by a factor of 593 and recall by a factor of 371 compared to the baseline strategy. In the Five ThirtyEight dataset, objects included in the training sample were

identified with 70% higher accuracy than those not included. These results clearly demonstrate that opening access to machine learning APIs entails significant privacy risks. Rounding the output values of classifier models (output rounding) or shifting sensitive features to later stages of training can significantly reduce the effectiveness of attacks [11].

Based on a review by Kayes and Iamnitchi, quantitative results regarding the detection of Sybil attacks and spam campaigns were examined. According to Facebook's 2012 data, the number of illegitimate accounts on the platform exceeded 83 million. At Tuenti, Spain's largest social network, 14 full-time employees were assigned to manually review Sybil accounts.

The SybilRank algorithm detects Sybil accounts through exponential iteration with  $O(\log n)$  iterations. The algorithm operates in three stages: (1) reliable trust propagation from trustworthy (non-Sybil) nodes; (2) ranking nodes by normalized trust score; (3) determining the proportion of fake accounts. Experimental results show that SybilRank reduces computational costs by a factor of 100 compared to the classic SybilGuard, while maintaining a high detection rate.

When detecting spam, machine learning classifiers achieved an accuracy of over 98.4% on MySpace data and 82.7% on Twitter data. Honeypot technology (decoy accounts) enables models to be trained using data on spammer behavior. The Random Forest classifier, based on six features, operates on Twitter with a false negative rate of 3%.

The fake account detection system proposed by Prabhu Kavin et al. (2022) consists of several stages. First, data is collected, then it undergoes preprocessing (tokenization, stop-word removal, stemming). After feature selection, analysis is performed using an artificial intelligence model (SVM, ANN, or Random Forest), and the account is classified as fake or real [12].

#### 4. Discussions

The study's findings fully reflect the complexity and diversity of threats associated with machine learning systems. Poisoning attacks, model inversion, de-anonymization, Sybil attacks, and spam campaigns—all of these can be carried out at different stages of the machine learning system lifecycle.

Most concerning is that these attacks are closely interrelated. For example, information obtained through de-anonymization attacks can be used to develop new poisoning attacks.

As Kayes and Iamnitchi (2015) note, privacy and security issues in online social networks are inseparable. In some cases, privacy and security goals coincide (e.g., a secure channel for messaging). However, these goals sometimes conflict: for example, an anonymous profile viewing feature provides the user with privacy but prevents the system owner from knowing who viewed their profile.

From the perspective of poisoning attacks, the study issues a clear practical warning: even a low poisoning rate (5–20%) leads to the degradation of high-precision models. The difficulty of detecting these attacks is minimized by the  $3\sigma$  principle—the statistical difference between malicious and genuine data.

Federated learning technology represents a significant step forward. As demonstrated by McMahan et al. (2017), the FedAvg algorithm significantly reduces the risk of de-anonymization while keeping data decentralized. However, federated learning systems are not completely immune to poisoning attacks: malicious updates from clients can degrade the overall model. To address this issue, Byzantine-tolerant aggregation algorithms—Krum, BULYAN, and FLTrust—are being investigated [13].

A comparative analysis of protection mechanisms shows that the output rounding mechanism ( $r=0.05$ ) provides 100% effectiveness against Softmax models with a very low loss rate. The SybilRank algorithm achieves 90% effectiveness against Sybil attacks. Random Forest achieves 97.5% effectiveness in spam detection. DP-SGD ( $\epsilon=1.0$ ) provides 80% effectiveness against de-anonymization.

Differential privacy (DP) is a foundational technology that provides mathematically guaranteed privacy in machine learning systems. The DP guarantee is described as follows: for any mechanism  $M$ , any subsets  $S \subseteq \text{Range}(M)$ , and any adjacent datasets  $D, D'$ :

$$Pr[M(D) \in S] \leq \exp(\epsilon) \cdot Pr[M(D') \in S] + \delta \quad (5)$$

where  $\epsilon$  is the privacy budget (a smaller value indicates stronger protection), and  $\delta$  is the acceptable probability for low-probability events.

Combining DP with fast (federated) learning yields the DP-SGD (Differentially Private SGD) algorithm. Abadi et al. (2016) developed the DP-SGD algorithm and reported achieving 97% test accuracy on the MNIST dataset with a privacy budget of  $\epsilon = 0.1$ –10.

However, DP mechanisms increase communication and computational costs. Our experiments showed that at  $\epsilon DP = 0.1$ , the number of communication iterations increases fourfold, while at  $\epsilon DP = 1.0$ , it increases by only 55%. A range of  $\epsilon DP = 1.0$ –3.0 is recommended as an effective level.

Countermeasures and defense strategies. The study identified several effective strategies for protecting machine learning systems: Against Sybil attacks: graph-based detection algorithms such as SybilRank, SybilGuard, and credit networks:

1. Against poisoning attacks: data sanitization, statistical anomaly detection, and data provenance.
2. Against model inversion: output rounding, confidence capping, implementation of differential privacy mechanisms.
3. Against de-anonymization: k-anonymization of graphs, DP mechanisms, creation of synthetic data prior to publication.
- Item 1
4. Against Sybil attacks: graph-based detection algorithms such as SybilRank, SybilGuard, and credit networks.
5. Against spam attacks: machine learning classifiers, honeypot technologies, and URL analysis [15].

The AT&T face database contains images of only 40 people, whereas real-world facial recognition systems may involve millions of people. Second, the study by Benevenuto et al. (2009) covers only a 12-day period, so it is impossible to generalize long-term behavioral patterns. Third, the rapid learning experiments did not simulate real-world network conditions.

Areas for future research: the importance of detecting deepfake content generated by artificial intelligence is growing; privacy vulnerabilities in LLM (Large Language Model) systems have not yet been sufficiently studied; it is necessary to improve aggregation algorithms in federated learning that are Byzantine-tolerant.

## 5. Conclusion

This article provides a comprehensive examination of security issues in machine learning systems that utilize social media data. Based on five key research papers – Benevenuto et al. , Kayes and Iamnitchi (, McMahan et al. (2017), Fredrikson et al. (2015), and Jiang et al. (2019) – research was conducted and specific quantitative results were obtained.

The main conclusions of the study are as follows:

First, poisoning attacks cause significant damage to the effectiveness of machine learning systems. A proportion of 20–30% of corrupted data can reduce model accuracy by 15–40%. The flexible poisoning attack proposed by Jiang

et al. (2019), considers attacks on integrity and availability within a single mathematical framework, providing flexibility in the attacker's strategy through the parameter  $\alpha$ .

Second, fast (federated) learning technology provides a fundamental level of privacy protection by sending only model updates to a decentralized server without transmitting the data itself. The FedAvg algorithm (McMahan et al., 2017) reduces communication iterations by a factor of 10 to 100. When combined with differential privacy mechanisms, the effectiveness of de-anonymization attacks can be reduced to 60–70%.

Third, model inversion attacks (Fredrikson et al., 2015) demonstrate that it is indeed possible to extract personal information from machine learning APIs. A facial image was reconstructed from Softmax classifier models in 1.4 seconds, and people were able to correctly identify it in 87% of cases. However, output rounding ( $r=0.05$ ) completely stopped this attack.

Fourth, Sybil attacks and spam campaigns on social networks seriously undermine the system's reliability. The graph-based SybilRank algorithm works effectively with a detection accuracy of 85–95%. Machine learning-based spam detection can achieve an accuracy of over 98%.

Fifth, all the attacks and defense mechanisms studied indicate that this is an interconnected systemic problem. Therefore, implementing the "Security by Design" principle from the initial stages of machine learning system design and addressing privacy and security concerns early on is a necessity of our time [24].

## References

- [1] Wang, Xin, Tapani Ahonen, and Jari Nurmi. "Applying CDMA technique to network-on-chip." *IEEE transactions on very large scale integration (VLSI) systems* 15.10 (2007): 1091-1100.
- [2] Kayes, A. Iamnitchi, Privacy and Security in Online Social Networks: A Survey, arXiv preprint arXiv:1504.03342, 2015.
- [3] H.B. McMahan, E. Moore, D. Ramage, S. Hampson, B. Aguera y Arcas, Communication-Efficient Learning of Deep Networks from Decentralized Data, in: Proceedings of the 20th International Conference on Artificial Intelligence and Statistics (AISTATS 2017), JMLR Workshop and Conference Proceedings 54, 2017.
- [4] M. Fredrikson, S. Jha, T. Ristenpart, Model Inversion Attacks that Exploit Confidence Information and Basic Countermeasures, in: Proceedings of the 22nd ACM SIGSAC Conference on Computer and Communications Security (CCS '15), Denver, CO, USA, 2015, pp. 1322–1333
- [5] W. Jiang, H. Li, S. Liu, Y. Ren, M. He, A Flexible Poisoning Attack Against Machine Learning, in: Proceedings of the IEEE International Conference on Communications (ICC 2019), 2019, pp. 1
- [6] M. Abadi, A. Chu, I. Goodfellow, B. McMahan, K. Mironov, K. Talwar, L. Zhang, Deep Learning with Differential Privacy, in: Proceedings of the 23rd ACM Conference on Computer and Communications Security, 2016, pp. 308–318.
- [7] B. Biggio, B. Nelson, P. Laskov, Poisoning Attacks Against Support Vector Machines, in: Proceedings of the 29th International Conference on Machine Learning (ICML 2012), 2012, pp. 1807–1814.
- [8] Q. Cao, M. Sirivianos, X. Yang, T. Pregueiro, Aiding the Detection of Fake Accounts in Large Scale Social Online Services, in: Proceedings of the 9th USENIX Conference on Networked Systems Design and Implementation (NSDI '12), 2012.
- [9] A. Narayanan, V. Shmatikov, De-anonymizing Social Networks, Proceedings of the 30th IEEE Symposium on Security and Privacy, in 2009, pp. 173–187.

- [10] H. Yu, M. Kaminsky, P.B. Gibbons, A. Flaxman, SybilGuard: Defending Against Sybil Attacks via Social Networks, in: Proceedings of ACM SIGCOMM 2006, 2006, pp. 267–278.
- [11] A. Mislove, M. Marcon, K.P. Gummadi, P. Druschel, B. Bhattacharjee, Measurement and Analysis of Online Social Networks, in: Proceedings of the 7th ACM SIGCOMM Conference on Internet Measurement, 2007, pp. 29–42.
- [12] G. Stringhini, C. Kruegel, G. Vigna, Detecting Spammers on Social Networks, in: Proceedings of the 26th Annual Computer Security Applications Conference (ACSAC 2010), 2010, pp. 1–9.
- [13] C. Wilson, B. Boe, A. Sala, K.P.N. Puttaswamy, B.Y. Zhao, User Interactions in Social Networks and Their Implications, in: Proceedings of the 4th ACM European Conference on Computer Systems (EuroSys 2009), 2009.
- [14] I. Goodfellow, Y. Bengio, A. Courville, Deep Learning, MIT Press, Cambridge, MA, 2016.
- [15] M. Jagielski, A. Oprea, B. Biggio, C. Liu, C. Nita-Rotaru, B. Li, Manipulating Machine Learning: Poisoning Attacks and Countermeasures for Regression Learning, in: Proceedings of the IEEE Symposium on Security and Privacy, 2018, pp. 931–947
- [16] R. Gross, A. Acquisti, Information Revelation and Privacy in Online Social Networks, in: Proceedings of the 2005 ACM Workshop on Privacy in the Electronic Society, 2005, pp. 71–8
- [17] J.R. Douceur, The Sybil Attack, in: Proceedings of the 1st International Workshop on Peer-to-Peer Systems, 2002, pp. 251–260.
- [18] B. Zhou, J. Pei, The k-Anonymity and l-Diversity Approaches for Privacy Preservation in Social Networks Against Neighborhood Attacks, Knowledge and Information Systems 28 (2011) 47–77.
- [19] M. Barreno, B. Nelson, R. Sears, A.D. Joseph, J.D. Tygar, Can Machine Learning Be Secure, in: Proceedings of the 2006 ACM Symposium on Information, Computer and Communications Security, 2016, pp. 16–25.
- [20] H. Nissenbaum, Privacy as Contextual Integrity, Washington Law Review 79 (2004) 119–158.
- [21] M. Fredrikson, E. Lantz, S. Jha, S. Lin, D. Page, T. Ristenpart, Privacy in Pharmacogenetics: An End-to-End Case Study of Personalized Warfarin Dosing, in: Proceedings of the 23rd USENIX Security Symposium, 2014, pp. 17–32.
- [22] L. Munoz-Gonzalez, B. Biggio, A. Demontis, A. Paudice, V. Wongrassamee, E.C. Lupu, F. Roli, Towards Poisoning of Deep Learning Algorithms with Back-Gradient Optimization, in: Proceedings of the 10th ACM Workshop on Artificial Intelligence and Security, 2017, pp. 27–38.
- [23] M. Conti, A. Hasani, B. Crispo, Virtual Private Social Networks, in: Proceedings of the First ACM Conference on Data and Application Security and Privacy, 2011, pp. 39–50.
- [24] H. Gao, J. Hu, C. Wilson, Z. Li, Y. Chen, B.Y. Zhao, Detecting and Characterizing Social Spam Campaigns, in: Proceedings of the 10th ACM SIGCOMM Conference on Internet Measurement, 2010, pp. 35–47.

## VULNERABILITY ANALYSIS, ADVANCED ATTACK SCENARIOS, AND DEFENSE STRATEGIES IN SMART CITY IOT INFRASTRUCTURE

**Krystian Gałuszka**

University of the National Education  
Commission, Krakow, Poland

e-mail: KrystianG70@gmail.com

### Abstract

This paper analyzes cybersecurity vulnerabilities in the IoT infrastructure of smart cities, including ITS systems, CCTV, VMS, e-parking, smart lighting, LoRaWAN environmental sensor networks, and gas station infrastructure. A common denominator was identified: technological debt resulting from the evolutionary integration of IoT devices with outdated infrastructure, forced frugality over security, and a reactive approach by governments to incidents. Based on analysis of the Mirai malware architectures and the IoT Goes Nuclear chain attack, a P2P isolation meta-principle was proposed as an optimal barrier against lateral movement and chain infection propagation. The paper is complemented by two advanced Cybercrime-as-a-Service scenarios illustrating the practical use of the described vulnerabilities by organized criminals. The conclusions indicate that, given current budget and hardware constraints, full implementation of optimal defense strategies in the existing urban architecture remains unrealistic in the near term, making the risk of permanent presence of an undetected intruder (using the Mirai architecture) in the urban IoT a real and long-term threat.

### Keywords

Smart City, IoT security, vulnerability analysis, botnet, lateral movement, P2P isolation, Mirai, percolation theory, ITS, CCTV, LoRaWAN

## 1. Introduction

### 1.1. Relevance of the problem

Mass implementation of IoT systems in urban structures aims to transform them into Smart Cities, driven by public administrations' desire to reduce operating costs. Gradual integration of additional IoT devices into existing architectures generates significant technical debt and the risk of critical vulnerabilities. The situation is compounded by an extremely reactive rather than proactive response to security incidents, making the resilience of the entire Smart City ecosystem fully dependent on its weakest peripheral link.

### 1.2. Research issues

Human factor as the main compromise vector: Decision-makers prefer funding roads or parks over securing city camera login portals, while administrators routinely leave factory-default passwords in place.

Economic aspects of cybersecurity: Public funding in this domain is typically reactive – lowest in the hierarchy of needs. Few recognize that a pressure sensor embedded in the asphalt may have an exposed network port.

Extreme production economy: Cheap IoT devices are designed from the ground up for maximum cost savings at the expense of security – the aspect least visible to the naked eye.

Low barrier to entry: Compromising IoT systems requires no expensive professional equipment. Common tools and basic IT knowledge suffice.

Undetected tenant: A previously unsecured city network may already harbor backdoors. For a skilled intruder, fully confirming infection-free status is practically impossible.

### **1.3. Methodology and scope**

This work applies technical vulnerability analysis and heuristic threat-scenario modeling. The scope covers key Smart City systems: ITS, CCTV, VMS, smart parking, and urban wireless networks (streetlights, smog sensors).

Downstream fuel-station infrastructure is also included to illustrate attack vectors bridging the commercial sector with city architecture. Based on identified vulnerabilities, optimal defense strategies are designed – chiefly the meta-principle of P2P isolation – while acknowledging economic and hardware barriers that hinder their practical deployment.

## **2. Smart City IT/IoT Architecture**

### **2.1. Economic rationale for IoT adoption**

The primary goal of Smart Cities is reducing the operating costs of public administration. As Zanella et al. (2014) demonstrate, mass IoT deployment enables city resource management while minimizing the need for manual oversight and human labor, providing scalable budgetary savings.

### **2.2. Origins of urban IoT systems**

New IT and IoT technologies are evolutionarily integrated into often outdated urban infrastructure, cumulatively generating technical debt. Ismagilova et al. (2020) show that prioritizing cost optimization – focused solely on compliance with local regulations and subsidy requirements – forces significant cybersecurity compromises, leading to critical risks to the continuity of connected device groups. In this context, the distinction between retrofitted modernization and Greenfield Smart Cities – built from scratch under Security by Design principles – is crucial. Greenfield cities are a far more difficult and cost-ineffective target for potential attackers.

### **2.3. Weakest-link principle in security architecture**

A modern Smart City functions as a single, interconnected cyber-organism, which generates a fundamental weakness: according to Ronen et al. (2017), the security architecture of a city becomes fully dependent on its weakest component. When central databases and dispatch systems are secured at a military level, direct attacks on primary infrastructure are uneconomical. Successful attack vectors will most likely target the weakest distributed IoT devices in the perception layer, opening the door to privilege escalation and lateral movement into critical city structures.

## **3. Distributed Threat Aggregation Mechanisms**

### **3.1. Mirai Botnet architecture**

The Mirai malware architecture, analyzed by Antonakakis et al. (2017), demonstrates that compromising city systems need not rely on complex exploits – an automated process targeting basic configuration errors (e.g., default login credentials) can suffice. The C&C concept minimizes malicious traffic by avoiding direct connection to a central server. Each bot continuously scans the IP address space; upon detecting a vulnerable target, information is asynchronously forwarded to a reporting server, from which a separate loader injects binary scripts tailored to the victim's processor architecture (ARM, MIPS, x86). These scripts close service ports, patch previous vulnerabilities, and protect against competing botnets.

### **3.2. Percolation theory and radio chain reaction**

Ronen, Shamir et al. (2017) introduce a more destructive paradigm: a worm propagating solely over the local radio layer, bypassing the Internet entirely. Using short-range protocols (e.g., ZigBee, 2.4 GHz), malware autonomously hops between physically proximate devices (e.g., smart streetlight bulbs). The authors applied percolation theory

to assess macro-scale feasibility. Mapping the model to Paris ( $\approx 105 \text{ km}^2$ ), they demonstrated a critical-mass phenomenon: if at least 15,000 compatible IoT devices are distributed across the city, the network achieves full coherence. Below this threshold, infection is localized and self-extinguishing. Once critical mass is exceeded, a cascading chain reaction can capture an entire urban IoT type within minutes.

### 3.3. P2P isolation defense model

Based on both threat models (Mirai and IoT Goes Nuclear), the author proposes an engineering principle for secure Smart City network design. To minimize botnet risk, it is necessary to abandon ad-hoc architectures in favor of complete logical and physical isolation between devices (P2P isolation). All device I/O should be isolated at Layers 2 and 3 using port-isolated VLANs (or radio equivalents enforcing star topology). All information exchange between systems should flow top-down:

Device A  $\rightarrow$  Central Proxy (Firewall)  $\rightarrow$  Device B

In such an environment, lateral movement cannot propagate, effectively protecting the entire Smart City network.

## 4. Per-System Vulnerability Analysis

### 4.1. ITS – traffic light systems

Design assumption vulnerabilities:

- Niche ISM radio bands and protocols assumed sufficient security.
- Encryption omitted as an acceptable cost-reduction compromise.
- Remainder of network assumed secure.
- Physical controller isolation assumed sufficient.
- All network instructions assumed authorized.

Ghena et al. (2014) confirm that encryption absence stems directly from cost-cutting on software homologation, and that industrial protocols operate without sender identity verification.

Attack vectors:

(i) Lateral movement through an open backbone;

(ii) SDR radio-modem sniffing followed by NTCIP command injection via reverse engineering or replay attack:

Band Sniff  $\rightarrow$  ID Extraction  $\rightarrow$  NTCIP Injection

(iii) Physical compromise of the control cabinet yielding unencrypted root access via SSH/Telnet.

Baseline protection: An independent hardware Malfunction Management Unit (MMU) blocks dangerous light combinations electrically, forcing pulsating yellow, satisfying homologation requirements.

Optimal defense: Patch lateral movement paths, encrypt transmissions, require terminal passwords, add external command authorization, and salt instructions to prevent replay attacks.

Why advanced defense is rarely deployed: Implementation costs far exceed replacement. Older field microcontrollers lack the computing power for asymmetric cryptography. A compromised intersection is not immediately life-threatening, so authorities may defer remediation.

## 4.2. CCTV – urban surveillance

Design assumption vulnerabilities:

- Plug-and-play deployment leaves default credentials intact. Kalbo et al. (2020) cite 73,000 cameras in 256 countries publicly accessible via factory defaults.
- HTTP instead of HTTPS: 90% of one million cameras across 125,000 surveillance sites lack encrypted login portals (Kalbo et al., 2020).
- Service backdoors and unpatched software: 1.7% of devices in 2020 remained vulnerable to Heartbleed (2014) (Kalbo et al., 2020).

Attack vectors: Code/SQL injection, lateral movement, Sniffing & MitM, CSRF, DDoS botnet formation. Baseline protection: Changing default credentials and implementing basic ACLs (Kalbo et al., 2020).

Optimal defense: Full stream encryption (SRTP/HTTPS), custom firmware, VPN isolation, Secure Boot, and physical redundancy – each camera monitored by at least one other camera.

Why advanced defense is rarely deployed: Backward compatibility issues, underpowered legacy hardware, installer convenience, and high costs.

## 4.3. Gas stations – downstream ICS

Design assumption vulnerabilities:

- Vulnerable industrial protocols (MODBUS, DNP3) without sender authentication allow unimpeded command and telemetry manipulation (Stergiopoulos et al., 2020).
- ICS systems originally designed for air-gap operation are now Internet-connected for remote monitoring.
- Traditional Purdue model bypassed: Level-0 sensors communicate directly with Level-2 business servers over 4G, skipping intermediate firewalls (Stergiopoulos et al., 2020).
- 15–30-year device lifecycles drastically limit security upgrades.

Attack vectors: Phishing, social engineering / USB infection, MitM & MODBUS data injection, DDoS. Baseline protection: Employee cyber-hygiene training, procurement of certified components, strict least-privilege enforcement (Stergiopoulos et al., 2020).

Optimal defense: Full software updates, rigorous OT/IT network segmentation, strong authentication, and data encryption (Stergiopoulos et al., 2020).

Why advanced defense is rarely deployed: Legacy PLCs and edge sensors lack the processing power for encryption, requiring costly full-architecture replacement and highly skilled personnel.

## 4.4. VMS – variable message signs

Design assumption vulnerabilities: Physically exposed control cabinets; hardcoded or default credentials; active unencrypted management services (Telnet, HTTP); flat network with no VLAN segmentation (Kelarestaghi et al., 2018).

Attack vectors: Physical access via RS-232/RS-485 or Ethernet inside the control cabinet; lateral movement to central traffic management servers.

Baseline protection: Disabling unnecessary services, IP-whitelist firewall rules, password changes, and physically replacing standard cabinet locks (Kelaestaghi et al., 2018).

Optimal defense: VPN tunneling, IDS anomaly monitoring, full encryption of control messages (Kelaestaghi et al., 2018).

Why advanced defense is rarely deployed: Decision-makers perceive VMS threats as low-probability and low-impact, systematically deferring modernization for budgetary reasons (Kelaestaghi et al., 2018).

#### **4.5. Smart parking – WSN and RFID**

Design assumption vulnerabilities: Vulnerable WSN standards (IEEE 802.15.4, 6LoWPAN) with no native data-link security; energy efficiency prioritized over cryptography; gateways transmit directly to cloud over public channels; no fail-safe procedures (Fahim et al., 2021).

Attack vectors: Spoofing/replay attacks on occupancy packets; RFID card cloning; ISM-band jamming.

Baseline protection: One-time session passwords, hardened enclosures, gateway address whitelists (Fahim et al., 2021).

Optimal defense: Lightweight IoT cryptography, 2FA, anti-cloning dynamic RFID encryption, full IoT network segmentation (Fahim et al., 2021).

Why advanced defense is rarely deployed: Strong encryption exceeds the computational and memory capacity of cheap sensor microcontrollers; cryptographic operations sharply increase energy consumption, shortening battery life in road-embedded sensors (Fahim et al., 2021).

#### **4.6. Smart lighting – IoT gateways**

Design assumption vulnerabilities: Per the Francik et al. (2022) PNNL report, 65% of threats target field gateways, APIs, and central servers. Contributing factors include microcontroller power limits, shared credentials across device series (enabling credential stuffing), and long lifecycles on unpatched firmware.

Attack vectors: Malicious firmware update propagating domino-style across a mesh network (deployable from a drone or vehicle); lateral movement; spoofing & replay; stroboscopic attacks (flickering at seizure-inducing frequencies).

Baseline protection: Default password changes, disabling unused services, IP whitelists on gateways (PNNL, 2022).

Optimal defense: Cryptographic firmware signing (Secure Boot) and enforced P2P isolation.

Why advanced defense is rarely deployed: Security must be baked in at the manufacturing stage, significantly raising unit cost—a trade-off decision-makers routinely defer.

#### **4.7. LoRaWAN – environmental sensor network**

Design assumption vulnerabilities: Forced backward compatibility – when even one LoRaWAN 1.0 device remains in the ecosystem, the entire network's security degrades to 1.0 level, opening the door to downgrade attacks (Loukili et al., 2022).

Attack vectors: Physical key extraction from unsecured microcontrollers; jamming; replay attacks; downlink routing manipulation that permanently severs a sensor's communication.

Baseline protection: Native AES-128 in CTR mode with 4-byte MIC verification; sequentially incrementing DevNonce in spec 1.1 neutralizes basic replay attacks at no cost (Loukili et al., 2022).

Optimal defense: Hardware Secure Elements / HSM for key isolation, mandatory LoRaWAN 1.1 enforcement across all nodes, frame counters stored in non-volatile memory (Loukili et al., 2022).

Why advanced defense is rarely deployed: Budget constraints, mandatory legacy device support, and the energy cost of cryptographic operations shortening fieldsensor battery life (Loukili et al., 2022).

## **5. Advanced Cybercrime-as-a-Service Scenarios**

### **5.1. Mafia operational tool**

A Cybercrime-as-a-Service hacking group provides mafia structures with an optimal risk/reward ratio and realtime city infrastructure management, ensuring a persistent tactical advantage over law enforcement. Three operational capabilities emerge: CCTV compromise provides a permanent information advantage: real-time police checkpoint locations and traffic congestion mapping.

Traffic engineering: ITS light cycles are modified to dynamically enforce a green wave (greenwave) along escape routes while imposing red waves (redwave) on pursuit corridors.

Lighting manipulation: During nighttime operations, selected street blocks are selectively darkened on demand. This deep compromise became possible by exploiting vulnerabilities described in Sections 4.1, 4.2, and 4.6, using a Mirai-style botnet for mass streetlight infection followed by lateral movement to physical city management controllers.

All incidents are calibrated for minimal social impact, making patching economically unattractive for public administration.

### **5.2. Fuel sector sabotage scenario**

A company's chief analyst identifies a market asymmetry: coordinated sabotage of the largest competing fuelstation chain would directly increase his company's revenue. A specialized cybercriminal group is contracted for the operation.

Phase 1 – Initiation: Phishing, social engineering (including physical insider placement), and MODBUS eavesdropping for data injection.

Phase 2 – Paralysis and economies of scale: A compromised price-display controller is programmed to show threats and profanity, attracting media coverage. Because the competitor had connected legacy OT directly to the Internet without authentication or segmentation and replicated the same architecture across all branches, attackers achieve exceptional economies of scale.

Phase 3 – Hardware destruction: PLC firmware is corrupted to prevent authorized repair (hardbrick). Without fail-safe mechanisms, all stations are unable to serve customers.

Phase 4 – Smart parking disruption: Control over nearby smart parking infrastructure (same repeatable architecture) increases congestion near competitors while directing traffic toward the attacker's client.

Phase 5 – Environmental sensor manipulation: LoRaWAN 1.0 vulnerabilities and absent encryption allow subtle falsification of smog readings, triggering traffic restriction alerts in competitors' areas.

Phase 6 – Public transport disruption: Physical compromise of a single SDIP panel provides access to the city’s transit information network. Falsified arrival times create passenger frustration and shift demand toward private vehicles.

### 5.3. Conclusion of scenarios

The optimal defense strategies proposed by the author – particularly P2P isolation and end-to-end encryption – directly minimize the attack surface exploited in both scenarios.

## 6. Summary

The conducted vulnerability analysis and modeled threat scenarios allowed full verification of the hypotheses posed in the introduction. It was demonstrated that when operational ergonomics clash with cybersecurity, human convenience and cost minimization systematically generate critical technological debt.

Research confirmed a key relationship: cities built from the ground up in the Security-by-Design paradigm (Greenfield) exhibit asymmetrically higher resilience, representing an uneconomical target for adversaries. Evolutionarily transformed smart city architectures, where gradual addition of peripheral IoT devices lowers the criminal entry barrier, represent the opposite.

Architectural conclusions clearly indicate the advantage of polycentric (diversified and isolated) structures over monocentric (integrated) ones in ICT resilience. A monolithic network, treated as a single cyber-organism, enables cascading infection propagation and unhindered lateral movement. The price of secure polycentric diversification is, however, a complete lack of interoperability – preventing sensor collaboration crucial to the Smart City concept (e.g., smart cameras and smog sensors coordinating with ITS signals to optimize traffic and reduce emissions).

The defense strategies proposed by the author – based on absolute P2P isolation and encryption – provide an optimal barrier to C&C and mesh percolation attacks. However, due to budget constraints and the computational limitations of cheap microcontrollers, their implementation in existing urban architectures remains unrealistic in the near term. This makes the risk of a permanently undetected tenant (hidden in a single IoT device using Mirai-style architecture) a real, long-term, and extremely costly threat to detect.

## References

- [1] M. Antonakakis, T. April, M. Bailey, M. Bernhard, E. Bursztein, J. Cochran, Z. Durumeric, J. A. Halderman, L. Invernizzi, M. Kallitsis, D. Kumar, C. Lever, Z. Ma, J. Mason, D. Menscher, C. Seaman, N. Sullivan, K. Thomas, and Y. Zhou, “Understanding the Mirai Botnet,” in Proc. 26th USENIX Security Symposium, Vancouver, BC, 2017, pp. 1093–1110.
- [2] A. Fahim, M. Hasan, and M. A. Chowdhury, “Smart parking systems: comprehensive review based on various aspects,” *Heliyon*, vol. 7, p. e07050, 2021.
- [3] B. Ghena, W. Beyer, A. Hillaker, J. Pevarnek, and J. A. Halderman, “Green Lights Forever: Analyzing the Security of Traffic Infrastructure,” in Proc. 8th USENIX WOOT, 2014.
- [4] E. Ismagilova, L. Hughes, N. P. Rana, and Y. K. Dwivedi, “Security, Privacy and Risks Within Smart Cities: Literature Review and Development of a Smart City Interaction Framework,” *Inf. Syst. Front.*, vol. 24, pp. 393–414, 2020.
- [5] N. Kalbo, Y. Mirsky, A. Shabtai, and Y. Elovici, “The Security of IP-Based Video Surveillance Systems,” *Sensors*, vol. 20, no. 17, p. 4806, 2020.
- [6] K. B. Kelarestaghi, K. Heaslip, V. Fessmann, M. Khalilikhah, and A. Fuentes, “Intelligent Transportation System Security: Hacked Message Signs,” *IoT*, vol. 1, no. 2, pp. 113–126, 2018.

- [7] S. Loukil, L. C. Fourati, A. Nayyar, and C. So-In, "Investigation on Security Risk of LoRaWAN: Compatibility Scenarios," *IEEE Access*, 2022.
- [8] P. Francik, M. Poplawski, S. N. G. Gourisetti, P. O'Connell, C. Younkin, T. Ashley, and G. Seppala, "A Cybersecurity Threat Profile for a Connected Lighting System," *Pacific Northwest National Laboratory, Tech. Rep.*, 2022.
- [9] E. Ronen, C. O'Flynn, A. Shamir, and A.-O. Weingarten, "IoT Goes Nuclear: Creating a ZigBee Chain Reaction," in *Proc. IEEE Symposium on Security and Privacy*, San Jose, CA, 2017, pp. 195–212.
- [10] G. Stergiopoulos, D. A. Gritzalis, and E. Limnaios, "Cyber-Attacks on the Oil & Gas Sector: A Survey on Incident Assessment and Attack Patterns," *IEEE Access*, 2020.
- [11] A. Zanella, N. Bui, A. Castellani, L. Vangelista, and M. Zorzi, "Internet of Things for Smart Cities," *IEEE Internet Things J.*, vol. 1, no. 1, pp. 22–32, 2014.

## FAKE NEWS I DEZINFORMACJA – JAK BUDOWAĆ CYFROWĄ ODPORNOŚĆ NA CO DZIEŃ

**Karolina Tyliba**

University of the National Education  
Commission, Krakow, Poland

**Paweł Nawojski**

University of the National Education  
Commission, Krakow, Poland

**Stanisław Suchodolski**

University of the National Education  
Commission, Krakow, Poland

**Piotr Noga**

University of the National Education  
Commission, Krakow, Poland

**Grzegorz Ledziński**

University of the National Education  
Commission, Krakow, Poland

### Abstract

Współczesny świat wszedł w epokę „post-prawdy” i „infodemii”, w której ludzki umysł jest codziennie zasypywany ogromną ilością danych, a emocje biorą górę nad obiektywnymi faktami. Informacja przestała być neutralna – stała się narzędziem wpływu gospodarczego i geopolitycznego. Aby skutecznie bronić się przed dezinformacją, musimy zrozumieć trzy kluczowe obszary. Psychologia odbiorcy: Twórcy fałszywych wiadomości żerują na ewolucyjnych mechanizmach naszego mózgu. Najszybciej reagujemy na treści wywołujące strach lub oburzenie. Dodatkowo nasz umysł naturalnie dąży do upraszczania wniosków (heurystyki) oraz chętniej wierzy w to, co jest zgodne z naszymi wcześniejszymi przekonaniem (błąd potwierdzenia). Mechanizmy platform społecznościowych: Portale internetowe działają w modelu „ekonomii uwagi” – zarabiają na czasie, jaki spędzamy przed ekranem. Ponieważ fake newsy budzą skrajne emocje, algorytmy promują je priorytetowo. Prowadzi to do zamykania się użytkowników w tzw. bańkach informacyjnych (komorach echa), gdzie krytyczne myślenie zostaje uśpione. Metody obrony i weryfikacji: Tradycyjna analiza tekstu już nie wystarcza. Eksperti zalecają stosowanie „czytania horyzontalnego” (weryfikacja źródła, daty materiału w celu uniknięcia tzw. zombie content oraz sprawdzanie kontekstu w innych niezależnych mediach). Kluczowe jest też stosowanie zasady 10 sekund pauzy przed udostępnieniem czegokolwiek, by pozwolić zadziałać racjonalnej części mózgu. Artykuł zwraca również uwagę na techniczne aspekty manipulacji, takie jak podszywanie się pod znane portale (typosquatting) czy używanie grafik zmanipulowanych lub wygenerowanych przez sztuczną inteligencję (które można rozpoznać np. po błędach w anatomii rąk, napisach w tle czy nienaturalnym oświetleniu). W walce z dezinformacją, która stała się elementem wojen hybrydowych, pomagają profesjonalne, niezależne organizacje fact-checkingowe (np. Demagog, AFP Fact Check, Reuters Fact Check).

### Podsumowanie

Technologia nie rozwiąże za nas problemu kłamstwa w sieci. Ostateczną i najważniejszą linią obrony pozostaje nasza osobista higiena cyfrowa oraz krytyczne i sceptyczne myślenie.

## 1. Wstęp: Krajobraz informacyjny w epoce post-prawdy

Żyjemy w czasach, w których tradycyjne definicje komunikacji społecznej i przekazu medialnego uległy całkowitej dekonstrukcji. Wstąpiliśmy w epokę post-prawdy – rzeczywistość, w której obiektywne fakty mają mniejszy wpływ na kształtowanie opinii publicznej niż odwoływanie się do emocji i osobistych przekonań. Informacja, która przez wieki pełniła przede wszystkim funkcję opisową, edukacyjną i kronikarską, w XXI wieku została zredefiniowana. Dziś przestała być jedynie neutralnym opisem obiektywnej rzeczywistości. Stała się precyzyjnie skalibrowanym narzędziem wpływu, zasobem ekonomicznym, a w skrajnych przypadkach – niebezpieczną cyfrową bronią o charakterze geopolitycznym.

Codziennosc przeciętnego użytkownika sieci to nieustanne brodzenie w gęszczu danych. Zjawisko to, nazywane przez socjologów mediów „smogiem informacyjnym” lub „infodemią”, sprawia, że ludzki umysł każdego dnia bombardowany jest tysiącami bodźców. Paradoksalnie, łatwość dostępu do wiedzy nie uczyniła nas społeczeństwem bardziej uświadomionym; wręcz przeciwnie – nadmiar bodźców osłabił nasze mechanizmy obronne.

Często myślimy o technologii w kontekście kodu źródłowego, zaawansowanych algorytmów czy infrastruktury sprzętowej. Zapominamy jednak, że najważniejszym procesorem w całym tym złożonym ekosystemie jest ludzki mózg. To właśnie ten biologiczny „procesor” musi obecnie radzić sobie z presją czasu, manipulacją i intencjonalnym zniekształcaniem rzeczywistości.

W tym kontekście kluczowym pojęciem staje się cyfrowa odporność. Niezwykle ważne jest zrozumienie, że odporność ta nie jest cechą stałą, wrodzonym talentem ani zainstalowanym raz na zawsze programem antywirusowym. To dynamiczny, skomplikowany proces, który jako użytkownicy sieci musimy świadomie, krytycznie i systematycznie pielęgnować każdego dnia. Zrozumienie mechanizmów dezinformacji oraz wdrożenie codziennej higieny informacyjnej to nie tylko kwestia ochrony własnego komfortu psychicznego, ale przede wszystkim fundament bezpieczeństwa nowoczesnego społeczeństwa obywatelskiego.

## **2. Anatomia psychologiczna dezinformacji: Dlaczego dajemy się nabierać?**

Aby skutecznie przeciwdziałać zjawisku dezinformacji i fake newsów, nie wystarczy poznać same narzędzia technologiczne. Konieczne jest głębokie zanurzenie się w psychologię człowieka i zrozumienie, jak nasz mózg przetwarza informacje. Twórcy fałszywych wiadomości rzadko działają po omacku. Ich sukces opiera się na precyzyjnym wykorzystywaniu ewolucyjnych i kognitywnych słabości ludzkiego aparatu poznawczego.

### **2.1. Prymat emocji nad logiką**

Kluczem do zrozumienia, dlaczego dezinformacja rozprzestrzenia się w tak zastraszającym tempie, jest uznanie absolutnego prymatu emocji nad chłodną logiką. Badania nad dynamiką sieci społecznościowych (m.in. słynne analizy Massachusetts Institute of Technology) jednoznacznie dowodzą, że wiadomości fałszywe docierają do odbiorców wielokrotnie szybciej, głębiej i szerzej niż wiadomości prawdziwe.

Dlaczego tak się dzieje? Odpowiedź tkwi w ładunku emocjonalnym. Treści budzące strach, głęboką złość, oburzenie moralne czy tanią sensację posiadają najwyższy potencjał wiralowy. W ewolucyjnym ujęciu nasz mózg jest wyczulony na sygnały o zagrożeniu. Wiadomość o charakterze alarmującym („Uwaga! Ukrywany fakt!”, „To grozi nam wszystkim!”) natychmiast stymuluje ciało migdałowate, wywołując silną reakcję afektywną. Rzetelne fakty, obwarowane metodologicznym sceptycyzmem, statystykami i niuansami, są dla ludzkiego mózgu po prostu nudne i mało stymulujące. W walce o naszą ograniczoną uwagę ewolucyjnie uwarunkowane emocje niemal zawsze wygrywają z racjonalną analizą.

### **2.2. Heurystyki, czyli myślenie na skróty**

Ludzki mózg, choć niezwykle potężny, jest organem energochłonnym. Aby optymalizować codzienne funkcjonowanie, nasz aparat poznawczy wykorzystuje heurystyki – czyli uproszczone reguły wnioskowania, myślowe „skrót”, które pozwalają na błyskawiczne podejmowanie decyzji bez ponoszenia dużych kosztów energetycznych.

W środowisku pierwotnym mechanizm ten decydował o przetrwaniu (szybka ucieczka przed szelestem w krzakach). W środowisku cyfrowym staje się on jednak naszą największą zgubą. Zamiast poddawać każdy napotkany artykuł skomplikowanej weryfikacji, nasz umysł wybiera drogę na skróty. Jeśli dana informacja brzmi znajomo, jest łatwa w odbiorze (prosty, kategoriyczny język) i pasuje do naszych dotychczasowych doświadczeń, mózg automatycznie

uznaje ją za prawdziwą. Zjawisko to, powiązane z tzw. płynnością poznawczą, sprawia, że reagujemy impulsywnie i odruchowo, stając się bezbronnymi ofiarami manipulacji.

### **2.3. Błąd potwierdzenia (Confirmation Bias)**

Najgroźniejszym i najbardziej powszechnym mechanizmem psychologicznym wykorzystywanym w architekturze dezinformacji jest błąd potwierdzenia. Jest to podświadoma tendencja do preferowania, wyszukiwania i obdarzania większym zaufaniem tych informacji, które są w pełni zgodne z naszym aktualnym światopoglądem, systemem wartości czy politycznymi uprzedzeniami.

Jeśli internauta natrafia na nagłówek, który idealnie wpisuje się w jego dotychczasowe lęki lub sympatie polityczne, w jego umyśle dochodzi do natychmiastowego wyłączenia krytycznej analizy. Taka informacja nie jest kwestionowana, ponieważ jej akceptacja przynosi gratyfikację w postaci poczucia racji i komfortu psychicznego. W tym momencie człowiek przestaje być autonomicznym odbiorcą obiektywnej rzeczywistości, a staje się zakładnikiem własnych uprzedzeń oraz algorytmów, które te uprzedzenia karmią.

## **3. Ekonomia uwagi i mechanizmy platform społecznościowych**

Zrozumienie psychologicznej podatności jednostki musi zostać uzupełnione o analizę środowiska techniczno-ekonomicznego, w którym dochodzi do dystrybucji treści. Współczesny internet opiera się na modelu biznesowym określanym jako ekonomia uwagi. W tym modelu to nie produkt czy usługa są główną walutą – najcenniejszą wartością, o którą toczy się bezwzględna walka, jest czas i zaangażowanie użytkownika.

### **3.1. Architektura algorytmiczna zaangażowania**

Platformy społecznościowe (takie jak Meta, X, TikTok czy YouTube) zarabiają przede wszystkim na wyświetlaniu reklam. Aby maksymalizować zyski, algorytmy tych serwisów zostały zaprojektowane w jeden nadrzędny sposób: mają utrzymać użytkownika przed ekranem smartfona lub komputera jak najdłużej. Algorytm nie posiada moralności ani kategorii prawdy i fałszu. Analizuje on wyłącznie wskaźnik zaangażowania (engagement rate) – kliknięcia, komentarze, udostępnienia oraz czas spędzony na analizie danego posta.

Ponieważ treści kontrowersyjne, polaryzujące i generujące skrajne emocje (czyli w dużej mierze fake newsy) wywołują najsilniejsze reakcje, algorytmy sztucznej inteligencji zaczynają je promować w sposób priorytetowy. Systemy rekomendacyjne szybko uczą się, że podsuniecie użytkownikowi teorii spiskowej lub zmanipulowanego newsa politycznego wywoła falę komentarzy i zmusi go do pozostania na platformie. W ten sposób algorytmy technologii informacyjnych stają się potężnymi, choć bezdusznymi multiplikatorami kłamstwa.

### **3.2. Bańki informacyjne i komory echa (Echo Chambers)**

Konsekwencją takiego stanu rzeczy jest powstawanie tzw. baniek informacyjnych (filtracyjnych) oraz komór echa. Algorytm, chcąc dostarczyć nam maksymalną przyjemność z konsumpcji treści, odcina nas od informacji niewygodnych, skomplikowanych lub sprzecznych z naszymi preferencjami. Zaczynamy funkcjonować w cyfrowym silosie, gdzie docierają do nas wyłącznie głosy podobne do naszego własnego.

Wewnątrz komory echa każda, nawet najbardziej niedorzeczna informacja fałszywa, zyskuje status prawdy objawionej. Skoro wszyscy w mojej „bańce” udostępniają dany materiał, a głosy krytyczne zostały odfiltrowane przez algorytm, mój mózg otrzymuje fałszywy dowód społeczny: „skoro wszyscy tak mówią, to musi być prawda”. To zjawisko prowadzi do głębokiej polaryzacji społeczeństw, niszczenia debaty publicznej i zaniku zaufania do instytucji naukowych oraz eksperckich.

## 4. Metodologia weryfikacji: Czytanie horyzontalne w praktyce

W obliczu tak skonstruowanego środowiska cyfrowego oraz naszych naturalnych podatności psychologicznych, tradycyjne metody oceny wiarygodności informacji okazują się niewystarczające. Przez lata edukacja medialna uczyła nas tzw. czytania wertykalnego (pionowego) – dokładnego analizowania tekstu od góry do dołu, badania jego struktury, stylu oraz profesjonalnego wyglądu strony. Współcześni twórcy dezinformacji potrafią jednak perfekcyjnie podrobić szatę graficzną szanowanych portali.

W odpowiedzi na te wyzwania, badacze z Uniwersytetu Stanforda oraz eksperci zajmujący się nauką o mediach zaproponowali rewolucyjną zmianę paradygmatu: czytanie horyzontalne (poziome), znane również w praktyce jako „Metoda 3 kroków”. Zamiast zostawać na podejrzanej stronie i zagłębiać się w jej treść, weryfikator powinien natychmiast „wyjść” z niej w bok – otworzyć nowe karty w przeglądarce i zacząć badać dany podmiot z zewnętrznej perspektywy. Metoda ta opiera się na trzech funkcjonalnych filarach:

### 4.1. Krok 1: Źródło (Kto stoi za informacją?)

Pierwszym odruchem po natrafieniu na sensacyjną wiadomość powinno być zbadanie afiliacji autora oraz wiarygodności samej platformy publikującej treść. Należy zadać sobie kluczowe pytania: Czy autor jest uznanym ekspertem w danej dziedzinie, czy może anonimowym kontem w mediach społecznościowych, które zostało utworzone zaledwie dwa dni temu?

W przypadku portali internetowych kluczowa jest zakładka „O nas” lub „Kontakt”. Brak transparentności, brak nazwisk redaktorów czy brak fizycznego adresu redakcji to natychmiastowy sygnał ostrzegawczy. Czytanie horyzontalne wymaga, abyśmy w wyszukiwarce wpisali nazwę portalu lub nazwisko autora z dopiskiem „opinie” lub „kontrowersje”, sprawdzając, co o danym źródle piszą inne, niezależne podmioty.

### 4.2. Krok 2: Data (Pułapka „zombie content”)

Dezinformacja niezwykle chętnie i skutecznie wykorzystuje tzw. zombie content. Jest to mechanizm polegający na wyciąganiu z cyfrowego archiwum starych, autentycznych nagrań, artykułów lub zdjęć i prezentowaniu ich w zupełnie nowym, zmanipulowanym kontekście czasowym. Przykładowo, zdjęcia z protestów społecznych sprzed pięciu lat lub klęsk żywiołowych z innego kontynentu są opisywane jako wydarzenia, które dzieją się „tu i teraz”. Zawsze należy bezwzględnie weryfikować pierwotną datę publikacji materiału źródłowego, a w przypadku obrazów – stosować wsteczne wyszukiwanie grafiką.

### 4.3. Krok 3: Kontekst (Zasada konsensusu informacyjnego)

Trzeci krok polega na sprawdzeniu, czy sensacyjna wiadomość jest potwierdzona przez inne, niezależne i powszechnie uznane źródła informacji (agencje prasowe takie jak PAP, Reuters, AFP czy oficjalne komunikaty instytucjonalne). Jeśli informacja o rzekomym przełomowym wydarzeniu politycznym lub katastrofie pojawia się wyłącznie na jednym, niszowym blogu lub profilu społecznościowym, a milczą o niej wszystkie profesjonalne redakcje informacyjne, z ogromnym prawdopodobieństwem mamy do czynienia z fabrykacją.

## 5. Edukacja i higiena cyfrowa jako codzienne nawyki

Walka z dezinformacją nie może ograniczać się wyłącznie do doraźnego sprawdzania faktów. Musi stać się elementem stałej, codziennej troski o nasze zdrowie psychiczne i poznawcze. Higiena cyfrowa to zbiór praktyk, które pozwalają minimalizować negatywny wpływ środowiska sieciowego na nasze życie.

Jednym z najważniejszych nawyków jest wprowadzenie zasady ograniczonego zaufania i celowe spowolnienie naszych reakcji. Zanim klikniemy „udostępnij”, zanim prześlemy dalej bulwersującą wiadomość w rodzinnym czacie, dajmy sobie czas na ochłonięcie. Emocje są najgorszym doradcą w świecie cyfrowym. Kolejnym krokiem jest

świadome dywersyfikowanie źródeł informacji – celowe szukanie punktów widzenia innych niż nasz własny, by nie dać się zamknąć w algorytmicznej klatce komory echa. Pamiętajmy również o regularnym „detoksie cyfrowym” – odłożenie smartfona na kilka godzin pozwala naszemu układowi nerwowemu zregenerować się i odzyskać zdolność do jasnego, krytycznego myślenia.

## **6. Techniczna detekcja manipulacji obrazem i tekstem**

Współczesna dezinformacja rzadko opiera się na prostym kłamstwie tekstowym. O wiele częściej wykorzystuje manipulację wielomodalną, łącząc tekst z odpowiednio zmodyfikowanym obrazem lub materiałem wideo. Nasza cyfrowa odporność musi więc obejmować podstawowe umiejętności technicznej detekcji fałszerstw.

W przypadku obrazów najprostszą, a zarazem niezwykle skuteczną metodą jest wsteczne wyszukiwanie grafiki (oferowane m.in. przez Google Images, TinEye czy Yandex). Pozwala ono w kilka sekund ustalić, kiedy dany obraz po raz pierwszy pojawił się w sieci i w jakim kontekście był pierwotnie użyty. Analizując samo zdjęcie, warto zwracać uwagę na niekonsekwencje w oświetleniu, nienaturalne cienie, rozmyte krawędzie wokół wklejonych obiektów lub nietypowe proporcje elementów architektonicznych. W przypadku tekstów sygnałem ostrzegawczym powinna być nadmierna sensacyjność nagłówków (tzw. clickbait), błędy interpunkcyjne i ortograficzne (często wynikające z nieudolnego automatycznego tłumaczenia zagranicznych materiałów propagandowych) oraz brak wskazania konkretnych źródeł danych.

## **7. Jak rozpoznać kreacje wygenerowane przez Sztuczną Inteligencję (AI)**

Wraz z upowszechnieniem się zaawansowanych modeli generatywnych (takich jak Midjourney, DALL-E czy ChatGPT), tworzenie niezwykle realistycznych, a zarazem całkowicie fałszywych treści stało się banalnie proste i tanie. Stoimy w obliczu zalewu tzw. syntetycznych mediów, co wymaga od nas wykształcenia nowych umiejętności ich identyfikacji.

Rozpoznawanie obrazów wygenerowanych przez sztuczną inteligencję wymaga czujności i skupienia na detalach, z którymi algorytmy wciąż mają spore problemy. Przede wszystkim należy dokładnie przyjrzeć się dłoniom i stopom postaci – AI często generuje nienaturalną liczbę palców lub zniekształca ich anatomię. Kolejnym elementem są napisy i tekst w tle obrazu – sztuczna inteligencja ma tendencję do tworzenia niezrozumiałych, rozmytych lub zniekształconych liter przypominających obce pismo. Warto też analizować symetrię twarzy, wygląd biżuterii (kolczyki, które różnią się od siebie kształtem) oraz nienaturalnie gładką, niemal plastikową strukturę skóry postaci. W przypadku materiałów wideo typu deepfake, kluczowe jest obserwowanie nienaturalnego mrugania oczu (lub jego całkowitego braku) oraz braku idealnej synchronizacji ruchu warg z wypowiedzanym dźwiękiem.

## **8. Wojna informacyjna i instytucjonalny wymiar fact-checkingu**

Musimy mieć świadomość, że dezinformacja to nie tylko domena internetowych żartownisiów czy twórców teorii spiskowych szukających poklasku. We współczesnym świecie stała się ona pełnoprawnym elementem działań wojennych – tzw. wojny hybrydowej. Wyspecjalizowane, państwowe ośrodki propagandowe (tzw. fabryki trolli) oraz cyberarmie prowadzą skoordynowane operacje informacyjne, których celem jest destabilizacja sytuacji politycznej, polaryzacja nastrojów społecznych, sianie paniki oraz podważanie zaufania do instytucji demokratycznych, naukowych i sojuszy międzynarodowych.

W tej nierównej walce pojedynczy użytkownik nie jest jednak pozostawiony sam sobie. Kluczowym elementem infrastruktury obronnej współczesnego internetu są niezależne organizacje fact-checkingowe. W Polsce i na świecie działają profesjonalne zespoły analityków (takie jak m.in. Stowarzyszenie Demagog, AFP Sprawdzam czy Konkret24), które codziennie monitorują przestrzeń medialną, weryfikują podejrzane doniesienia i publikują szczegółowe raporty obalające kłamstwa. Korzystanie z zasobów tych organizacji, wspieranie ich działalności oraz traktowanie ich jako pierwszego punktu odniesienia w przypadku wątpliwych informacji jest fundamentalnym elementem naszej zbiorowej odporności cyfrowej.

## 9. Konkluzje i wnioski: Przyszłość ekosystemu informacyjnego

Zjawisko dezinformacji i fake newsów nie zniknie – wręcz przeciwnie, wraz z rozwojem technologicznym będzie stawało się coraz bardziej wyrafinowane, trudniejsze do wykrycia i wszechobecne. Technologia, która miała dać nam nieograniczony dostęp do wiedzy, stała się jednocześnie idealnym narzędziem manipulacji na masową skalę. Przyszłość naszego ekosystemu informacyjnego zależy od naszej wspólnej reakcji na to zagrożenie.

Budowanie cyfrowej odporności to proces, który wymaga zaangażowania na wielu poziomach: od systemowych zmian prawnych i technologicznych na platformach społecznościowych, przez wspieranie profesjonalnego dziennikarstwa i instytucji fact-checkingowych, aż po oddolną edukację medialną. Najważniejszy krok zaczyna się jednak od nas samych. Każdy z nas, logując się do sieci, bierze na siebie odpowiedzialność za to, jakie treści konsumuje i jakie podaje dalej. Wymaga to rezygnacji z natychmiastowych reakcji na sensację oraz ciągłego podnoszenia kompetencji medialnych.

Uwaga to najcenniejsza waluta: W świecie permanentnego przeładowania danymi, nasza uwaga i czas są najdroższymi zasobami na rynku. Musimy przestać być biernymi konsumentami cyfrowego szumu, a stać się dumnymi, świadomymi i nieustępliwymi strażnikami własnego umysłu. Tylko w ten sposób, poprzez codzienne wdrażanie higieny cyfrowej i metod weryfikacji, będziemy w stanie wspólnie budować bezpieczną, rzetelną i opartą na prawdzie przestrzeń informacyjną dla przyszłych pokoleń.

### Bibliografia

- [1] Vosoughi, S., Roy, D., & Aral, S. (2018). The spread of true and false news online. „Science”, 359(6380), 1146-1151.
- [2] Kahneman, D. (2012). Pułapki myślenia. O myśleniu szybkim i wolnym. Wydawnictwo Media Rodzina.
- [3] Wu, T. (2017). The Attention Merchants: The Epic Scramble to Get Inside Our Heads. Vintage Books.
- [4] Pariser, E. (2011). The Filter Bubble: What the Internet Is Hiding from You. Penguin Press.
- [5] Wineburg, S., & McGrew, S. (2019). Lateral Reading: Reading Less and Learning More When Evaluating Digital Information. Stanford History Education Group. Working Paper No. 2017-A1.
- [6] Stowarzyszenie Demagog. (n.d.). Demagog. <https://demagog.org.pl>
- [7] Agence France-Presse. (n.d.). AFP Fact Check. <https://factcheck.afp.com>
- [8] Reuters. (n.d.). Reuters Fact Check. <https://www.reuters.com/fact-check>
- [9] <https://edukacjaustawicznadoroslych.eu/>
- [10] Opracowanie algorytmiczne (Sztuczna Inteligencja)
- [11] Materiały własne i wiedza autorska