https://arxiv.org/html/2603.23438v1. (дата звернення: 26.04.2026)
6. Dimitri Galli, Andrea Venturi, Dario Stabili, Mauro Andreolini, Mirco Marchetti. (2025). Defending Network Intrusion Detection Systems Based on Graph Neural Networks Against Structural Adversarial Attacks. – URL: https://ieeexplore.ieee.org/document/11261632. (дата звернення: 27.04.2026)

## On Evidence Deficits in Kleptography and the Application of Artificial Intelligence for Their Mitigation

UDK 004.4:056.57

Mykhailo Shelest[1], Yuliia Tkach[2], Oleksandr Polevod[3]

*National University "Chernihiv Polytechnic",*
*[1]mishel3141@gmail.com, [2]tkachym79@gmail.com,*
*[3]oleksandr.polevod23@gmail.com*

Modern information systems are increasingly viewed not only as targets of attacks, but as potentially controllable environments in which hidden intervention may occur without explicit violation of functionality. Within the kleptographic paradigm [1], the central problem is not merely the fact of compromise, but the *deficit of evidence*, i.e., the inability to reliably detect and prove the presence of hidden influence.

This work proposes a formalization of a class of evidence deficits in kleptography and introduces their classification as a distinct object of cybersecurity analysis. The proposed approach enables a transition from descriptive consideration of kleptographic threats to a systematic analysis of the limitations of the evidentiary base and the conditions under which such limitations manifest.

Unlike classical information security incidents, kleptographic interventions may leave no unambiguous technical traces, which creates fundamental constraints for their detection and attribution. In this context, we propose to treat these constraints as formalized evidence deficits that define the boundaries of applicability of traditional analysis methods.

Several key classes of such deficits can be identified.

First, the *reproducibility deficit* manifests itself in the instability of anomalies that arise only under specific conditions, such as particular system states or environmental configurations [2]. This prevents their reliable reproduction in laboratory settings. For example, in systems with dependencies on third-party libraries, behavioral deviations may occur only under narrowly defined input conditions that are not captured by standard testing procedures. A typical case involves a modified software component that exhibits correct behavior during testing but demonstrates selective or controlled behavior in real-world deployment, making it practically non-reproducible.

Second, the *causality deficit* refers to situations where an observable effect exists, but the underlying mechanism cannot be isolated or distinguished from normal system behavior.

Third, the *provenance deficit* is associated with the inability to reliably verify the origin of software artifacts, which is particularly critical in the context of supply chain attacks.

In addition, the *invariant deficit* limits the formalization of "normal" system behavior, while the *subject attribution deficit* complicates the identification of the responsible entity behind a potential intervention.

Traditional cybersecurity methods rely on the assumption that explicit indicators of compromise exist. However, within the kleptographic model, such indicators may be absent or intentionally masked. Hidden controllability may be implemented through conditional activation, rare triggers, or selective modification of system behavior [2], rendering classical incident-based approaches ineffective.

Under these conditions, a fundamental shift occurs from the model of "incident detection" to the model of "analysis of potential controllability". This implies that the object of study is not only the fact of a security breach, but the very possibility of hidden influence, even in the absence of observable incidents. Such a shift transforms the classical logic of cybersecurity and requires the development of new analytical methods.

In this context, we propose an approach to the use of artificial intelligence as a tool for partial compensation of evidence deficits. AI enables the analysis of high-dimensional and weakly structured data, allowing the detection of latent patterns and anomalies that are not captured by classical methods.

To address the reproducibility deficit, clustering and anomaly detection techniques can be applied to identify unstable patterns. The causality deficit can be partially mitigated through the analysis of statistical dependencies between system parameters and behavior. In provenance-related tasks, graph-based models can be used to analyze dependencies and detect anomalous components within supply chains.

AI plays a particularly important role in the analysis of intelligent systems, where kleptographic controllability may manifest not as a technical anomaly, but as a semantic shift or selective system behavior [3]. In such cases, AI acts as a hypothesis-generation tool for identifying potential mechanisms of hidden influence.

At the same time, the use of AI has significant limitations. Models themselves may be subject to manipulation or contain embedded backdoors, and their outputs are inherently probabilistic. Therefore, AI should not be considered a tool for proof, but rather a means of supporting analytical reasoning.

Thus, kleptography establishes a new research paradigm in which the central challenge is not the detection of incidents, but overcoming the limitations of evidence. The proposed classification of evidence deficits provides a structured framework for analyzing hidden controllability, while the integration of artificial intelligence methods opens new opportunities for detecting anomalies and forming well-grounded hypotheses. Kleptography shifts cybersecurity from the problem of detecting attacks to the fundamentally harder problem of proving that hidden control exists.

1. Shelest M. Ye., Tkach Yu. M. Kleptography: From Backdoor to the Politics of Trust in the Digital Era. Nizhyn, 2025.
2. Böhme R., Freedman M. et al. Toward Systematic Classification of Cybercrime Events. WEIS, 2015.
3. Sommer R., Paxson V. Outside the Closed World: On Using Machine Learning for Network Intrusion Detection. IEEE S&P, 2010.