Cybersecurity from Within: How Embedded Backdoors Undermine Trust in State and International Systems ⋆

Oleksandr Polevod[1,*,†], Mykhailo Shelest[1,†], Yuliia Tkach[1,†]

[1] *Chernihiv Polytechnik National University, 14000 Chernihiv, Ukraine*

Abstract

The article examines the phenomenon of kleptography—the deliberate implantation of backdoors into cryptographic algorithms, protocols, hardware and software components, as well as into certification and institutional mechanisms of digital trust. It traces the historical roots of kleptography in intelligence-agency practices, reviews emblematic cases (Crypto AG, Dual_EC_DRBG, etc.), and highlights the geostrategic dimension of the problem. The paper proposes the concept of "klepto-hygiene" as a preventive approach to trust management in public policy and international relations. It also formulates recommendations on independent auditing, supply-chain transparency, the right to audit, and prioritizing open standards.

Keywords

kleptography, backdoor, digital trust, supply chain, Zero Trust, klepto-hygiene, national security, international standards.

1. Introduction

Digital trust has become a strategic resource of the 21st century. It determines the continuity of public services, the reliability of security and defense communications, the resilience of financial transactions, and the effectiveness of international institutions. Traditional cybersecurity paradigms focused on external threats, whereas a critically dangerous vector has turned out to be internal—the implantation of controlled vulnerabilities (backdoors) during the design, standardization, assembly, and certification stages of components [1–2]. In this context, we view kleptography as a systemic practice of creating "secure" technologies with hidden access mechanisms. This practice undermines the very foundation of trust—not through accidental errors, but through architectural decisions in which control is embedded as a property of the design.

2. Origins of kleptohraphy in intelligence practices

Kleptography stems from the needs of intelligence and security agencies to secure privileged access to protected communication channels. Historically, U.S. agencies (NSA) influenced cryptographic standards (the Dual_EC_DRBG episode) [3–4], promoted hardware solutions with special access mechanisms (Clipper Chip) [2]; GCHQ played a role in international crypto-politics [2]; Soviet/Russian agencies (KGB/FAPSI/FSB) centralized the licensing of encryption tools and controlled their export [2]. Common features of these approaches included:

- influence over standardization and certification processes;
- management of supply chains;
- covert integration of control into hardware/software modules.

⋆ *SMICS'25: Workshop on Cryptology and Data Security, October 16-18, 2025, Lviv, Ukraine*

1,* Corresponding author.

† These authors contributed equally.

oleksandr.polevod23@gmail.com (O. Polevod); mishel3141@gmail.com (M. Shelest); tkachym79@gmail.com (Y. Tkach)

0009-0007-0885-8625 (O. Polevod); 0000-0001-7110-4876 (M. Shelest); 0000-0002-8565-0525 (Y. Tkach)

Thus, kleptography is not a “market anomaly” but an institutional strategy that eventually spread into commercial products and global standards.

3. Definition of kleptography: from narrow to broad

Originally, in the formulation of A. Young and M. Yung [1–2], kleptography was understood as “cryptography against cryptography”: a hidden mechanism (backdoor) is deliberately embedded in a cryptographic algorithm or protocol, enabling a third party to recover keys or messages while remaining practically undetectable through conventional analysis. This “narrow” core of kleptography focuses on mathematical constructs and protocol steps.

However, the experience of the past decades has shown that the viability of such backdoors depends not only on mathematics but also on the socio-technical context: who approves standards, how certification procedures are organized, how supply chains and build processes are structured, how updates and keys are managed, which CI/CD tools and repositories are used, and who has access to firmware and microcode. It is precisely these “extra-cryptographic” layers that often create the conditions under which a backdoor can appear unnoticed, persist for a long time, and spread widely. Therefore, we propose moving from a narrow to a broad understanding of kleptography that encompasses managerial, organizational, and institutional practices [2; 14, 17–19].

Narrow definition. Kleptography is the discipline concerned with the design and integration of cryptographic algorithms and protocols with deliberate backdoors that allow a third party to recover keys or messages without providing the owner with any means of detection [1].

Broad definition. Kleptography is the totality of technical, organizational, and institutional practices of covertly embedding and exploiting controlled vulnerabilities in any layer of digital infrastructure: algorithms and protocols; software libraries and SDKs; firmware and microcode; hardware modules (HSM/TPM/SoC); build, signing, and update systems; and procedures of standardization, certification, and procurement [2; 17–19].

Principal distinctions between the narrow and broad approaches:

1. Object of control: from purely cryptographic constructs → to socio-technical systems (code + processes + institutions).
2. Mechanism of access: from hidden parameters/protocol steps → to organizational and supply-chain levers (supply chain, standards, trusted centers).
3. Detection methods: from cryptanalysis of artifacts → to process auditing (SBOM, reproducible builds, certification transparency, update traceability).
4. Consequences and accountability: from a technical “anomaly” → to a politico-legal problem of trust and institutional responsibility.

Key implication of the broad perspective: trust itself becomes an attack surface. When the reputation of a standard or supplier is used as cover, traditional code audits are insufficient—the “trap” may be embedded at the level of procedures and rules of the game. Hence, the logical continuation is our framework of klepto-hygiene: policies and practices that render trust verifiable and manageable, rather than merely declarative [14, 19].

4. Typology and lifecycle of a backdoor

One of the most dangerous characteristics of a kleptographic backdoor is its ability to remain unnoticed for extended periods of time. The absence of behavioral anomalies, formal compliance with cryptographic standards, and the use of certified components—all of these allow the hidden mechanism to evade detection by both technical audits and institutional oversight. Concealment is an inherent phase of a backdoor’s lifecycle, serving the strategic function of preserving invisibility during deployment, activation, and exploitation.

Methods of concealing kleptographic backdoors can be classified across three main levels: architectural, software-cryptographic, and operational.

1. **Architectural concealment.** This involves embedding the backdoor at the level of design or standards, where the very existence of a hidden mechanism remains invisible, and in some cases even formally legitimized:
    - Exploiting permissible options within a standard. For example, in the case of Dual_EC_DRBG, the generation mechanism itself was formally allowed in NIST SP 800-90A, but the specific parameters enabled potential exfiltration.
    - Embedding within a multi-component system where responsibility for security is distributed across several modules. This blurs audit boundaries and complicates the attribution of a vulnerability to a particular block.
    - Service or maintenance interfaces presented as part of technical support but containing uncontrolled access mechanisms.
2. **Software-cryptographic concealment.** At this level, techniques of encryption, encoding, or obfuscation are used to make the backdoor mechanism opaque to auditors:
    - Code obfuscation or the use of excessive complexity (e.g., embedded crypto-libraries with multiple levels of calls, conditional access, or dynamically generated dependencies).
    - Masking as standard cryptographic mechanisms. A backdoor may use a common protocol (e.g., TLS) but with atypical parameters or altered generation points enabling covert leakage.
    - Use of one-way triggers that leave no traces in audit logs and lack explicit signatures for IDS/IPS systems.
3. **Operational concealment.** This type is realized during system operation, with the aim of avoiding detection in real use:
    - Data exfiltration through nonstandard channels: DNS queries, HTTP/HTTPS patterns, timing steganography, parameters of TLS Client Hello, or distortion of entropy profiles.
    - Imitation of normal behavior. For example, the backdoor may only respond to an exact sequence of actions (challenge-response), minimizing the chance of accidental activation.
    - Self-destruction or "sleeping" mode, allowing the component to erase itself after activation or disable functionality if anomalous monitoring is detected.

Specific cases of concealment include:

- *Backdoor.Juniper (2015):* the backdoor in the NetScreen VPN component maintained full encryption functionality but allowed decryption by a third party with knowledge of generator parameters. Concealment relied on closed-source code and a controlled update chain.
- *X.509 certificate injection:* backdoors in certificate chains exploit the trust in Certificate Authorities, with concealment achieved by creating formally legitimate but pre-controlled certificates.

Concealment is not merely a technical trick but part of a long-term strategy to retain control without exposure. In kleptography, it acquires an especially deep character: by merging with the norm, the backdoor ceases to resemble a "hack" and becomes an "architectural condition." This is why detection policies must consider anomalous patterns, not only suspicious signatures or atypical functions. To highlight where controlled vulnerabilities may hide, we propose the following classification:

1. Algorithmic — traps in mathematics/parameters (e.g., DRBG, curves) enabling key recovery [1, 3–4].
2. Protocol-level — modifications in handshakes, entropy/IV management, or timing that enable controlled bypass [1–2].
3. Software — backdoors in libraries/SDKs, update mechanisms, and dependencies (supply chain) [5–6, 11–12].

4. Firmware — hidden hooks in networking/IoT firmware, bootloaders, cryptomodules/PRNGs [5–6].
5. Hardware — modified IP blocks, controllers, additional coprocessors (HSM/TPM/SoC) [2].
6. Institutional — standards/certification/procurement practices that legitimize opaque components while minimizing real audits [3–4, 14].

The lifecycle model of a backdoor helps in planning detection and organizing control and countermeasures. It consists of the following phases:

1. Design: selecting control points and activation conditions (key/version/configuration).
2. Integration: embedding into code, firmware, hardware, or even standardization procedures.
3. Propagation: distribution through dependencies, updates, certification, or supply chains.
4. Activation: delayed or conditional, in order to preserve plausibility.
5. Exploitation: low-noise access/metadata collection, avoidance of artifacts in logs.
6. Concealment/deactivation: "fix" patches, branch replacements, rebranding, or appeals to "compatibility."

The lifecycle of a kleptographic backdoor is deliberately optimized for a low probability of detection and a high strategic impact [19].

5. Case review

Kleptography is a weapon of strategic scale: it grants control over the infrastructure and metadata of partners or adversaries without the need for overt "hacks" [1–2, 7–12]. Below are several of the most well-known cases of kleptographic implants:

- **Crypto AG.** Modified cipher machines provided the "appearance" of security while in fact enabling intelligence services to monitor the traffic of more than 120 countries. The exposure of Operation Rubicon (CIA/BND cooperation) was documented through journalistic investigations and official Swiss reports [7–10].
- **Dual_EC_DRBG.** A recommended deterministic random bit generator standard was later withdrawn from NIST guidelines after the discovery of potential backdoor risks, though it had already seen some deployment [3–4].
- **Juniper ScreenOS.** Vulnerabilities and backdoor mechanisms in authentication and key generation were discovered in the firmware of network equipment [5–6].
- **XZ Backdoor (2024).** A backdoor in the popular compression library was introduced under the guise of a routine patch—an incident that confirmed the fragility of the trust chain even in open source ecosystems [11–12].

6. Strategic Implications for Public Policy and International Relations

Kleptography is not limited to isolated technical incidents; it represents a systemic risk in which technology, trust institutions, and political power intersect [17–20]. Embedded backdoors reshape the architecture of trust: they infiltrate supply chains, pass certification processes, accumulate hidden presence in critical services, and eventually become tools of influence. To illustrate this evolution from technological cause to geopolitical consequence, we apply a four-step causal framework:

1. Supply chain — point of entry (cause). Any component, from a crypto-library to a semiconductor factory, can serve as the insertion point of a backdoor. Through networks of dependencies, updates, and code reuse, a single change can invisibly scale across entire ecosystems, compromising even "flawless" products [11–12, 14, 17].
2. Digital sovereignty — domestic consequence. When state services (communications, registries, finance) rely on compromised chains, hidden access mechanisms and external levers of influence over critical infrastructure emerge, resulting in partial loss of sovereign control [7–12].

3. International trust — external consequence. Once technical trust is undermined, institutional trust collapses as well: standards and certifications cease to be perceived as impartial, and cooperation requires mutual independent audits, transparent SBOMs, and verifiable build processes [10, 14].

4. Political weapon — strategic dimension. Control over infrastructure and metadata transforms kleptography into an instrument of asymmetric influence: technological advantages are converted into bargaining power, regulatory leverage, and intelligence capabilities [1–2, 7–10].

Illustration (MAX, Russia, 2025). The launch and mandatory pre-installation of the state messenger MAX in Russia creates a "super-app" deeply integrated into public services and payments. Such concentration of communication channels and data establishes an architecture of digital control and demonstrates kleptography in the broad sense—at the institutional level of trust [15].

7. Klepto-hygiene: Prevention and Trust Management

To shift the discussion from merely stating threats to establishing managed trust, a framework is needed that makes trust verifiable, traceable, and accountable across all layers—from code to institutions.

Such a framework is klepto-hygiene: a coordinated set of principles, processes, and evidential artifacts that reduce the probability of hidden backdoors and increase the likelihood of their early detection [13–14].

Principles (their meaning)

- Transparency — observability of artifacts (code, parameters, builds).
- Traceability — SBOM and provenance for every dependency.
- Accountability — clear responsibility and audit trails.
- Least privilege — minimal necessary access/keys.
- Separation of trust — segmentation and supplier diversification.
- Verification over declaration — independent assessment/certification instead of "certified = secure" [13–14].

Levels of Implementation

Technical level. At the technical level, klepto-hygiene is implemented through embedded tools, development practices, and verification methods that prevent or detect hidden interventions. Key mechanisms include:

- reproducible builds — ensuring identical outputs regardless of environment;
- dependency control — lock files, hash verification, SBOM review;
- cryptographic attestation — digital signatures at build, update, and delivery stages;
- integration of Software Composition Analysis (SCA) and CI/CD pipeline reviews with provenance checks;
- built-in trusted logging mechanisms.

This level ensures formal verifiability, engineering integrity, and digital traceability—the epistemological foundation of system assurance. Independent crypto-audits and protocol analysis; SBOM with version/provenance verification; reproducible/"hermetic" builds; update policies with rollback protection; entropy/key monitoring; binary transparency and artifact signing; periodic fuzzing/cryptanalysis—all address supply-chain risks (as incidents like XZ have demonstrated) [11–12, 14, 16].

Organizational level. At the organizational level, klepto-hygiene manifests as internal information security policies encompassing DevSecOps, supply-chain risk management, and access models.

Core components include:

- formalization of Secure Development Lifecycle (SDL) policies;
- segregation of duties and privilege control (e.g., least-privilege access to build servers);

- mandatory audits of critical changes — peer review with external reviewers;
- capacity-building — training personnel to recognize kleptographic patterns, using supplier checklists;
- regular testing of trust mechanisms — red team backdoor modeling, fuzzing, anomaly analysis.

This level transforms principles into everyday procedures and builds a security culture where klepto-hygiene is part of operations, not external enforcement. Examples: Zero Trust by design; dual control/four-eyes for critical actions; supplier verification by two independent labs; requirement of open implementations for critical crypto modules; mandatory peer review and red teaming for hidden-channel detection; clear SLO/SLA and metrics (e.g., reduction of MTTD/MTTR, SBOM coverage) [14, 16].

Institutional level. At the institutional level, klepto-hygiene establishes political and legal infrastructure that sets the rules of the game for suppliers, regulators, independent experts, and end users.

Key directions:

- recognition of the right to audit as a condition of market entry for critical digital infrastructure suppliers;
- regulation of transparency requirements for software and hardware (e.g., EU Cyber Resilience Act);
- mandatory SBOM declaration for all components used in public or critical sectors;
- security certification procedures with open criteria (e.g., Common Criteria, FIPS, ETSI EN 303645);
- independent expert centers (public audit labs, academic clusters) engaged in peer review, incident analysis, and standards testing.

This level enforces systemic accountability, compels transparency, and creates economic and legal incentives for klepto-hygienic practices.

For klepto-hygiene to be effective, vertical alignment across all three levels is critical. Technical measures without organizational support are ineffective; policies without technical implementation are declarative; regulation without audit is formal. Examples: legal right to audit code/firmware/microcode for the public sector and critical infrastructure; independent certification centers with public reports; international transparency mechanisms for transnational components; watchlists of risky dependencies with clear inclusion/exclusion criteria; integration of SBOM and reproducible-build requirements into government procurement. Reference frameworks: NIST SP 800-207 (Zero Trust), ENISA on supply-chain, NTIA SBOM [13–14, 16].

Operationalization (how to manage in practice)

- Policy: who is responsible (CISO / certification body / vendor).
- Process: what and how to measure (SBOM coverage, % reproducible builds, update audits).
- Artifacts: what to collect as evidence (audit reports, build logs, supplier attestations).

Without such evidence of trust, klepto-hygiene becomes a declaration; with them, it becomes a verifiable standard of practice that reduces the space for kleptography [13–14, 16].

8. Recommendations for Public Policy

In the era of digital transformation, a state's level of cybersecurity is determined not only by technical tools, regulatory mechanisms, or geopolitical alliances, but also by the level of awareness among citizens, experts, and civil servants about digital risks, threats, and countermeasures. Thus, national cyber literacy emerges as a component of digital sovereignty and a prerequisite for the effective implementation of klepto-hygiene policies.

Kleptographic interventions, due to their deeply architectural nature, cannot be fully neutralized by rigid technical solutions. They require critical awareness, trust verification, and independent thinking—across all levels, from developers and civil servants to end users. For this reason, national

cyber literacy stands as one of the key components for preventing kleptographic threats and legitimizing digital sovereignty.

Traditionally, most users perceive digital infrastructure as a “black box” that works “automatically.” Such a mindset enables architectural interventions precisely because of the absence of a critical mass of users capable of noticing, questioning, or initiating audits. Moving from “naïve trust” to “verified trust” means forming a new digital culture in which the user is not merely an object but also a subject of security.

Kleptographic threats often disguise themselves as technical details, invisible to the average user or even to specialists without targeted training. This creates the risk of uncritical decision-making, blind reliance on “authoritative” suppliers, and formal belief in certificates. For this reason, broad educational campaigns on digital hygiene must accompany all cybersecurity reforms.

Key directions for digital awareness include:

- developing basic understanding of the lifecycle of backdoors, supply chains, and parameters of cryptographic robustness;
- clarifying the risks of using unaudited services, applications, and SDKs;
- fostering skills in critically evaluating digital solutions (among citizens, journalists, civil servants);
- implementing mandatory digital hygiene courses in the public sector for staff handling protected information, including modules on backdoor detection, phishing risks, and classification of cryptographic protocols;
- public campaigns explaining the dangers of “black box” mobile apps, uncertified SDKs, and promoting self-audit practices in government agencies.

Societies lacking independent experts capable of auditing suppliers, analyzing cryptographic implementations, or verifying CAs and critical updates effectively delegate their digital security to external actors. This creates a structural asymmetry in which formal independence is not backed by intellectual verification mechanisms.

Institutional cyber literacy is impossible without expert hubs capable of conducting independent analysis of cryptographic implementations, backends, firmware, and digital signatures. Such functions may be performed by:

- technical analysis centers (CERT/CSIRT),
- university laboratories,
- civic oversight initiatives with open audit mandates.

These hubs not only strengthen the state’s capacity for independent security evaluation but also form the long-term human resource foundation for klepto-hygiene, insulated from commercial and political influence. It is therefore essential to support:

- academic laboratories for digital analysis (including firmware/SDK reverse engineering),
- national peer-review platforms with open access,
- community-based fuzzing and SBOM indexing,
- training programs for state auditors in digital infrastructure security.

The role of universities, communities, and peer review. Institutions of higher education must act as providers of cyber literacy, embedding not only technical skills but also the values of transparency, verification, and collaborative audit. Within this approach, cybersecurity ceases to be a narrow specialization and becomes part of a broader culture of integrity.

Universities should serve as centers for generating new models of digital integrity, integrating topics such as supply-chain security, architectural transparency, open audit, and risk analysis into curricula for engineers, IT specialists, and public administrators.

Critical roles include supporting:

- open peer-review communities,
- student initiatives in digital auditing,
- public platforms for sharing knowledge, cases, and vulnerabilities,
- projects such as SBOM indexing, open fuzzing, and crowdsourced reverse engineering.

National cyber literacy is not simply digital awareness; it is the intellectual infrastructure of digital independence. It combines critical thinking, technical expertise, and ethical responsibility in the digital space. In the context of kleptographic risks, cyber literacy acquires a new quality: it becomes a safeguard against digital naïveté and a form of societal architectural self-defense. Raising the national level of cyber literacy is not an optional add-on but a strategic requirement for states seeking to preserve digital independence and control over their information architecture.

Policy Recommendations:

1. National audit of critical components (cryptography, KMS/HSM, network firmware, mobile platforms) [14, 16].
2. State registry of risky dependencies, integrated into public procurement procedures [14, 16].
3. Legal right to independent auditing of code, firmware, and microcode in the public sector and critical infrastructure; reproducible builds as a mandatory requirement [16].
4. Prioritization of open standards and implementations for critical modules (with justified exceptions) [14].
5. Institutionalize klepto-hygiene in cybersecurity education programs and civil servant training [13–14, 16].

Kleptography has evolved into a systemic instrument of digital influence that demands a paradigm shift: trust must not be a precondition but the result of verification. The policies of states and international institutions must move from “certification by default” to transparent verification and managed trust in supply chains. Klepto-hygiene offers an actionable framework—from technical procedures to institutional guarantees—for restoring legitimate trust in the digital age.

Declaration on Generative AI

During the preparation of this work, the authors used ChatGPT to: translate certain text fragments into English, perform grammar and spelling checks, and paraphrase or reword content. After using these tools, the authors carefully reviewed and edited the content as needed and take full responsibility for the publication’s content.

References:

[1] A. Young, M. Yung. Kleptography: Using Cryptography Against Cryptography. In: EUROCRYPT ’97, LNCS 1233. Springer, 1997.

[2] A. Young, M. Yung. Malicious Cryptography: Exposing Cryptovirology. Wiley, 2004.

[3] NIST. NIST Removes Cryptography Algorithm from Random Number Generator Recommendations. Press release, 21 Apr 2014. URL: https://www.nist.gov/news-events/news/2014/04/nist-removes-cryptography-algorithm-random-number-generator-recommendations

[4] NIST CSRC. SP 800-90A Rev.1 announcement: Dual_EC_DRBG removed; Hash/HMAC/CTR_DRBG recommended. 2015. URL: https://csrc.nist.gov

[5] D. Goodin. Researchers confirm backdoor password in Juniper firewall code. Ars Technica, 21 Dec 2015. URL: https://arstechnica.com

[6] Rapid7. CVE-2015-7755: Juniper ScreenOS Authentication Backdoor. 20 Dec 2015. URL: https://www.rapid7.com

[7] G. Miller, P. Mekhennet, et al. The intelligence coup of the century. The Washington Post, 11 Feb 2020. URL: https://www.washingtonpost.com

[8] Reuters. Switzerland probes reports CIA and BND used Swiss firm Crypto AG to crack codes. 11 Feb 2020. URL: https://www.reuters.com

[9] The Washington Post. Swiss report reveals new details on CIA spying operation (Crypto AG). 10 Nov 2020. URL: https://www.washingtonpost.com

[10] Reuters. Swiss cabinet blames intelligence community for Crypto AG affair. 28 May 2021. URL: https://www.reuters.com
[11] CISA. Reported Supply Chain Compromise Affecting XZ Utils (CVE-2024-3094). 29 Mar 2024. URL: https://www.cisa.gov
[12] Red Hat Blog. Understanding Red Hat's response to the XZ security incident. 30 Apr 2024. URL: https://www.redhat.com
[13] S. Rose, O. Borchert, S. Mitchell, S. Connelly. Zero Trust Architecture. NIST SP 800-207, 2020. doi:10.6028/NIST.SP.800-207
[14] ENISA. Good Practices for Supply Chain Cybersecurity. 2023. URL: https://www.enisa.europa.eu
[15] Reuters. Russia orders state-backed MAX messenger app to be pre-installed on phones and tablets. 21 Aug 2025. URL: https://www.reuters.com
[16] NTIA (U.S. DOC). The Minimum Elements for a Software Bill of Materials (SBOM). 12 Jul 2021. URL: https://www.ntia.gov
[17] Polevod, O. (2024). *Kleptography in the context of information protection: Classification of kleptographic attacks.* Technical Sciences and Technologies, 4(38), 208–213. https://doi.org/10.25140/2411-5363-2024-4(38)-208-213
[18] Tkach, Y., Shelest, M., & Synenko, M. (2023). *The history of the emergence of kleptography and its place in information security.* Technical Sciences and Technologies, 3(33), 150–161. https://doi.org/10.25140/2411-5363-2023-3(33)-150-16
[19] Shelest, M. Ye., & Tkach, Y. M. (2025). *Kleptography: A new direction of cybersecurity in the digital age.* Monograph. Chernihiv. 283 p.